ANALYSIS OF FLOWERING TIME, HYBRID VIGOR, YIELD, AND LODGING IN MAIZE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Sara Johanna Larsson

January 2013

ANALYSIS OF FLOWERING TIME, HYBRID VIGOR, YIELD, AND LODGING IN MAIZE

Sara Johanna Larsson, Ph. D.

Cornell University 2013

Maize (*Zea mays* L.) is an important crop and an excellent model organism to study genetic systems. It captures remarkable diversity, which can be observed on both the genotypic and phenotypic level. Because of its diversity, maize responded very effectively to artificial selection during domestication and improvement. Maize adapted to very diverse environments. This adaptation has been possible through heritable changes in flowering time, responses to photoperiod and temperature, and plant architecture. Understanding the underlying architecture of these traits will allow us to utilize all the variation offered and increase productivity for a more sustainable agriculture. The following studies focus on analysis of three different traits.

First, is a reanalysis of one of the first generation structured association mapping studies of the *Dwarf8* locus with flowering time, using new mapping populations and statistical approaches. This trait is highly correlated with population structure, and we found that basic structured association methods overestimate the phenotypic effect in the region, while mixed model approaches perform better. Combined with analysis of the maize NAM population, it is concluded that the QTL effects at the general location of the *d8* locus are from extended haplotypes and that *d8* is not associated with flowering time.

Second, hybrids were developed using the NAM inbred population crossed to a common tester to examine hybrid vigor in terms of plant height and flowering time, as well as yield. A number of QTL were identified for all three traits using joint linkage mapping. Additionally, reasonable prediction accuracies (~0.55) were obtained using ridge regression in the hybrids. This study gives us a better understanding of yield and hybrid vigor.

Last, damage caused by lodging is a significant problem in maize production, resulting in 5–20 % annual loss in yield. In this study, more than 1,500 diverse inbred lines crossed to a common tester were evaluated across multiple environments. Due to a large sample size and despite multiple environments with lodging events occurring at different points in time, we were able to utilize joint linkage mapping to identify a number of QTL with small effects for lodging.

BIOGRAPHICAL SKETCH

Sara enrolled at the Swedish University of Agriculture Sciences with a major in horticulture. During this time she took courses in plant genetics and biotechnology and her interest and inspiration in genetics and breeding increased. She also did undergraduate work and her bachelor thesis in Dr. Margareta Welander's laboratory, mentored by Dr. Li-Hua Zhu. Sara's work focused on the evaluation of the *doa1* gene as a selectable marker in transformation of apple rootstock M26, as well as transformation of a vector containing *GA20 oxidase* gene into Agrobacterium. After receiving her Bachelor of Science in Horticulture she pursued a Masters of Science at the same university, continuing to build on part of the research she did during her undergraduate studies. Sara's interest in plant breeding continued to develop and she left the field of gene transformation to explore and get more exposure in quantitative genetics.

During her first years as a master student, Sara applied and was awarded a full scholarship for one year at Cornell University. There, she had the opportunity to attend courses, especially those focused on plant breeding, population genetics, and quantitative genetics. It was also during this year she had the opportunity to meet Dr. Edward Buckler and learn about the research performed in his laboratory at Cornell University. Sara spent the summer in the cornfield working along side the inspiring researchers in the Buckler lab, as well as learning about the interesting projects they were pursuing, and getting hands-on experience with amazingly diverse germplasm. She got an extension on her scholarship and the opportunity to stay at Cornell to finish

her research on flowering time in maize and its impact of the *Dwarf8* gene.

After receiving her Masters of Science as well as her degree as horticulturist from the Swedish University of Agriculture Sciences, she worked briefly as a laboratory technician in the Buckler Laboratory before she was admitted to the Department of Plant Breeding and Genetics at Cornell University and joined the laboratory of Dr. Edward Buckler as a graduate student.

During her time at Cornell, Sara worked primarily on examining the genetic architecture underlying heterosis of plant height and flowering time, as well as yield. She developed an extensive hybrid population using the inbred lines in the nested association mapping (NAM) population and managed the field evaluation for these hybrids in widespread yield trials in nine environments over two years. Sara has facilitated the collection of over 150,000 points of phenotypic data. She has gained experience and skill in a large number of statistical tools used to analyze genetic data: from the use of mixed models to generate best linear unbiased predictions (BLUPs) for phenotypes to control for environmental field effects, to joint linkage mapping for identification of QTL, to different strategies for association mapping and controlling for population structure. Furthermore, Sara has used genomic selection models to estimate marker effects across the genome for hybrid vigor to examine prediction accuracies.

I thankfully dedicate this dissertation to my parents, Lars-Åke and Eva Larsson, for their unconditional support of my pursuit of knowledge.

Tack pappa och mamma för att ni alltid finns där för stöd och uppmuntran.

ACKNOWLEDGMENTS

I would like to express my gratitude to the many individuals who have, in one way or another, contributed to this dissertation as well as all the significant empirical and analytical work that has been performed over the last five years.

A very special thanks to my advisor, **Dr. Edward S. Buckler,** for giving me the opportunity to join his lab as a summer intern six years ago. During these years in the Buckler Lab, I have gotten the chance to learn high-throughput phenotyping and genotyping technologies, and to work with large datasets and statistical tools among many other things. This exposure to a diverse group of people (whether they be plant biologists, computer programmers, or statisticians) has exposed me to a much wider breadth of knowledge than most. I am extremely grateful for the opportunity Dr. Buckler has provided me to broaden my professional experience and prepare me for future challenges.

I am very fortunate and particularly grateful to all of the scientists who I have had the chance to work with and who have so generously shared their wide experience and skills (from designing experiments to managing data and analyses). In particular, thank you to our field manager, **Nicholas Lepak,** for sharing his knowledge, resources, and assisting with everything field related; from seed storage, to planting, making tools for phenotyping, and harvest. Many collaborators from the Maize Diversity Project at Cornell and at other institutions provided significant phenotyping efforts contributing to the data necessary for this dissertation. I would like to thank all of you. Special thanks go to **Dr. James Holland** at North Carolina State University, **Dr. Sherry Flint-Garcia** and **Dr. Michael McMullen** at University of Missouri, **Dr.**

Mitchell Tuinstra at Purdue University, **Dr. Jode Edwards** at Iowa State University, and **Dr. Elhan Erzos** at Syngenta Seeds. Without these collaborators assisting in planting, managing, and harvesting the yield trial I would never have been able to perform experiments on the NAM hybrids. I would also like to thank my colleagues: **Kelly Swarts** and **Alberto Romero,** as well as **Drs. Jason Peiffer, Cinta Romay, Denise Costich, Feng Tian, Nengyi Zhang, Fei Lu**, and all the undergraduate interns who have assisted with seed counting and packing, planting, phenotyping, and harvesting over the years. I would like to thank the teams behind HapMap I and II, as well as the Genotyping by Sequencing (GBS) pipeline for generating large datasets of high quality genotypic data. In particular, **Robert Elshire, Drs. Sharon Mitchell, Michael Gore, Jer-Ming Chia, Jeffery Glaubitz, Qi Sun**, the members of the Institute of Genomic Diversity.

These projects have included large sample sizes and generated large amounts of phenotypic and genotypic data. I would like to thank the bioinformatics and statistical genetic groups at Cornell who have shared their experience and skills with me. **Dallas Kroon** for managing databases for seed storage and phenotypes. **Terry Casstevens** for his development of software tools. **Dr. Peter Bradbury** for his extensive help with tools for genetic mapping and SNP imputation. **Dr. Alexander Lipka** for his help and input on statistical analysis as well as always making the time to answer my million questions about how to code in R. **Drs. Zhiwu Zhang** and **Jeffery Endelman** for our discussions regarding statistics which helped in my understanding and analysis of the genomic data.

I am particularly thankful to my supporting advisors **Drs. Timothy Setter** and **Margaret Smith** for their knowledge in the fields of plant physiology and plant

breeding, respectively. I very much appreciate the continued input and instruction from my committee members.

Finally, I would like to thank **Sara Miller** for the help she has given me during the past years and all the time she spent editing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BLUE – Best Linear Unbiased Estimation

BLUP – Best Linear Unbiased Prediction

bm1- brown midrib1

bm3- brown midrib3

bp – base pair

CAD - cinnamyl alcohol dehydrogenease

CesA – cellulose synthase

Chr – Chromosome

cM – centi Morgan

COMT- caffeic O-methyl transferase

D2A – days to anthesis

D2S – days to silk

d8 – Dwarf8

FDR – False Discovery Rate

GAI – Gibberellic Acid Insensitive

GBS – Genotyping by Sequencing

GDD – Growing Degree Day

GEBV – Genomic Breeding Values

GLM – General Linear Model

GWAS – Genome Wide Association Study

HapMap – Haplotype Map

HMM – Hidden Markov Model

Indel – insertion/deletion

k – subpopulations determined by software such as STRUCTURE

K – Kinship matrix

kb – kilobases

LD – Linkage Disequilibrium

LSmeans – Leased Squares means

MAF – Minor Allele Frequency

MAS – Marker Assisted Selection

Mb – Mega bases

Mg/ha – mega grams / hectare

MITE – Miniature Inverted-repeat Transposable Element

MLM – Mixed Linear Model

NAM – Nested Association Mapping

NCBI – National Center for Biotechnology Information

NIR – Near Infra Red

NSS – Non-Stiff Stalk

PCA – Principle Component Analysis

PVP - Plant Variety Protection

Q – Population structure

QTL – Quantitative Trait Locus

RefGen_v2 – Reference Genome version 2

RIL – Recombinant Inbred Line

RPR – Rind Puncture Resistance

SA – Structured Association

SH2-like domain – Src Homology 2-like domain

SNP - Single Nucleotide Polymorphism

SSS – Stiff Stalk Synthetic

tb1 - teosinte branched1

vgt1 - vegetative to generative transition 1

CHAPTER 1

INTRODUCTION

Maize (*Zea mays* L.), a New World crop, is the largest production crop in the world today, contributing billions of dollars to agriculture [1], and has outpaced Old World crops such as wheat and rice. Maize is currently grown on 97 million acres in the United States [2], and used for protein, oil, starch, animal feed, ethanol, and other bio-based products. In addition, maize is a model organism. The maize genome is one of the most complicated genomes, but maize also allows for controlled cross-pollinations and is easy to evaluate for a number of phenotypes. The species captures remarkable diversity; it is roughly 11-fold more diverse than humans [3]. This diversity can be observed on both the genotypic and phenotypic level. Maize varies in height from less than one meter to over six meters and ranges vastly in biomass and carbon allocation. In the same manner, maize kernels vary approximately 10-fold in their oil and protein content [4].

The domesticated maize being cultivated today retains about 57% of the diversity of its ancestor, teosinte [5], and 77% of the diversity of landraces [3]. During domestication and adaptation, maize went through genetic bottlenecks due to selection pressure. As a result linkage disequilibrium decay varies across the genome, as well as across subpopulations. The diversity of maize arises from millions of years of mutations and recombinations, and because of the high level of diversity it has responded very effectively to artificial selection over time. Maize has been able to adapt to very different environments, from northern Europe and Canada to the lowland

tropics and to the high Andes. This adaptation has been possible through heritable changes in flowering time, responses to photoperiod and temperature, and plant architecture [6].

Archaeological and genetic findings suggest that maize was domesticated in the Balsas River Basin of southwestern Mexico roughly 10,000 years ago, around the same time period as most major crop plants [7–9]. Domestication was driven by growers selecting preferred seeds based on size and other advantageous characteristics over multiple generations. This resulted in increased allele frequencies of favorable traits, as well as reducing overall diversity [10]. During the domestication process from its wild ancestor teosinte (*Zea mays* ssp. *parviglumis*) to maize (*Zea mays* ssp. *mays*), the plant architecture dramatically changed from widely branched with multiple inflorescences with dispersible seeds, to a single stalk plant with seed attached on one single inflorescence. Maize moved from Mexico via the American westward expansion across North American to the Corn Belt where it is the dominant crop today [11].

Maize lines have been under extensive improvement since domestication, and look far removed from their wild ancestors. Despite this dramatic change in phenotype, the effect on the genome wide diversity and mean haplotype length has been insignificant in improved lines. Modern breeding has been focused on optimizing traits that have been relatively easy to select for [6]. This suggests that significant progress can still be made by increasing diversity at loci that have not been under improvement in the past by introducing new germplasm into the breeding programs or breaking up already present haplotypes.

The majority of the germplasm used in today's breeding programs arises from one of the around 300 races of maize, and most of the inbred lines can be traced by pedigree back to two open pollinated populations, Reid Yellow Dent and Lancaster [12]. In addition, practically all commercial hybrids on the US market in the late 1980s were founded by six public lines or close relatives to them, namely C103, Mo17 and Oh43 belonging to the Lancaster variety, and B37, B73 and A632 from the Reid variety [13]. The level of diversity has continued to decrease as time goes on. A study from the mid-2000s shows that the US commercial germplasm is now based on a mere seven lines: B73, LH82, LH123, PH207, PH595, PHG39 and Mo17 [14].

At present, commercial maize in the US is almost entirely hybrids. The inbred-hybrid breeding system we currently use was introduced by Shull and East at the turn of the 20th Century [15–17]. At first, the performance of the hybrid combinations was not improved enough to outweigh the financial risk of switching from the traditional open pollinated populations. Hybrid maize was not widely used until Jones introduced the double cross hybrids [12]. Additionally, the use of hybrids was delayed by the Great Depression. Farmers were resistant to invest in the new hybrid seed that they could not save for the next season as they were accustomed to with the open pollinated populations. However, the New Deal gave farmers access to capital to buy the improved hybrid seed and thus the modern seed companies were born [11].

As maize is a hybrid crop, breeding programs are divided into the development of inbred lines and hybrid development. Genetic improvement is made in the inbred development by utilizing recombination to create new allelic combinations. The

combining ability between lines is evaluated in the hybrid development by crossing lines from opposite heterotic groups. The main heterotic groups in US maize breeding are the Stiff Stalk Synthetic (SSS) and Non-Stiff Stalk (NSS). Since the establishment of these heterotic groups, the genetic distances and allele frequency differences have increased between the two groups due to improvements by selection [12].

The main breeding objectives for modern maize improvement are machine harvestability, increasing yield, reducing negative effects of biotic and abiotic stresses, increasing overall plant health, and environmental adaptation [18]. The work in this dissertation focuses on flowering time, hybrid vigor, yield, and lodging. Flowering time is essential for maize since it is an outcrossing species and needs to have synchronized flowering with neighboring plants to secure fertilization, and it has to flower early enough in the season to reach full maturity before the first frost. Flowering time is a quantitative trait controlled by a large number of loci with small effects [19]. Despite its importance, only one locus has been cloned in maize so far [20].

Yield is the highest valued trait when it comes to maize breeding and it is relatively easy (although costly) to measure, but genetically very complex, most likely influenced in one way or another by all 40,000 genes in the maize genome. While much attention has been devoted to heterosis / hybrid vigor, it is not the only explanation for the increase in maize yield; in addition, there have been genetic improvements of inbred lines and significant changes in farm management. Scientists have researched hybrid vigor for over a hundred years and a number of hypotheses to explain it have been proposed, and it is still a very active field of research.

The last trait discussed in this dissertation is lodging. Lodging is a serious problem in corn production, resulting in lost yield. A large effort has been made to breed for lodging resistance. But few studies have been performed. It is a very difficult trait to evaluate and replicate under natural field conditions at full maturity, which makes genetic studies problematic.

Over time, a dramatic portion of maize improvements has been made by phenotypic selection. However, breeding strategies, breeding objectives, and growing environments are not constant. More recently, technologies to generate and handle large sets of genotypic data have becoming available. A number of statistical tools are developed to utilize the genotypic and phenotypic resources to make further improvements.

One of the first association studies in plants [21] was published over a decade ago and although association studies are not always straightforward, due to population structure and rare haplotypes [22,23], there are ample successful studies [24–32]. Another approach is linkage mapping using bi-parental populations rather then association populations and few markers, which results in mapping with lower resolution [19,33]. For breeding purposes, association and linkage mapping can be used for two main objectives: to understand the genetic architecture of a trait and to identify loci across the genome contributing to a trait of interest. The knowledge of these loci can be used to develop genetic markers linked to the underlying genes. Genotypes in a breeding program can be screened genetically and genome wide selection and predictions can be made even on the seeds before field evaluations are needed. A successful example of marker assisted selection (MAS) [34] is the work

done on increasing Vitamin A content in maize, where association mapping was used to identify genetic markers linked to genes in the pathway [28]. The marker information is now used in the HarvestPlus program in Africa to improve local germplasm for Vitamin A content (Personal communication T. Rocheford).

With the continued reduction in the cost of genotyping, field evaluation is, in many cases, becoming the more costly and limiting factor in breeding programs and genetic research. An alternative approach or complement to association and linkage mapping is genomic selection, using genome wide markers to predict a trait. Genomic selection was first applied in animal breeding [35] using high-density markers all treated as random effects across the genome. The advantages of genomic selection are that a training population is established and extensively genotyped and phenotyped to train the statistical models. The models are used to predict the phenotype on lines with only genotypic data. This allows shortening of the breeding cycle as well as the ability to make selections based on genotypes before lines are evaluated in the field [36,37].

Plant breeding has been the important practical application of genetics in the 20th Century. Now, with a radically expanded genetic toolset, we are able to focus on breeding on a whole different level and at an entirely different speed. No longer are we limited to large phenotypic variation and growing seasons. We can dissect and understand the underlying characteristics of traits controlled by large numbers of loci with small effects. We can mine the ever-expanding genotypic data sets for advantageous allelic combinations to improve crops. As the world's population increases, arable land shrinks, and the climate changes, the ability to rapidly improve crops will be vital.

REFERENCES

1. FAO (2010) FAO Statistical Databases.

2. National Agricultural Statistics Service (2012).

3. Tenaillon MI, Sawkins MC, Long a D, Gaut RL, Doebley JF, et al. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proceedings of the National Academy of Sciences of the United States of America 98: 9161–9166.

4. Bennetzen J, Hake S (2009) Handbook of Maize: It's Biology.

5. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, et al. (2005) The effects of artificial selection on the maize genome. Science (New York, NY) 308: 1310–1314.

6. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, et al. (2012) Comparative population genomics of maize domestication and improvement. Nature genetics 44: 808–811.

7. Piperno DR, Ranere AJ, Holst I, Iriarte J, Dickau R (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 of the United States of America 106: 5019–5024.

8. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez G J, Buckler E, et al. (2002) A single domestication for maize shown by multilocus microsatellite genotyping. Proceedings of the National Academy of Sciences of the United States of America 99: 6080–6084.

9. Heerwaarden JV (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proceedings of the National Academy of Sciences of the United States of America.

10. Wang RL, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. Nature 398: 236–239.

11. Troyer A (2009) Development of hybrid corn and the seed corn industry. Handbook of Maize II: 87–114.

12. Lee E, Tracy W (2009) Modern maize breeding. Handbook of Maize II.

13. Goodman MM (1990) Genetic and germplasm stocks worth conserving. Journal of Heredity 81: 11–16.

14. Mikel M A., Dudley JW (2006) Evolution of North American Dent Corn from Public to Proprietary Germplasm. Crop Science 46: 1193.

15. Shull G (1909) A pureline method of corn breeding. American Breeders' Association 5: 51–59.

16. Shull G (1908) The Composition of a Field of Maize. Journal of Heredity 4: 296–301.

17. East EM (1908 Inbreeding in corn. Conneticut Agric Exp Stn Rep: 419–428.

18. Troyer A (1996) Breeding widely adapted, popular maize hybrids. Euphytica.

19. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, et al. (2009) The genetic architecture of maize flowering time. Science (New York, NY) 325: 714–718.

20. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, et al. (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proceedings of the National Academy of Sciences of the United States of America 104: 11376–11381.

21. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, et al. (2001) *Dwarf8* polymorphisms associate with variation in flowering time. Nature genetics 28: 286–289.

22. Platt A, Vilhjálmsson BJ, Nordborg M (2010) Conditions under which genome-wide association studies will be positively misleading. Genetics 186: 1045–1052.

23. Larsson SJ, Lipka AE, Buckler ES (in press) Lessons from *dwarf8* on the strengths and weaknesses of structured association mapping. PLoS genetics.

24. Wilson LM, Whitt SR, Ibanez AM, Rocheford TR, Goodman MM, et al. (2004) Dissection of Maize Kernel Composition and Starch Production by Candidate Gene Association. The Plant Cell 16: 2719–2733.

25. Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172: 1165–1177.

26. Beló A, Zheng P, Luck S, Shen B, Meyer DJ, et al. (2008) Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. Molecular genetics and genomics : MGG 279: 1–10.

27. González-Martínez SC, Huber D, Ersoz E, Davis JM, Neale DB (2008) Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. Heredity 101: 19–26.

28. Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, et al. (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science (New York, NY) 319: 330–333.

29. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, et al. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465: 627–631.

30. Yan J, Kandianis CB, Harjes CE, Bai L, Kim E-H, et al. (2010) Rare genetic variation at *Zea mays crtRB1* increases beta-carotene in maize grain. Nature genetics 42: 322–327.

31. Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, et al. (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nature genetics 43: 163–168.

32. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, et al. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature genetics 43: 159–162.

33. Peiffer J, et al. (in prep.) The Genetic Architecture of Plant Height. PloS one.

34. Xu Y, Crouch JH (2008) Marker-Assisted Selection in Plant Breeding: From Publications to Practice. Crop Science 48: 391.

35. Meuwissen T, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 49.

36. Heffner EL, Sorrells ME, Jannink J (2009) Genomic Selection for Crop Improvement. Crop Science 49: 1–12.

37. Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics 9: 166–177.

CHAPTER 2

LESSONS FROM *DWARF8* ON THE STRENGTHS AND WEAKNESSES OF STRUCTURED ASSOCIATION MAPPING

ABSTRACT

The strengths of association mapping lie in its resolution and allelic richness, but spurious associations arising from historical relationships and selection patterns need to be accounted for in statistical analyses. Here we reanalyze one of the first generation structured association mapping studies of the *Dwarf8* (*d8*) locus with flowering time in maize using the full range of new mapping populations, statistical approaches, and haplotype maps. Because this trait was highly correlated with population structure, we found that basic structured association methods overestimate phenotypic effects in the region, while mixed model approaches perform substantially better. Combined with analysis of the maize nested association mapping population (a multi-family crossing design), it is concluded that most, if not all, of the QTL effects at the general location of the *d8* locus are from rare extended haplotypes that include other linked QTLs and that *d8* is unlikely to be involved in controlling flowering time in maize. Previous independent studies have shown evidence for selection at the *d8* locus. Based on the evidence of population bottleneck, selection patterns, and haplotype structure observed in the region, we suggest that multiple traits may be strongly correlated with population structure and that selection on these traits has influenced segregation patterns in the region. Overall, this study provides insight into

how modern association and linkage mapping, combined with haplotype analysis, can produce results that are more robust.

AUTHOR SUMMARY

Eleven years ago, association mapping was a cutting-edge tool used to identify regions of a genome associated with phenotypic variation. One of the first association studies performed in plants was reported in Thornsberry, et al. (2001). Since then, researchers continued to develop new and improved genotyping, phenotyping, and statistical methods to examine the relationship between genotype and phenotype. Reanalysis of the old data for the *d8* locus and flowering time, as well as new and improved data sets, gives us a unique opportunity to examine the strengths and weaknesses of association studies. These new analyses reveal that the results reported in 2001 significantly overestimated the association between genotype and phenotype, in particular the estimated effect size. The key issues with the Thornsberry et al. (2001) study were lack of control for population structure and relatedness between individuals, as well as a potential confounding between the phenotype and the population structure examined. The new analysis demonstrates a marginal association between *d8* and flowering time, and a minimal effect (if any).

INTRODUCTION

Association mapping, which was developed as a necessity for large-scale human studies, is commonly used in conjunction with family (linkage) mapping in plant and animal genetic studies. The application of association mapping for plants was originally assessed in Thornsberry J.M. (2001) [1] with Buckler as senior author. It was concluded that association mapping offers higher resolution than linkage mapping due to quicker linkage disequilibrium (LD) decay, that structured association mapping is crucial for controlling false positives arising from population structure, and that *Dwarf8* (*d8*) (RefGen_v2 position: Chr. 1; 266,094,769-266,097,836 bp) is associated with flowering time. This initial study has been cited extensively, and has been the basis of several reanalyses of *d8*. New data and statistical tools give us the opportunity to reevaluate this locus. Results show that the *d8* associations reported by Thornsberry et al. (2001) are likely false positives (i.e., spurious associations), which resulted from insufficient correction of population structure. Indeed, the application of association mapping to animal and plant studies has been very successful, culminating in many important findings [2–10]. In this light, Thornsberry et al (2001) was a pioneering study, which has attracted a lot of interest to the area and led to more studies and the development of techniques to control for population structure and familial relatedness. When the phenotype is strongly correlated with population structure (e.g., flowering time), it is often difficult to obtain statistically significant results when the models used include covariates accounting for population structure. This leads to uncertainty when determining which associated sites are causative. Thus, linkage mapping is a

valuable complementary approach in these situations, and in maize, large-scale connected mapping populations issued from diverse founders have been developed [11,12] in order to conduct joint linkage-association analyses [10,13,14].

A major issue with association studies is false positives. In particular, indirect associations that are not causal will not be eliminated by increasing the sample size or the number of markers [15]. The main sources of such false positives are linkage between causal and noncausal sites, more than one causal site, and epistasis. These indirect associations are not randomly distributed throughout the genome and are less common than false positives arising from population structure. This makes them more difficult to control for than false positives arising from population structure.

The identification of a statistically significant association between a genotypic marker and a trait is considered to be proof of linkage between the phenotype and a casual site. This assumption is true for random mating populations with fast LD decay [16]. However, it is important to consider that population structure is typically present in association panels and it has an impact on the results. Population structure exists among all species in forms such as colonies, ethnic groups, and other subdivisions based on selection or geography. Typically, population structure leads to spurious associations between markers and the trait [17].

The ability to account for population structure in a given data set is influenced by the population size, the number of markers, the level of admixture, and the divergence in allele frequency between the subpopulations [18]. One commonly used method for controlling population structure is structured association (SA), which relies on randomly selected markers from the genome to estimate population structure. This

estimate is then incorporated into the association analysis [16,18,19]. Another methodology for controlling population structure is to conduct a principal component analysis (PCA) [20,21]. This approach summarizes the variation observed across all markers into a smaller number of underlying component variables. One interpretation of these principal components relates them to separate, unobserved subpopulations from which the individuals in the data set originate. The loadings (i.e., coefficient values) of the individuals for each principal component describe their relationship to the subpopulations. Both SA and PCA are limited to correcting for spurious associations by clustering on a global level of genetic variation. Thereby, they do not adequately capture the relatedness between individuals.

Correcting for population structure is not sufficient to eliminate all false positives. Therefore, the unified mixed linear model (MLM; also called the Q+K model) [22] was developed to further reduce the false positive rate by controlling for both population structure and cryptic familial relatedness. This approach uses a mixed model framework that has traditionally been used by animal geneticists [23,24]. Specifically, covariates accounting for population structure are included as fixed effects (Q), and the individuals in the association panel are included as random effects. A kinship matrix (K) is calculated to estimate the variance-covariance between the individuals. Typically, the covariates used in the unified MLM are either principal components of the markers or covariates from SA approaches (e.g., STRUCTURE [17]). The advantages of the MLM are that it crosses the boundary between family-based and population-based samples. However, not all associations that are eliminated

will be false. If a polymorphism is perfectly correlated with population structure, it is not possible to differentiate between true and false positives.

The initial study by Thornsberry et al. (2001) identified nine polymorphisms within *d8* [25] that were associated with variation in flowering time in an association panel consisting of 92 diverse inbred lines. The most significant site was an 18 bp deletion (RefGen_v2 position: Chr. 1; 266,094,529 bp) in the promoter region. A 6 bp indel (RefGen_v2 position: Chr. 1; 266,095,483) was also identified. This allele is over-represented in Northern Flint lines and is located near a Src Homology 2-like domain, which is an important binding domain within this class of transcription factors. The initial association analysis was performed using logistic regression analyses, accounting for population structure. Population structure was estimated as a modification of SA using STRUCTURE software [18] with k=3.

Using a general linear model (GLM) without population structure, Andersen et al. (2005) obtained similar results for six of the nine *d8* polymorphisms identified by Thornsberry et al. (2001). However, when including population structure in the model, (using STRUCTURE with both k=2 and k=3 subpopulations), it was found that the association results were overestimated. Each subpopulation was also analyzed separately, and a spurious association was still detectable [26].

Camus-Kulandaivelu et al. (2006) examined the association between *d8* and flowering time using a panel of 375 inbred lines (including the 92 from the initial study) as well as a panel consisting of 275 traditional landraces from American and European origins [27]. Population structure was estimated using STRUCTURE, and association analysis was performed using both GLM and logistic regression. Their

analysis revealed that the 6 bp indel at 266,095,483 bp (identified in Thornsberry et al., 2001) was spuriously associated with flowering time when covariates accounting for population structure were not included. In contrast, no association between *d8* and flowering time was detected in the inbred panel when accounting for population structure. However, this spurious association was still detectable in the traditional landraces panel, including Andean material that has no relationship to the Northern Flint material.

The *d8* gene produces a signaling factor involved in the gibberellin pathway. Gibberellins are types of endogenous plant growth regulators [28]. Maize *d8* and wheat *Rht-B1/Rht-D1* have been shown to be orthologous of the *GAI* gene [25]. Mutants of *d8* have severe height phenotypes due to alterations of the DELLA domain. In maize, these are dominant, gain-of-function mutations, suggesting that *d8* is a negative regulator. Conversely, recessive mutants of the *GAI* gene in *Arabidopsis* result in loss-of-function, specifically in polypeptides truncated upstream of the SH2-like domain. As a consequence, the gene product does not function as a negative regulator, resulting in normal height phenotypes [1].

Two evolutionary processes have likely impacted the *d8* locus. First, the associated allele, specifically the 6 bp indel reported in Thornsberry et al. (2001), is related with Northern Flint maize. Maize originated from southern Mexico, where there are long growing seasons and high temperatures. As maize agriculture expanded from Mexico through the Southwestern United States to the Eastern United States (with its shorter growing season and lower temperatures), a severe bottleneck occurred in maize diversity, resulting in the Northern Flint subpopulation [29]. The bottleneck

created extensive long range LD in this subpopulation. Northern Flints were substantially isolated from all other maize subpopulations [29] until the introduction of the Southern Dents in the 1600s [30].

Additionally, the *d8* locus is located only 347,057 bp from the *tb1* (*teosinte branched1*) locus (RefGen_v2 position: Chr. 1; 265,745,979-265,747,712 bp), which is one of the key genes involved in maize domestication [31]. The *tb1* locus lost much of its diversity during the domestication process [31,32]. The original *d8* study [1] identified evidence of purifying selection with substantial diversity loss; however, there was little LD identified in the region between *d8* and *tb1*. Although unconfirmed, some Northern Flint allied germplasm (e.g. sweet corn, P39) have a morphology that looks like the undomesticated *tb1* phenotype. It is likely that the region around *d8* and *tb1* has been through a bottleneck with multiple selective sweeps, resulting in complex extended haplotypes.

Most of the loci controlling flowering time in maize have been identified through QTL studies. Of these, only *d8* and *vegetative to generative transition 1* (*vgt1*) have been confirmed with association and fine mapping [33]. Located on chromosome 8, *vgt1* is arguably the most important flowering time locus in maize. It contains an APETALA2-like gene, *ZmRap2.7*, which is controlled by an enhancer region about 70 kb upstream [33]. The association between *vgt1* and flowering time is supported by a study conducted in the maize nested association mapping (NAM) population, where a major QTL was identified in this region [11]. This study also detected an allelic series at this QTL, suggesting that more than one causative allele is present. One of these alleles is from northern germplasm and is in linkage with a MITE whose association

with early flowering time was confirmed in the NAM population [11]. Although the lack of the *vgt1* early flowering allele did not completely explain the late flowering time, a SNP identified in the *ZmRap2.7* gene showed association with the late flowering effect [11].

An association study by Ducrocq et al. (2008) reported *P*-values for *vgt1* association several magnitudes lower that those obtained by Salvi et al. (2007). Both studies accounted for population structure. Compared to Salvi et al. (2007), Ducrocq et al. (2008) used a more genetically diverse and larger association panel, including a higher number of lines from Northern Flint and European germplasm [34]. In the case of *d8*, the association between the site and the trait becomes less significant, and even undetectable, when increasing the number of lines examined. This supports no association between the 6 bp indel in *d8* and flowering time in maize. Including *d8* in the model when performing association mapping for flowering time does not change the result for the SNPs in *vgt1* [34]. This indicates that there is no interaction between the two loci.

The purpose of this study was to reanalyze the work of Thornsberry et al. (2001) utilizing some of the latest association mapping methodologies and data sets. This study compared association results from various statistical approaches using a maize diversity panel and the NAM population [11,12]. Single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) from recent genotyping efforts (e.g., HapMap sequencing from Gore et al. 2009 [35] and Chia et al. 2012 [36]) were used to evaluate these various approaches and the *d8* association.

RESULTS

Association Study

The results from the Thornsberry et al. (2001) study showed significant association at both the 18 bp deletion (266,094,829 bp) in the promoter region and the 6 bp indel (266,095,483 bp). Our reanalysis of the two sites using the Q model and a significantly larger association panel (consisting of 282 lines) resulted in less significant associations at both loci (Table 2.1). By increasing the number of lines we are able to obtain a larger sample size within each of the subpopulations and thus, more accurately estimate the underlying population structure (i.e., Q).

Sampling has a larger effect on some sites than others. The 6 bp indel is more significantly associated with flowering time in the smaller population (92 lines) than it is in the 282 association panel analyzed with MLM (K model) without controlling for population structure, but controlling for familial relatedness. The site is, in fact, carried by Northern Flint lines, which are underrepresented in the smaller population. The results for the 282 association panel suggest that the GLM (Q) approach overestimates the association. In contrast, the MLM (Q+K) approach, which accounts for both population structure and relatedness between individuals, gives a moderately significant association between the 6 bp indel (*P*-value = 0.0127) and flowering time variation (Table 2.1).

Table 2.1. Association between polymorphisms at the *d8* locus and variation in flowering time in the 92 and 282 association panel, and association between polymorphisms in the region between *d8* and *tb1* (*d8*/*tb1*) and variation in flowering time in the 282 line association panel.

	SNP	Population	Naïve Model[a]	K[b]	Q[c]	Q+K[d]
			p-value	p-value	p-value	p-value
d8	18 bp	282	0.0775	0.3996	0.8422	0.3676
		92	7.95×10^{-4}**	0.1585	0.0021*	0.0847
	3 bp	282	2.32×10^{-5}**	0.5759	0.0077*	0.8697
		92	0.1649	0.1411	0.3143	0.2403
	6 bp	282	9.70×10^{-6}**	0.0142*	0.0018*	0.0127*
		92	2.51×10^{-5}**	3.23×10^{-4}**	0.0012*	0.0017*
d8/tb1	36	282	0.0017*	0.309	0.005*	0.062
	72	282	0.003*	0.3123	0.0259*	0.062
	174	282	5.38×10^{-5}**	0.1354	0.0084*	0.046*
	287	282	5.38×10^{-5}**	0.1357	0.0084*	0.046*

[a] A general linear model not controlling for population structure.
[b] A mixed linear model controlling for kinship but not population structure.
[c] A general linear model controlling for population structure *(k*=5).
[d] A mixed linear model controlling for both population structure *(k*=5) and kinship.

The proportion of the genetic variation explained by the different models varies significantly. In this study, the best models are the Q+K and K models (the latter being a MLM that only includes familial relatedness between individuals as random effects) because they explain the highest amount of the genetic variance (Table 2.2). The reason for the minimal difference between the two models is that K most likely controls for the majority of the relatedness between individuals.

This study confirms the weak association between the 6 bp indel in *d8* and flowering time analyzed using both GLM and MLM approaches (Table 2.1). However, the association is not as significant as previously reported by Thornsberry et al. in 2001. Additionally, the GLM and MLM analyses of the 282 association panel imply there is no association between the 18 bp deletion in *d8* and flowering time (Table 2.1). The initial study by Thornsberry et al. (2001) found this site to be the

most significant. Our results from the Q+K and K models yielded a more significant *P*-value for the 92 association panel than the 282 association panel.

We also sequenced a 3 bp indel (266,097,198 bp), which is present in tropical late-flowering lines when we examined sequences available at NCBI. However, new genotypic data for the 282 association panel suggest that there is no association between this site and variation in flowering time in maize (Table 2.1).

Table 2.2. Genetic variance explained by respective model used for the association study.

Model	SNP	R^2 Model
Naïve	none	0.000
	6 bp	0.084
	18 bp	0.022
Q	none	0.446
	6 bp	0.461
	18 bp	0.456
K	none	0.898
	6 bp	0.916
	18 bp	0.917
Q+K	none	0.898
	6 bp	0.914
	18 bp	0.914

Our study confirms the results presented by Camus-Kulandaivelu et al. (2008) [37], that there are regions between *d8* and *tb1* associated with variation in flowering time (Table 2.1) (Supplemental Figures 2.3 and 2.4). However, these sites are moderately significant at α=0.05 when using the K and Q+K models. Association mapping of *d8* on other traits results in a number of weak associations with other traits, in addition to flowering time (e.g., plant height, ear height, and node number) (Table 2.3). All the associations are in the same range of significance as flowering time. No clear pattern can be observed between correlation among traits except for what can be expected (

e.g., the high correlation between days to silk and days to anthesis) (Figure 2.1). Collectively, these results undermine the conclusion that *d8* is of more importance for flowering time than any of the other traits.

Table 2.3. Results from association study between polymorphisms within *d8* and a range of traits using MLM (Q+K).

	P-value		
	6 bp	18 bp	3 bp
Days to Silk	0.0013		
Days to Tassel	0.0014		
Number of Ears		0.0469	
Number of Nodes Tassel to Ear		0.0473	0.0083
Middle Leaf Angle	0.0072		
Number of Nodes Ear to Roots	0.0185		
Ear Height	0.0207		
Cob Color	0.0343		
Plant Height	0.0478		
Number of Brace Root Nodes			0.0035
Tassel Branch Length			0.0171

	TasselBranch Length	#OfNodes RootToEar	#OfEars	#OFNodes TasselToEar	#OfBraceRoots	PlantHeight	EarHeight	Middel LeafAngel	D2A
TasselBranch Length									
#OfNodes RootToEar	0.30								
#OfEars	0.11	0.29							
#OFNodes TasselToEar	0.20	0.55	0.11						
#OfBraceRoots	0.34	0.65	0.14	0.53					
PlantHeight	0.32	0.68	0.22	0.51	0.54				
EarHeight	0.36	0.83	0.29	0.57	0.57	0.88			
Middel LeafAngel	0.21	0.47	0.06	0.39	0.37	0.46	0.46		
D2A	0.38	0.79	0.22	0.62	0.61	0.62	0.73	0.47	
D2S	0.38	0.78	0.20	0.63	0.61	0.60	0.71	0.44	0.99

Figure 2.1. Pearson correlation coefficient between multiple traits. No clear pattern can be observed between correlation between traits and its association to the 6 bp and 18 bp deletions in *d8*.

From a genome-wide perspective, there were a large number of sites with a similar degree of association (from the MLM approach) with flowering time as *d8* (Figure

2.2A). The contrasting results from the various models fitted at the SNPs in the genomic regions surrounding *d8* and *tb1* are illustrated in Figures 2B and 2C. In particular, the GLM model overestimated the significance of the results in comparison to the Q+K and K models. GWAS of flowering time detected SNPs within *d8* that have a weak statistically significant association at α=0.05 (Figure 2C).

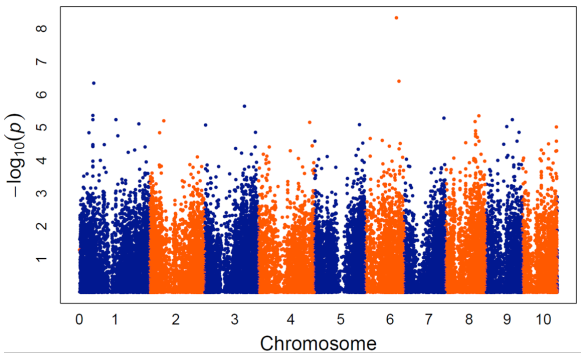


Figure 2.2A. Genome wide association results for flowering time (days to silking) in the 282 association panel using genotyping by sequencing (GBS) and 55k SNPs. The Q+K mixed linear model was fitted at each SNP to account for population structure (Q) and kinship (K).

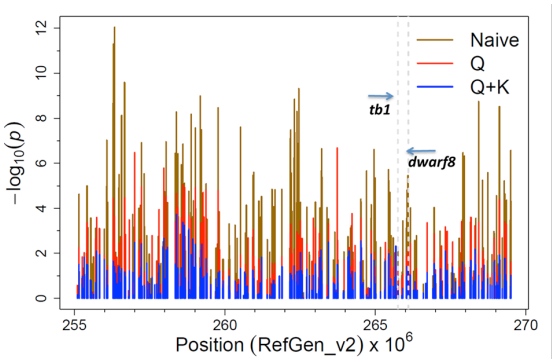


Figure 2.2B. GWAS results for flowering time (days to silking) using three models in the chromosomal region surrounding *tb1* (Chr. 1; 265,745,979-265,747,712 bp) and *d8* (Chr. 1; 266,094,769-266,097,836 bp). All GBS and 55K SNPs between 255 Mb and 270 Mb on Chr. 1 are included in the figure. Brown lines indicate results from naïve model, red lines indicate results from Q model, and blue lines indicate results from Q+K model.

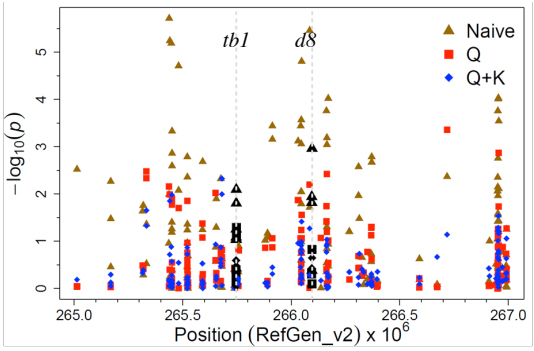


Figure 2.2C. GWAS results for flowering time (days to silking) using three models in the chromosomal region surrounding *tb1* (Chr. 1; 265,745,979-265,747,712 bp) and *d8* (Chr. 1; 266,094,769-266,097,836 bp). All GBS and 55K SNPs between 265 Mb and 267 Mb on Chr. 1 are included in the figure. Black markers on the right are significant SNPs located within *d8*. Black markers on the left are significant SNPs located within *tb1*. Triangles indicate results from naïve model, squares indicate results from Q model, and diamonds indicate results from Q+K model.

Linkage Mapping

Linkage mapping of flowering time in the NAM population detected a number of QTL. A small QTL (*P*-value = 0.0127) colocalized with *d8* (RefGen_v2 position: Chr. 1; 269,321,476-269,322,794 bp), supporting the association identified by association mapping (Figure 2.3). In the initial study by Thornsberry et al. (2001), the effect of *d8* was estimated to be between 7-10 days. The *d8* polymorphism should be in three of the mapping families, and modest effects are seen in the right direction for all three, but the estimated effect is always less than half a day.

Additionally, many of the subfamilies appear to have other QTL along this section of chromosome 1 (RefGen_v2 position: Chr. 1; 231,701,106-231,703,173 bp and Chr. 1; 286,977,415-287,063,457 bp), but the favored positions are millions of base pairs away. It is quite possible that the mapping position of these joint linkage QTL could be synthetic, but there is little to no support for a QTL in this exact region.

A GWAS in the NAM population for flowering time using 26.5 million segregating SNPs was performed [35,36]. This approach in the NAM population offers in-depth power and resolution because it utilizes both historic and recent recombination. No significant sites were identified in the region of *d8* (Supplemental Figure 2.5). This supports the result that *d8* is not associated with flowering time.

Figure 2.3. Effect estimates in days for NAM subpopulations carrying QTL in the region of *d8*, *P*-value < 0.05. Light gray bar shows QTL effect estimate at marker position 116 (RefGen_v2 position: Chr. 1; 231,701,106-231,703,173 bp) (137.6 cM) on the NAM map. Dark gray bar shows QTL effect estimate at marker position 155 (181.3 cM). *d8* is located closest to marker 135 (RefGen_v2 position: Chr. 1; 269,321,476-269,322,794 bp), (RefGen_v2 position: Chr. 1; 286,977,415-287,063,457 bp), (162.2 cM). * indicates taxa with the 6bp deletion in *d8*.

Haplotype Structure

Hapmap data [35] [36] suggest extended haplotypes for Northern Flint lines in the region of *d8*. Data show modest F_{st} between temperate and tropical subpopulations. However, there could potentially be differences in diversity between these two groups and Northern Flint lines. Hapmap data are only available for a few Northern Flint lines, which limits these studies.

GBS SNPs were used to examine the range of LD decay within the different subpopulations (Northern Flint, stiff stalk, non-stiff stalk, and tropical) of the 282 association panel. Extended LD is observed for the Northern Flint lines compared to the other subpopulations. Likewise, the stiff stalk lines, which were only founded from

16 inbred lines, also show a pattern of extended haplotypes, although not as extreme as the Northern Flints (Figure 2.4). The extended haplotype pattern in the Northern Flints make it difficult to control for false positives and to identify the causative SNP using association mapping.

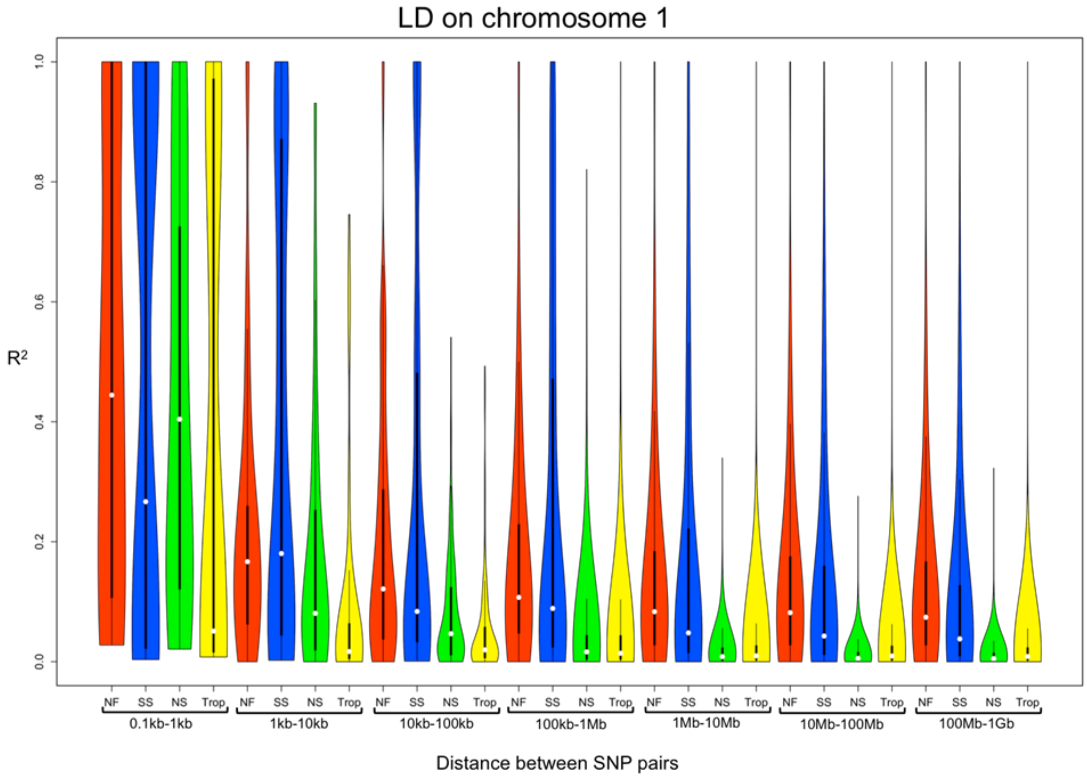


Figure 2.4. LD on chromosome 1 for the subpopulations, Northern Flint (red), stiff stalk (blue), non-stiff stalk (green), tropical (yellow), of the 282 association panel. White dot indicates the median R^2 for the bin. This graph shows that there is more extended LD in Northern Flint than in other subpopulations.

The 6 bp indel in *d8* is carried by Northern Flint lines. When we examine the LD between the 6 bp indel and the 13,815 high coverage GBS SNPs on chromosome 1 (Figure 2.5), an extended area around the 6 bp exhibits fairly high values of R^2. This is additional evidence that extended haplotypes exist in the Northern Flint lines in the *d8* region. In fact, there are two regions with high LD at 20 Mbp and 0.9 Mbp away, which contain previously identified domestication gene candidates (i.e.,

GRMZM2G034217 RefGen_v2 position: chr. 1 246,720,001 – 247,030,000, a mitochondrial transcription termination factor) [38]. In contrast, LD between the MITE in *vgt1* and the 7,539 GBS SNPs on chromosome 8 (Supplemental Figure 2.6) show sites with high R^2 values close to the position of the MITE, but LD decays much more rapidly. To test for two-way interaction between the 6 bp and 18 bp indels and the MITE, a series of mixed models including two-way interaction terms were fitted. The most significant interaction was between the 18 bp indel and the MITE (*P*-value 0.0418). However, this association is not likely to be statistically significant after controlling for the multiple testing problem across the entire genome.

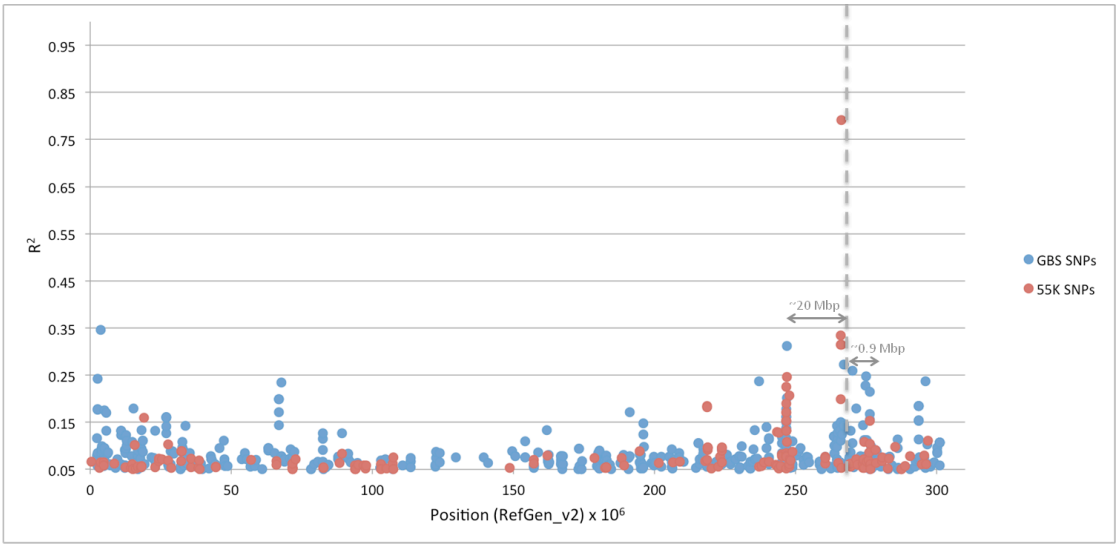


Figure 2.5. R^2 between the 6 bp indel in *d8* and all the other sites on chromosome 1. Blue dots indicate results from 13,815 GBS SNPs present in 200 or more of the 282 lines. Red dots indicate results from 7,695 55K SNPs present in 200 or more of the 282 lines.

DISCUSSION

Association Study

The results underscore the importance of properly accounting for population structure in association studies. The analysis in Thornsberry et al. (2001) divided their 92 association panel into three subpopulations. This subdivision did not fully account for population structure, and thus, the effect of the *d8* allele carried by Northern Flint lines was overestimated. In contrast, our study accounted for population structure using both $k = 3$ (stiff stalk, non-stiff stalk, and tropical) and $k = 5$ (stiff stalk, non-stiff stalk, tropical, sweet corn, and popcorn) subpopulations. This was important for sites such as the 6 bp indel within *d8*, which is present at a higher frequency in Northern Flint lines (which includes sweet corn), and this signature of population structure was unaccounted for when $k=3$ was used.

Of all the models tested, the Q+K model [22] was the most suitable approach for analyzing the 282 association panel because it controls for both population structure and cryptic familial relatedness. It is especially important to control for the latter with traits such as flowering time, which is highly correlated with population structure. Additionally, the Q+K model is beneficial for association studies because Q and K capture different types of long range LD [22,39].

In general mixed models sufficiently account for population structure and familial relatedness. In contrast, false positives arising from other sources, although rare, are typically unaccounted for in association studies. For example, spurious associations could arise from markers that are in long-range LD with causative

polymorphisms. These associations violate a basic assumption made in GWAS; namely, independence between markers. Additionally, causative polymorphisms for one trait may not necessarily be causal for another highly correlated trait (and, hence a spurious association), but will be statistically associated with both traits. Finally, when a trait is controlled by multiple loci in LD, it is likely that the site with the largest effect is an indirect association. One reason for this result arises from differing minor allele frequencies among the causal sites. All three of these types of false positives do not occur randomly across the genome and thus, they are more challenging to eliminate. Haplotype-based association studies is one approach for addressing many of these issues. Nevertheless, multiple sites, selection for multiple traits, and population structure result in spurious associations and these need to be accounted for when performing association studies.

Contrast with Linkage Mapping

Association mapping is limited when the trait analyzed is correlated with population structure. However, linkage mapping can overcome this problem by crossing individuals with known relatedness, spurious associations can be broken.

In this study, we were able to detect a small QTL at the general location of *d8*. However the favored QTL locations are on both sides of *d8*: RefGen_v2 position Chr. 1; 231,703,173 - 287,063,457. Association results suggest that the majority of the QTL effects detected around *d8* are from rare extended haplotypes that include other linked QTLs. Another possible explanation for the weakness of the QTLs detected is

the population sampled. The associated haplotype is present only in Northern Flint lines, which are underrepresented in the population.

Haplotype Structure

Flowering time is strongly correlated with population structure. Our study showed that *d8* had a very small effect on flowering time. One possible explanation for this result is that *d8* is associated with another trait that was selected along with flowering time (e.g., cold tolerance). This hypothesis is supported by the extended haplotype pattern observed in Northern Flint lines as well as the associations that are detected with traits like plant height and node number. Northern Flint lines are underrepresented in both the 282 association panel and the NAM population, and it is difficult to scrutinize associations with low allele frequencies. Northern Flint lines have been shown to be distinctive compared to other subpopulations, especially in regions like *d8* and *tb1,* which have been under selection pressure.

Consistent with the findings of previous studies on the *d8* locus, we observed a strong correlation between *d8* and population structure. Thus, the functional site of *d8* is not likely to be involved in flowering time. Indeed, *d8* and *tb1* are strong integrators in plant signals that are adjacent to each other on chromosome 1. However, our results demonstrated that these signals are a signature of population structure instead of true biological signals. The extended LD in the Northern Flint lines around *d8* supports the hypothesis that this gene is regulated in a similar manner as *tb1* and *vgt1*. For both *tb1* and *vgt1*, cis-acting regulatory sites located more than 50 kb from the actual genes have been shown to be the functional regions and not the genes themselves [31,33,40].

Signatures of selection on *d8* have been observed in teosinte [41]. Because apical dominance (*tb1*) and gibberellin signaling (*d8*) have both played key roles for domestication phenotypes, it is likely that the genomic region surrounding *d8* and *tb1* has been under selection since early maize domestication. Northern Flint lines differ from Corn Belt dent lines in a number of traits such as leaf angle, plant height, and cold tolerance. Thus, the long range LD block around *d8* could be a signature of selection from the development of Northern Flint lines that happens to be associated with one of these traits distinguishing Northern Flint from Corn Belt dent. Consequently, it has been possible to detect a weak association between flowering time and *d8* because of the correlation between flowering time and the Northern Flint specific traits due to population structure. Using this rationale, it may be possible to detect associations between the *d8* locus and phenotypes such as carbon allocation and harvest index, when considering the differences in the usage of Northern Flints (sweet corn and silage) and Corn Belt dent.

CONCLUSION

The basic *d8* associations identified in Thornsberry *et al.* (2001) have been replicated by other independent groups [26,27], but population structure has always remained a consistent issue. This reanalysis using SNPs within and near *d8* suggests that these associations are either incorrect or vastly overestimated. This work implemented more powerful statistical approaches, germplasm resources, and whole genome sequencing data, enabling a more thorough understanding of this locus.

This analysis underscores the importance of controlling for population structure. All three previously published studies on the *d8* locus illustrate how naïve association results overestimated effect sizes. In our study, we used the unified MLM to control for both population structure and relatedness between individuals, which are more accurate in effect estimation and give a truer level of significance. Even in species with rapid LD decay, like maize, it is possible to have subpopulations that can exhibit LD many orders of magnitude greater than the average length. This long range LD resulted in the extended haplotype lengths observed in Northern Flint lines for the genomic region surrounding *d8*. Northern Flint lines are underrepresented in the association panel, which makes it difficult to accurately account for the population structure of this subpopulation. Another issue is the strong correlations between traits. It is very likely that in the case of *d8* there has been selection for other correlated traits, such as cold tolerance. Because of the correlation between the population structure and flowering time, we can detect a weak association between flowering time and *d8,* but *d8* does not actually have an effect on time of flowering. Genes like *d8* have been targets of strong selection and, as such, are among the hardest to identify in GWAS and accurately estimate their effect size. NAM-like linkage populations with bi-parental crosses in a reference design to minimize population structure may be necessary for dissecting the most structured traits.

Although our results strongly suggest that the previously reported association between *d8* and flowering time is an artifact of population structure, further research on this complex locus is warranted. The long range LD present at *d8* for Northern Flint lines is a signature of selection, and it is important to determine the traits that are

regulated by this gene. By applying the appropriate statistical models, we have shown that flowering time is not one of these traits.

MATERIAL AND METHODS

Germplasm

The association panel consists of 282 diverse maize lines that have been previously described [42]. These lines can be subdivided into five major subpopulations, namely stiff stalk, non-stiff stalk, tropical or semitropical lines (related to the non-stiff stalk lines), sweet corn and popcorn. The association panel includes the 25 founder lines of the NAM population. The maize NAM population consists of 5,000 RILs (Recombinant Inbred Lines) derived from crossing B73 with 25 diverse maize inbred lines, and then selfing for 5 generations [43].

Phenotypic Data

Phenotypic data were collected from the NAM population and the 282 association panel, grown in eight environments including Ithaca, NY, Clayton, NC, Champaign, IL, and Colombia, MO, during the summers of 2006 and 2007. Flowering time was measured separately for female flowers (number of days-to-silk) and male flowers (days-to-anthesis) from the day of planting. The flowering date was defined as the day when the anthers or silk were visible on 50% of all plants within a row.

Sequencing Data

DNA sequence data were obtained for *d8* from Thornsberry et al. (2001), available at NCBI. Primers were designed for PCR amplification of gene fragments of interest from the 282 lines in the association panel.

Each PCR product was cleaned by treating the samples with Exonuclease (ExoI) and Shrimp Alkaline Phosphatase (SAP) and incubated at 37°C for 3 min followed by 80°C for 10 min. The samples were prepared for sequencing using a mixture with a total volume of 10 µl containing 0.7 µl forward primer, 0.7 µl reverse primer (5 pmol/µl), 0.5 µl Big Dye terminator, 1.7 µl 5x sequencing buffer, 7.1 µl distilled water and the PCR product. The thermal cycler was set on the following program: Initial denaturation at 96°C for 4 min, followed by 30 cycles at 96°C for 10 sec, 50°C for 5 sec and 60°C 4 min, with a final, incubation at 10°C. Sanger(3730XL) DNA sequencing was performed using an Applied Biosystems Automated 3730 DNA Analyzer. The software BioLign alignment and multiple contig editor with codon code phred-phrap analysis was used for alignment using consensus sequence contigs and sequence quality scores.

The alignments from NCBI were also used to reanalyze the results published in the initial study by Thornsberry et al. (2001).

To obtain sequence data for the region between *d8* and *tb1*, the same protocol was used as described above. However, primer sequences were obtained from Camus-Kulandaivelu et al. (2008)

Statistical Analysis

Field Spatial Correction

Best linear unbiased predictions (BLUPs) of the lines in the 282 association panel and the NAM population were the same as those reported in Buckler et al. (2009). These were obtained from a random effects model fitted in ASREML version 2.0 software [44] that accounts for spatial correlation and field effects.

Association Mapping – Candidate gene study

TASSEL (Trait Analysis by aSSociation, Evaluation, and Linkage) was used for data processing analysis [45], and results were confirmed by using SAS [46]. Association between polymorphisms and phenotypes were evaluated using General Linear Model (GLM) and Mixed Linear Model (MLM) by incorporating phenotypic and genotypic data, population structure (Q) and kinship matrix (K).

Population structure was predicted using a Bayesian approach that estimates the relationship between subpopulations by grouping genotypic correlations at unlinked markers within the population with the software STRUCTURE [18] as described in [42]. This approach uses the proportion of an individual's genome that originated from each subpopulation to estimate the genetic background matrix (Q).

In MLM, the familial relatedness between the individuals is taken into consideration through a kinship matrix. This model corrects for spurious associations arising from population structure and familial relatedness [47]. In this study we used marker-based kinship, which was determined on the basis of the definition that random pairs of inbreds are unrelated. Kinship was calculated using the software

package SPAGeDi [48]. It has been suggested that marker based kinship is more appropriate for association studies than kinship based on pedigree records [19,40]. The same set of markers was used to create the population structure and kinship matrix.

The GLM approach in this study for phenotype, **y**, is:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \quad (1)$$

Where, the $\mathbf{X\beta}$ term represents the fixed effects, including genotypes and population structure, Q, and $\boldsymbol{\varepsilon}$ is a vector of residual effects following a multivariate normal distribution with mean 0 and variance-covariance matrix $\sigma^2_\varepsilon I$. The naïve model is the same as GLM without the population structure effect.

The MLM approach in this study is the same model as used by Yu et al., 2006. Mixed model, for phenotype, **y**, is:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z\mu} + \boldsymbol{\varepsilon} \quad (2)$$

Where, the $\mathbf{X\beta}$ term represents the fixed effects, including genotypes and population structure, Q, and the $\mathbf{Z\mu}$ term represents random line effects, including the matrix of kinship coefficients, K, and vector of polygene background effects. $\boldsymbol{\varepsilon}$ is a vector of residual effects following a multivariate normal distribution with mean 0 and variance-covariance matrix $\sigma^2_\varepsilon I$.

Genome Wide Association Study

Genome-wide association studies (GWAS) were carried out in the 282 association panel using 51,741 SNPs obtained from the Illumina MaizeSNP50 BeadChip and

591,552 SNPs from the genotyping by sequencing (GBS) protocol [49]. Three different approaches that take into account varying degrees of population structure and familial relatedness were undertaken. The first approach, called the naïve approach, uses a model similar to the one presented in Equation (1), except that the Q matrix representing population structure is not included among the fixed effects. The next approach is the GLM approach, which uses the model in Equation (1), with the first five principal components (PCs) of the non-industry subset of the Illumina MaizeSNP50 BeadChip SNPs (34,368 SNPs) included as fixed effects to represent population structure. The final approach is the MLM approach, with the aforementioned first five PCs representing population structure, and a kinship matrix calculated from the non-industry subset of these SNPs for the variance-covariance matrix of the random line effects. This kinship matrix is calculated using the method of [50]. In each approach, these models are fitted to each SNP. After all SNPs with minor allele frequencies (MAFs) less than 0.05 are removed from the analysis, the Benjamini-Hochberg [51] procedure adjusts for the multiple testing problem by controlling the false discovery rate (FDR) at 0.05. This phase of the anlaysis was conducted using the genome association and prediction integrated tool (GAPIT) package in the R programming language [52].

Joint-Linkage Mapping

Joint Linkage Mapping of BLUPs for the phenotype across environments was performed using the proc GLMSelect in SAS, as described in Buckler et al. (2009)

[11]. BLUPs were calculated for each phenotype and used together with imputed genetic marker intervals and stepwise regression to identify QTLs. Missing marker data were imputed by utilizing genetic distance between missing and flanking markers. A permutation procedure was implemented to obtain empirical α=0.05 thresholds for including and excluding terms in the joint linkage model [53].

Linkage Disequilibrium

To calculate the linkage disequilibrium between the SNPs within *d8,* tested in this association study, against the rest of the genome the LD function SitebyAll in the TASSEL software was used [45]. For genotypic data, the Illumina MaizeSNP50 Beadchip was used, as well as 458k GBS (Genotyping by Sequencing) SNPs [49].

REFERENCES

1. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, et al. (2001) *Dwarf8* polymorphisms associate with variation in flowering time. Nature genetics 28: 286–289.

2. Wilson LM, Whitt SR, Ibanez AM, Rocheford TR, Goodman MM, et al. (2004) Dissection of Maize Kernel Composition and Starch Production by Candidate Gene Association. The Plant Cell 16: 2719–2733.

3. Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172: 1165–1177.

4. Beló A, Zheng P, Luck S, Shen B, Meyer DJ, et al. (2008) Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. Molecular genetics and genomics : MGG 279: 1–10.

5. González-Martínez SC, Huber D, Ersoz E, Davis JM, Neale DB (2008) Association genetics in Pinus taeda L. II. Carbon isotope discrimination. Heredity 101: 19–26.

6. Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, et al. (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science (New York, NY) 319: 330–333.

7. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, et al. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465: 627–631.

8. Yan J, Kandianis CB, Harjes CE, Bai L, Kim E-H, et al. (2010) Rare genetic variation at Zea mays crtRB1 increases beta-carotene in maize grain. Nature genetics 42: 322–327.

9. Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, et al. (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nature genetics 43: 163–168.

10. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, et al. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature genetics 43: 159–162.

11. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, et al. (2009) The genetic architecture of maize flowering time. Science (New York, NY) 325: 714–718.

12. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, et al. (2009) Genetic properties of the maize nested association mapping population. Science (New York, NY) 325: 737–740.

13. Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. Nature Reviews Genetics 11: 867–879.

14. Li H, Bradbury P, Ersoz E, Buckler ES, Wang J (2011) Joint QTL linkage mapping for multiple-cross mating design sharing one common parent. PloS one 6: e17573.

15. Platt A, Vilhjálmsson BJ, Nordborg M (2010) Conditions under which genome-wide association studies will be positively misleading. Genetics 186: 1045–1052.

16. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. American journal of human genetics 65: 220–228.

17. Pritchard JK, Stephens M, Rosenberg N a, Donnelly P (2000) Association mapping in structured populations. American journal of human genetics 67: 170–181.

18. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

19. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

20. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick N a, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38: 904–909.

21. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723.

22. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38: 203–208.

23. Henderson CR (1984) Applications of linear models in animal breeding. Guelph: University of Guelph.

24. George a W, Visscher PM, Haley CS (2000) Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. Genetics 156: 2081–2092.

25. Peng J, Richards DE, Hartley NM, Murphy GP, Devos KM, et al. (1999) "Green revolution" genes encode mutant gibberellin response modulators. Nature 400: 256–261.

26. Andersen JR, Schrag T, Melchinger AE, Zein I, Lübberstedt T (2005) Validation of *Dwarf8* polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.). TAG Theoretical and applied genetics 111: 206–217.

27. Camus-Kulandaivelu L, Veyrieras J-B, Madur D, Combes V, Fourmann M, et al. (2006) Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. Genetics 172: 2449–2463.

28. Harberd NP, King KE, Carol P, Cowling RJ, Peng J, et al. (1998) Gibberellin: inhibitor of an inhibitor of...? BioEssays : news and reviews in molecular, cellular and developmental biology 20: 1001–1008.

29. Doebley JF, Goodman MM, Stuber CW (1986) Exceptional Genetic Divergence of Northern Flint Corn. American Journal of Botany 73: 64–69.

30. Doebley J, Wendel JD, Smith JSC, Stuber CW, Goodman MM (1988) The origin of cornbelt maize: The isozyme evidence. Economic Botany 42: 120–131.

31. Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011) Identification of a functional transposon insertion in the maize domestication gene tb1. Nature genetics 43: 1160–1163.

32. Wang RL, Stec a, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. Nature 398: 236–239.

33. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, et al. (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proceedings of the National Academy of Sciences of the United States of America 104: 11376–11381.

34. Ducrocq S, Madur D, Veyrieras J-B, Camus-Kulandaivelu L, Kloiber-Maitz M, et al. (2008) Key impact of *Vgt1* on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics 178: 2433–2437.

35. Gore M a, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, et al. (2009) A first-generation haplotype map of maize. Science (New York, NY) 326: 1115–1117.

36. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, et al. (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nature genetics 44: 803–807.

37. Camus-Kulandaivelu L, Chevin L-M, Tollon-Cordet C, Charcosset A, Manicacci D, et al. (2008) Patterns of molecular evolution associated with two selective sweeps in the *Tb1-Dwarf8* region in maize. Genetics 180: 1107–1121.

38. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, et al. (2012) Comparative population genomics of maize domestication and improvement. Nature genetics 44: 808–811.

39. Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, et al. (2007) An Arabidopsis example of association mapping in structured samples. PLoS genetics 3: e4.

40. Clark RM, Linton E, Messing J, Doebley JF (2004) Pattern of diversity in the genomic region near the maize domestication gene *tb1*. Proceedings of the National Academy of Sciences of the United States of America 101: 700–707.

41. Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS (2004) Selection versus demography: a multilocus investigation of the domestication process in maize. Molecular biology and evolution 21: 1214–1225.

42. Flint-Garcia SA, Thuillet A-C, Yu J, Pressoir G, Romero SM, et al. (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. The Plant journal : for cell and molecular biology 44: 1054–1064.

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

44. Gilmour AR, Gogel BJ, Cullis BR, R T (2005) ASReml User Guide.

45. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, et al. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics (Oxford, England) 23: 2633–2635.

46. SAS Institute (2004) SAS/STAT user's guide. Version 9.2. SAS Inst., Cary, NC

47. Stich B, Möhring J, Piepho H-P, Heckenberger M, Buckler ES, et al. (2008) Comparison of mixed-model approaches for association mapping. Genetics 178: 1745–1754.

48. Hardy OJ, Vekemans X (2002) spagedi : a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2: 618–620.

49. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6: e19379.

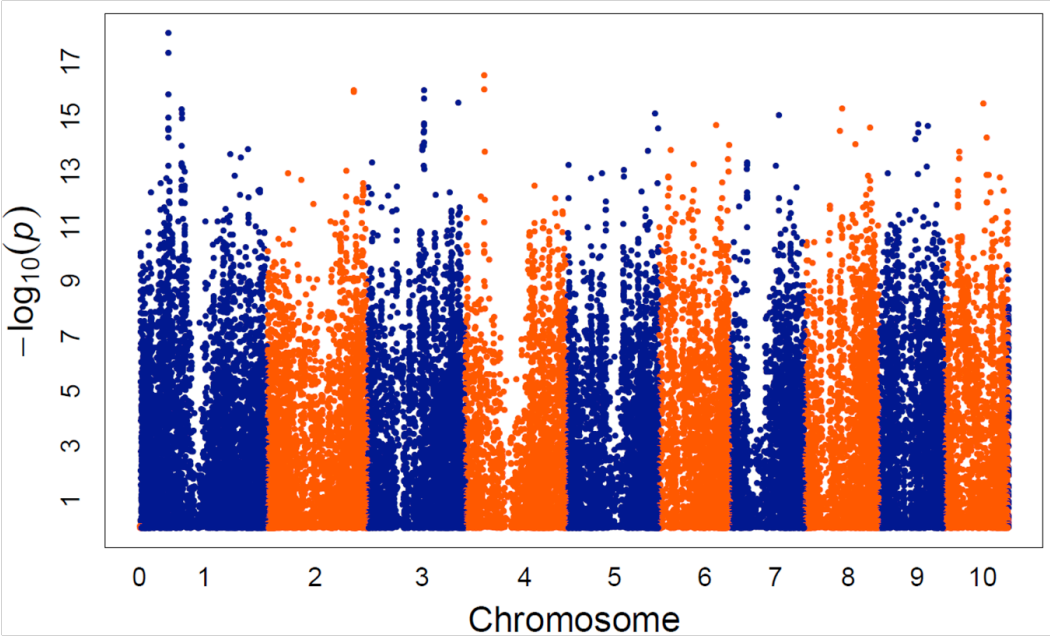
50. Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial Genetic Structure of a Tropical Understory Shrub, *Psychotria officinalis* (Rubiaceae). American Journal of Botany 82: 1420–1425.

51. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 57: 289–300.

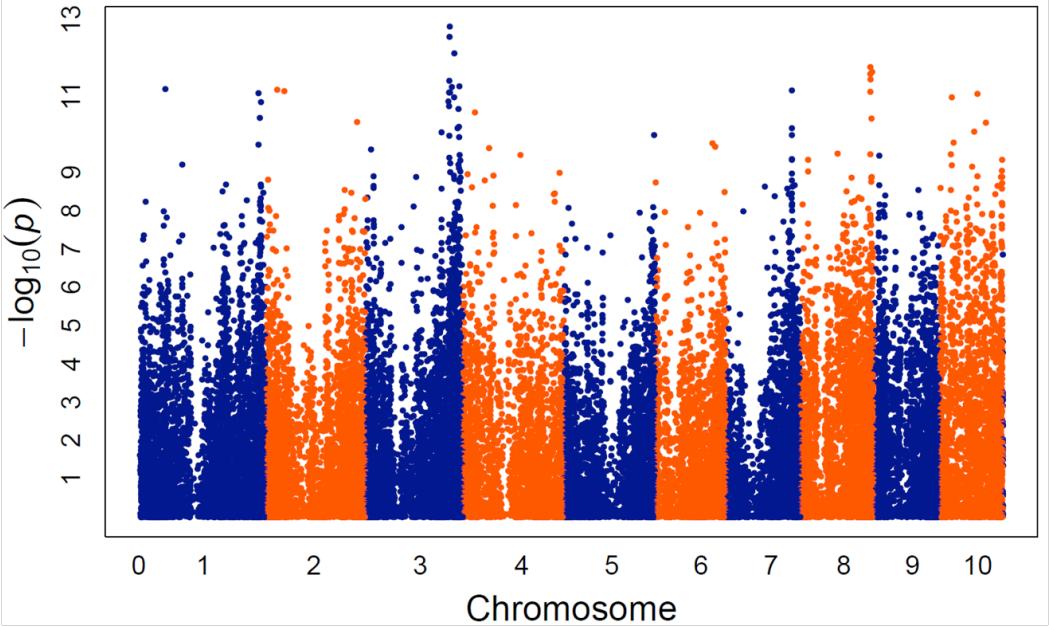
52. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, et al. (2012) GAPIT: Genome Association and Prediction Integrated Tool. Bioinformatics (Oxford, England): 1–2.

53. Churchill G a, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971.

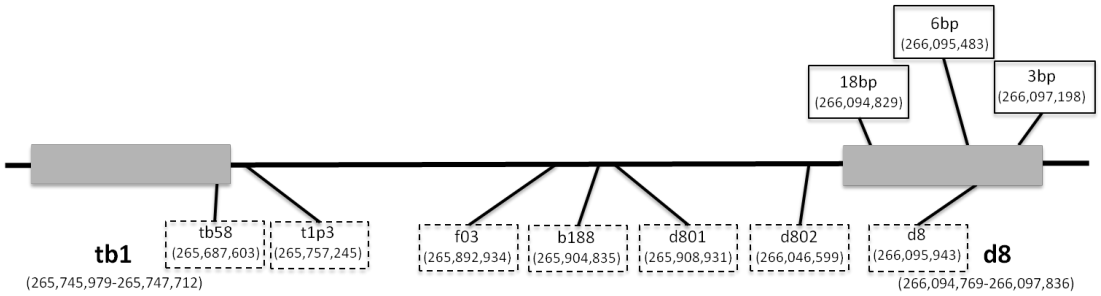
SUPPLEMENTAL MATERIAL



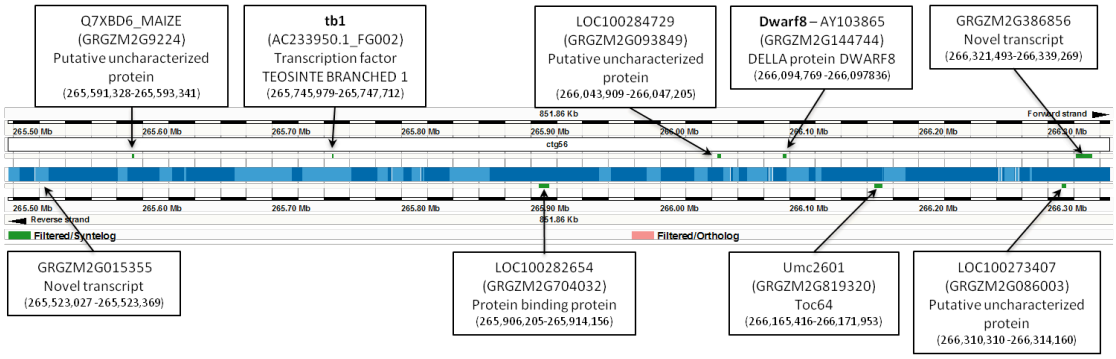
Supplemental Figure 2.1. Genome wide association results for flowering time (days to silking) in the 282 association panel using genotyping by sequencing (GBS) and 55k SNPs. The naïve model, which does not account for population structure, was fitted at each SNP.



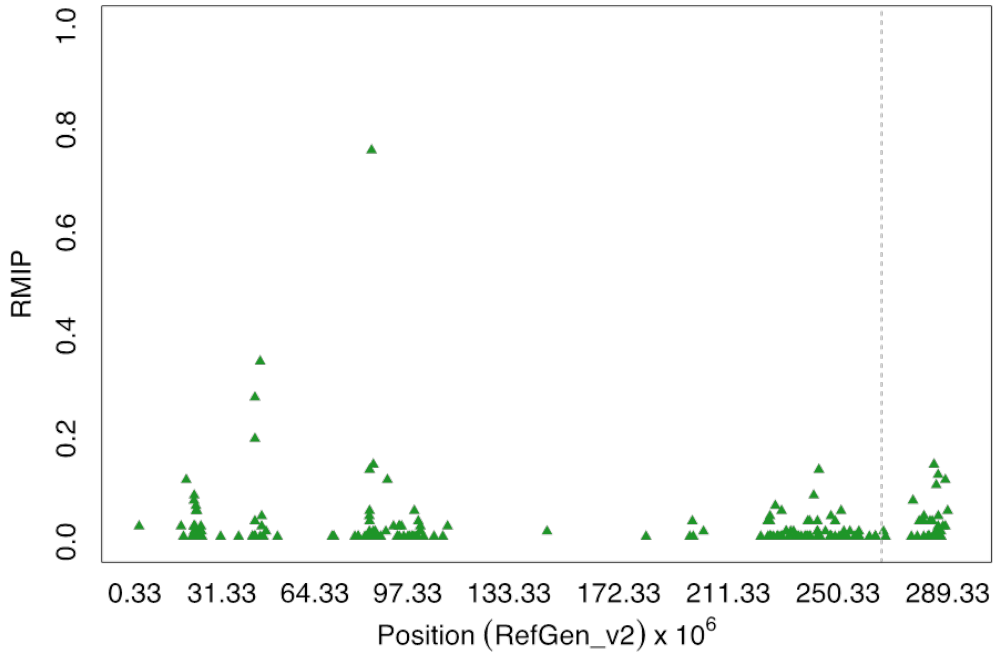
Supplemental Figure 2.2. Genome wide association results for flowering time (days to silking) in the 282 association panel using genotyping by sequencing (GBS) and 55k SNPs. The Q model was fitted at each SNP to account for population structure (Q).



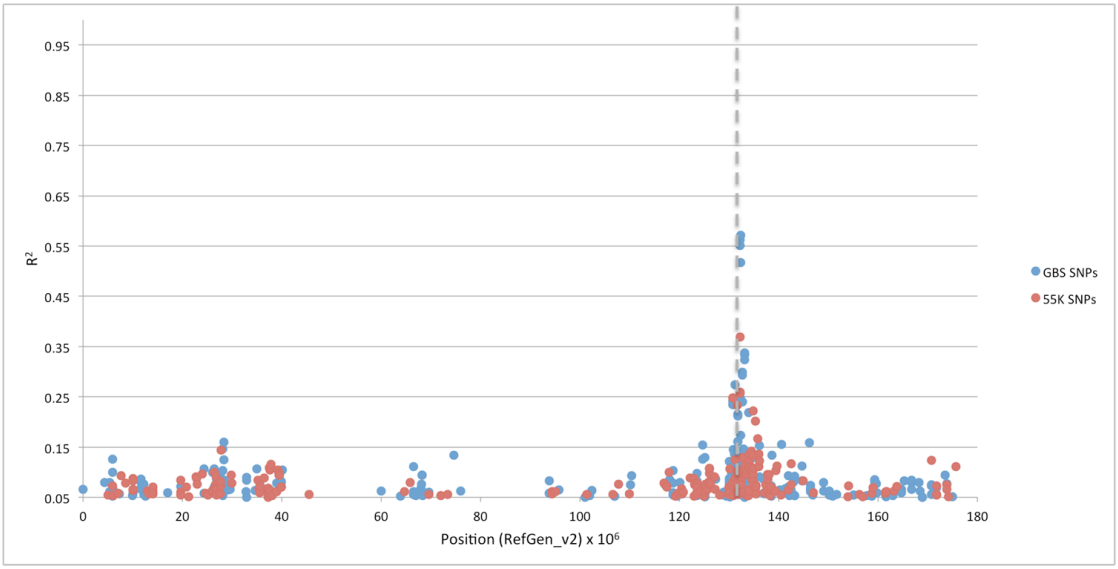
Supplemental Figure 2.3. Physical positions of *tb1* and *d8* on RefGen_v2. Positions of SNPs are obtained by blasting primer sequences using www.maizesequence.org and are approximate. Sites above the line in solid black boxes are evaluated in this study. Sites below the line in dashed boxes are from the study by Camus-Kulandaivelu et al. (2008).



Supplemental Figure 2.4. The region around *tb1* and *d8* on chromosome 1 (265,495,979 – 266,347,836 RefGen_v2), and all identified gene transcripts available at www.maizesequence.org. There are no other obvious candidate gene for flowering time in the region.



Supplemental Figure 2.5. Genome wide association results for flowering time (days to silking) in the NAM population using maize HapMapv1 and HapMapv2 SNPs. There are no significant sites identified in the region of *d8* (indicated by the gray line).



Supplemental Figure 2.6. R^2 between MITE in *vgt1* and all the other sites on chromosome 8. Blue dots indicate results from 7,539 GBS SNPs present in 200 or more of the 282 lines. Red dots indicate results from 4,197 55K SNPs present in 200 or more of the 282 lines.

CHAPTER 3

ANALYSIS OF HYBRID VIGOR AND YIELD IN DIVERSE MAIZE HYBRIDS

ABSTRACT

Hybrids created from the maize Nested Association Mapping (NAM) population were developed to examine hybrid vigor and yield. To better understand where in the genome loci affecting these traits are located and the relationship with recombination rate, this large hybrid population was evaluated in trials in nine environments over multiple years. All hybrids in the study show better phenotypic values than their inbred parents, expressing heterotic effect. The use of joint linkage mapping enables the identification of QTL associated with yield, as well as plant height and flowering in hybrids and their respective heterotic effects. A number of the mapped QTL are located in or on the edge of the pericentromeric regions with restricted recombination rate. Ridge regression was used to calculate marker effect estimates across the genome to predict breeding values in the hybrids. Considering the modest sample sizes within each subfamily, reasonable prediction accuracies for plant height and flowering were obtained using a five-fold cross-validation.

INTRODUCTION

The most important goal for maize breeding is grain yield. Over the last century, approximately half of improvement in yield is due to genetic improvement by plant

breeders, while the remainder is improvements in agronomic practices [1]. Yield is a measure of overall plant health, and it is frequently a result of genotypic stress tolerance. The majority of the increase in yield due to genetic improvements is through increased stress tolerance [2], and better interaction between genotype and agronomic management [1], rather than through specific focus on increased yield potential or heterotic effect. The increase in stress tolerance is a consequence of selection for improved yield stability under a range of environmental conditions. Under optimal conditions the yield potential per plant has not changed significantly over the last 20-40 years, and yield potential is at least three times greater than the average US yield [2]. Since the introduction of the use of hybrids in maize further increases in the heterotic effect has not appeared to have a major influence on yield increase because the heterotic effect seems to have been almost constant since the 1930s [3]. Still, there has been a yield increase of a mean of 0.1 Mg/ha per year since hybrid introduction [4].

An enormous breakthrough in the breeding of maize was the development of hybrids. Shull first defined heterosis in maize over a century ago [5], as 'the superiority of heterozygous genotypes with respect to one or more characters in comparison with the corresponding homozygotes'. Since then maize breeders have turned a simple field observation into the basis of almost all modern maize breeding [6]. In addition, since maize is a natural outcrossing species, plant breeders have developed an "unnatural" state of inbred maize, which cannot be found in the wild. This might be why maize hybrids express a relatively high heterotic effect. The heterotic effect of a hybrid relates to the genetic distance between the two parents

[7,8], and the level of adaptation to the growing environment of the parents. These adaptations include altered flowering time, disease resistance, growth habits and stress tolerance.

Over the past hundred years, four main hypotheses have been proposed as the explanation of heterosis, and none of them are necessarily mutually exclusive. First, epistasis being the interaction between genes at two or more loci affecting the expression of the phenotype of interest and may lead to superior performance. Epistasis is most likely less important on a general level but of more importance in specific hybrid combinations. Epistasis seems to be more important to selfing species, like rice [9]. Second, the dominance model was first proposed by Bruce (1910) [10] and Keeble and Pellew (1910) [11] and is based on the complementary effect of dominance factors introduced from each parent. Heterosis has been explained as partial or complete dominance in maize [12]. Third, the overdominace model proposes the heterozygous combination of alleles at the same locus to express greater phenotypic values than either homozygous combination [9]. In the overdominance model no linkage is required between the acting loci, nor is the involvement of multiple loci necessary to express a heterotic effect in the hybrid. Last, pseudo-overdominance extends the concept of dominance by including linkage [13,14]. A locus can seem to act in an overdominant fashion when it in fact is a number of genes that are linked due to lack of recombination in the region, the Hill-Robertson effect [15,16]. Locations of identified QTL for heterosis frequently correspond with pericentromeric regions (e.g. [17–19]).

The contribution of the different hypothesis proposed to control heterosis is most likely influenced by reproductive system [9]. For example, self-pollination supports maintaining of epistatic networks throughout the genome, but also exposes recessive and deleterious alleles to selection pressure in a relatively short period of time. In outcrossing species, deleterious recessive alleles remain in the genome for a longer time in a heterozygote state, often in the regions with restricted recombination. Inbred lines in modern maize breeding may have inherited this genetic load. While deleterious alleles in the inbreds are complemented for in hybrids, it might not fully explain the superiority of hybrids. Most genes in hybrids have a gene expression equal to the mid-parent value [20], but a significant proportion of the genes do not follow that assumption [21,22].

The chromosomal arms, regions near the end of the chromosome, exhibits higher recombination rates relative to physical distance, and the opposite is true in the pericentromentic regions [16,23,24]. Correlations of 35% between residual heterozygosity and the inverse of recombination rate suggest that large effects for heterosis are located in regions with low recombination. This proposes that recombination rate is the major factor determining residual heterozygosity and contributor of maintaining a higher level of heterozygosity. Regions with increased residual heterozygosity had more than 30% of all genes and nearly average diversity [24]. The suppressed level of recombination in these genetically divergent centromeric regions hinders the most optimum allelic combination to be formed [24]. This evidence argues for pseudo-overdominance as a major source of the heterotic effect due to linkage between loci in repulsion [25,26].

In this study hybrids of the NAM population have been developed to further examine the hypothesis proposed that pericentromeric regions with low recombination rate and higher than average level of heterozygosity have a large effect on hybrid vigor based on the pseudo-overdominance model. With the use of joint linkage mapping, QTL have been identified for yield as well as plant height and flowering time in inbreds, hybrids, and their heterotic effects. Here, we get a better understanding of the relationship between effect estimates in hybrids and recombination rate by analysis on a haplotype level. In addition, we evaluate whether this diversity can be modeled through genomic prediction approaches to predict agronomic traits in hybrids.

MATERIAL AND METHOD

Germplasm

In this study we are using the maize nested association mapping (NAM) population developed by the Genetic Architecture of Maize and Teosinte Project consortium [16]. NAM was created by selecting 25 inbreds to maximize diversity, and crossing them to the reference inbred, B73. From each of the 25 subfamilies, 200 progeny were chosen, selfed for five generations and subsequently sib-mated. This resulted in a mapping population of nearly 5,000 RILs. From these RILs, we selected a subset of 60-70 lines from each subfamily (except the popcorn subfamily Hp301), and created hybrids by crossing to the male tester PHZ51. PHZ51 is a non-stiff stalk line developed by DuPont Pioneer with expired Plant Variety Protection (PVP), it is a yellow dent

classified as “mixed” developed by crossing PH814 and PH848 [27]. The subset of NAM female lines for hybrid development was based upon flowering time. The earliest RILs from the late subfamilies and the latest RILs from the early subfamilies were selected to reduce flowering time variation and make hybrid production using isolation plots more manageable.

Field Evaluation

Hybrids were grown and evaluated in Sandhills NC, Bradford MO, West Lafayette IN, and Slater IA in the summer of 2010, as well as Kinston NC, Bradford MO, West Lafayette IN, Ames IA, and Aurora NY in the summer of 2011. All nine environments were cultivated in a conventional manner with respect to fertilization, weed, and pest management. Hybrids were planted in two-row plots with a single replication per environment, except for the field in NY 2011, which was planted in single rows and only developmental traits were measured. The experiment was blocked by subfamily to avoid competition for space and light interception resulting from height variation. Entries were randomized within blocks, and blocks were randomized within environments. Due to weather conditions, five environments possessed substantial variation for root and stalk lodging. In 2010, Bradford MO, Slater IA, and Sandhills NC were damaged by lodging, and in 2011 the Bradford MO environment and West Lafayette IN environment were also damage by lodging. Because of the damage, yield data was obtained from only six of the eight fields planted in two-row plots for yield evaluation (excluding the 2011 Bradford MO environment and 2011 West Lafayette IN environment which were damage to such an extent that it was impossible to harvest

the fields). Data for plant height were collected from eight of the nine environments evaluated for developmental traits (excluding the 2011 Bradford MO environment which was severely damaged by lodging early in the season). Data for flowering were collected from seven of the nine environments. Excluding the 2011 Bradford MO environment for the same reason as for plant height as well as the 2010 Sandhills NC environment because of shortage of personnel at that particular time of the season.

Phenotyping

All phenotypic data of the hybrids were collected on a plot basis. Number of days from planting until half the plants in a plot shed pollen or had a visible silk was used as the criterion to measure days to anthesis and days to silk, respectively. All other traits were measured at full maturity after flowering. Plant height was measured as the distance from the soil line to the base of the flag leaf. Yield was measured using a two-row combine and moisture was measure automated on the combine. Yield was adjusted to15.5% moisture content and reported in tons per hectare.

In this study, when reporting results for inbreds it refers to the female inbreds (NAM RILs). The flowering time data for the NAM RILs were obtained from the Buckler et al., 2009 study [28], and plant height data was obtained from Peiffer et al., [29]. Data for leaf traits (length, width and angle) is from the Tian et al., 2011 study [30] and traits for female and male inflorescence were obtained from the Brown et al 2011 study [31]. All other traits were collected by the Genetic Architecture of Maize and Teosinte Project but have yet not been made public.

Based on previous collected data for the inbreds and data collected for the hybrids in this study, best-parent heterosis and mid-parent heterosis were calculated. Best-parent heterosis is the difference between the hybrid value and the better value from the female and male inbred, and mid-parent heterosis is the difference between the hybrid value and the average value between the female and male inbred.

Genotyping

Genotypic data for joint linkage mapping was collected as previously described [16,28]. In total, 1,106 markers were scored on an Illumina GoldenGate Assay across the NAM RILs.

To estimate marker effects across the genome and calculate prediction accuracies using ridge regression, the NAM RILs were genotyped at low coverage using the GBS platform [32]. As a result of the relatively low coverage, about 80% of the data for individual markers was missing and about 80% of the heterozygous loci were called as homozygotes. An HMM (Hidden Markov Model) algorithm was used to correct the heterozygote calls to about 99.8% accuracy [33]. Following that a set of markers to be used in fitting a regression model was imputed from the GBS data at 0.2cM intervals based on flanking markers. The marker values identified the parent of origin, with B73 coded as 0 and the non-B73 parent coded as 2. Heterozygous loci were coded as 1. For the imputed sites, if the flanking markers were identical, the site was set to that value. If the flanking markers were not identical, an interpolated value was used based on the relative genetic distance from each of the flanking markers.

Statistical Analysis

Statistical analyses were performed using SAS [34] software, and R [35] scripts and libraries. Best Linear Unbiased Estimations (BLUEs) for all phenotypes across environments were calculated as LSmeans with range position as fixed effect for each genotype within block using SAS software.

After deriving BLUE, the base package in R was used to calculate Pearson correlation coefficients and to study relationships between hybrid yield and 28 different phenotypes evaluated in the inbreds at the genotypic level across and within the NAM subfamilies.

To characterize genetic architecture, joint linkage mapping of BLUEs across environments was performed using proc GLMSelect in SAS. After adjusting for differences between NAM subfamilies, an imputed set of 1,106 markers were nested within each NAM subfamily and regressed against BLUEs for each of the phenotypes across in a stepwise manner, as described in Buckler et al. (2009) [28]. For the stepwise procedure, model inclusion and exclusion of subfamily nested markers were discerned by comparison with a null distribution based on permutation testing. The *p*-value derived from the null distribution for model inclusion was 0.001 at an alpha level of 0.05.

To assess our ability to predict breeding values from all assayed genetic diversity, we performed genomic prediction by ridge regression BLUP in R using the rrBLUP package [36]. Genomic relationship matrices were calculated using 7,389 marker intervals for the NAM population. To calculate effect estimates for each

marker across the genome for each phenotype, genetic matrices were regressed against each phenotypic BLUE including all phenotypic values within each subfamily. The same calculations were made to estimate genomic estimated breeding values (GEBVs) for each phenotype. These calculations were also made within each subfamily. Accuracy for the prediction was determined by Pearson correlation of the predicted GEBVs and the observed values. This was accomplished by fitting all genotypes into a ridge regression model. To assess overfitting and the robustness of the modeling approach, each subfamily was randomly divided into five mutually exclusive subpopulations, and four of the five subpopulations were used to construct a model, which was fitted to the last remaining subpopulation. Accuracy of the prediction was averaged across the five subpopulations. The process was repeated with 20 different randomizations of the five subpopulations. This was performed for each trait.

RESULTS

Phenotypes

Due to weather conditions such as rain and strong winds, five environments possessed substantial root and stalk lodging (Table 3.1) (For details refer to chapter four). From 85-99 percent of the plots in the five environments had one or more plants lodged, and 13-59 percent of the plants had visible damage. As a result, the IN11 and MO11 environments were not harvestable and the final dataset for yield consisted of six environments. The MO11 environment was also excluded from the dataset of plant height, and NC10 and MO11 were excluded form the dataset of flowering time,

leading to a total of eight environments for plant height and seven environments for flowering time.

Table 3.1. Percent of plots and percent of plants per environment damaged by root lodging, stalk lodging, and total lodging.

	IA10	IN11	MO10	MO11	NC10
% damaged plots root	96	78	75	6	7
% damaged plots stalk	11	87	67	99	83
% damaged plots total	96	98	91	99	85
% damaged plants root	21	11	18	0.2	12
% damaged plants stalk	0.6	48	10	41	0.3
% damaged plants total	21	59	8	41	13

In this study, all evaluated hybrid genotypes express hybrid vigor for plant height in terms of best-parent heterosis as well as mid-parent heterosis (Figure 3.1; Supplement table 3.1). In terms of flowering time, this was not true for a small number of subfamilies (Supplement table 3.2).

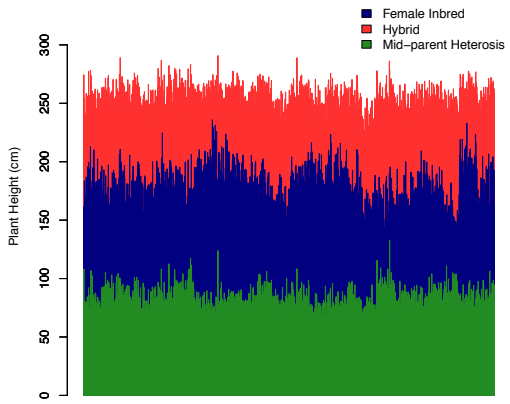


Figure 1. Distribution of plant height values for the female inbred (blue), corresponding hybrid (red), and mid-parent heterosis (green).

Trait correlation

Across the hybrid NAM population, days to anthesis and days to silk were strongly correlated (Table 3.2). Flowering time was moderately positively correlated with plant

height but negatively correlated with yield (Table 3.2). There was no correlation between plant height and yield across the NAM population. By subfamilies the correlation for all the traits varied between r = -16 to r = 0.46, except for the two flowering traits which varied between r = 0.40 to r = 0.98. At least one subfamily or more showed very low or no correlation between the traits.

Table 3.2 Correlation between yield (T/ha), plant height (cm), days to anthesis, and days to silk. Top table reports correlations across NAM. Middle table reports highest and lowest directional correlation within the subfamilies. Bottom table reports highest and lowest absolute correlation within subfamilies. * indicates a significance of <0.01 and ** indicates a significance od <0.001.

Correlation across NAM

	Yield	**Plant height**	**d2a**	**d2s**
Yield	----------	0.00	-0.23**	-0.26**
Plant height		----------	0.19**	0.22**
d2a			----------	0.89**
d2s				----------

Directional correlation within subfamilies

		high			
		Yield	**Plant height**	**d2a**	**d2s**
low	**Yield**	----------	0.34*	0.70**	0.70**
	Plant height	-0.33*	----------	0.44*	0.46**
	d2a	-0.70**	-0.16	----------	0.98**
	d2s	-0.70**	-0.17	0.40**	----------

Absolute correlation within subfamilies

		high			
		Yield	**Plant height**	**d2a**	**d2s**
low	**Yield**	----------	0.34*	0.70**	0.70**
	Plant height	0.02	----------	0.44*	0.46**
	d2a	0.00	0.00	----------	0.98**
	d2s	0.00	0.00	0.40**	----------

The NAM inbred populations have been scored for numerous traits by the maize community over the last several years. These NAM hybrids provide an opportunity to

examine how predictive these measures are to hybrid performance. In general no strong correlations were found across the NAM populations between hybrid yield and developmental traits measured in the inbreds (Table 3.3). Exceptions are the negative correlations with days to silk (r = -0.25) and days to tassel (r = -0.22), as well as plant height (r = -0.10) and ear height (r = -0.14). The strongest positive correlation was with 20 kernel weight, r = 0.11. Analysis within subfamilies show that all traits had very low, or no, impact on yield in at least one of the subfamilies. However, every trait also had a modest correlation (r = 0.22 – 0.50), positive or negative, with yield in all of the subfamilies. Each of the 28 traits examined had both positive and negative correlations with yield. The same traits had a modest negative correlation in one subfamily and positive correlation in another. The strongest correlation with yield was with plant height (r = 0.44) and ear height (r = 0.46). In addition the strongest negative correlation was between days to tassel and yield (r = -0.50). Maximum correlation for ear weight (r = 0.27) and 20-kernel weight (r = 0.30) was relatively modest.

For plant height, the correlation between inbred and hybrid was 0.51. By subfamily it ranged between 0.28 and 0.72. For days to anthesis, the correlation across NAM was 0.48, and it varied from 0.19 to 0.71 within subfamilies. Same patterns were observed for days to silk and days to anthesis.

Table 3.3. Correlation between hybrid yield and traits measured in the corresponding female inbreds. Results are shown across all the NAM population, as well as directional and absolute correlation within subfamilies. * indicates a significance of <0.01 and ** indicates a significance od <0.001.

		Within subfamily			
		Directional correlation		Absolute correlation	
	Across NAM	max	min	max	min
Germination Count	0.02	0.29	-0.25	0.29	0.01
Stand Count	0.05	0.24	-0.29	0.29	0.00
Days to Silk	-0.25**	0.16	-0.48**	0.48**	0.00
Days To Anthesis	-0.22**	0.20	-0.50**	0.50**	0.01
Tassel Length	-0.08*	0.28	-0.19	0.28	0.00
Main Spike Length	-0.05	0.32	-0.09	0.32	0.02
Tassel Primary Branches	-0.03	0.12	-0.29	0.29	0.03
Ear Number	0.10**	0.37*	-0.19	0.37*	0.00
Number of Nodes Ear - Roots	-0.17**	0.28	-0.27	0.28	0.01
Number of Nodes Tassel - Ear	-0.09**	0.22	-0.41*	0.41*	0.01
Number of Brace Root Nodes	-0.06	0.38*	-0.31	0.38*	0.00
Plant Height	-0.10**	0.44**	-0.33*	0.44**	0.00
Ear Height	-0.14**	0.46**	-0.11	0.46**	0.01
Leaf Length	-0.13**	0.25	-0.28	0.28	0.02
Leaf Width	-0.06	0.26	-0.26	0.26	0.00
Upper Leaf Angle	-0.02	0.29	-0.22	0.29	0.01
Middle Leaf Angle	-0.03	0.33*	-0.20	0.33*	0.01
Tillering Index	-0.03	0.35*	-0.19	0.35*	0.00
Cob Diameter	-0.04	0.26	-0.33*	0.33*	0.00
Cob Weight	-0.02	0.29	-0.28	0.29	0.01
Ear Diameter	0.00	0.21	-0.34*	0.34*	0.02
Ear Length	0.02	0.22	-0.20	0.22	0.00
Seed Set Length	0.01	0.28	-0.18	0.28	0.01
Ear Row Number	0.03	0.20	-0.28	0.28	0.00
Total Kernel Volume	0.04	0.24	-0.26	0.26	0.02
Ear Weight	0.05	0.27	-0.15	0.27	0.00
Ear Rank Number	-0.01	0.28	-0.28	0.28	0.00
20 Kernel Weight	0.11**	0.30	-0.14	0.30	0.00

Heritability

Heritability for the traits measured in the hybrids was calculated using mixed model with spatial correction in ASReml [37]. The heritability across the environments varied between 0.65 and 0.79 for the three traits, yield, flowering time, and plant height (Table 3.4).

Table 3.4. Heritability estimates for the yield, flowering time and plant height measure in the hybrids, within and across environments.

	Yield	Flowering time	Plant height
Across env	0.66	0.79	0.65
NC10	0.67	Na	0.64
MO10	0.65	0.84	0.67
WL10	0.65	0.73	0.64
IA10	0.64	0.77	0.66
NC11	0.68	0.86	0.66
IA11	0.66	0.84	0.67
WL11	Na	0.79	0.65
NY11	Na	0.73	0.63

Joint linkage mapping

Using this smaller subset of genotypes, we remapped QTL in the inbreds to evaluate the change in statistical power. Fourteen QTL were mapped for the inbreds for both plant height and flowering time (Table 3.5). This is a smaller number than previously reported [28,29]. An even smaller number of QTL were identified for hybrids (10 for plant height, 4 for days to anthesis, and 5 for days to silk). For best-parent heterosis and mid-parent heterosis a larger number of QTL was identified for plant height (7 for best-parent heterosis, and 5 for mid-parent heterosis) than flowering time (2 for days to anthesis best-parent heterosis, 2 for days to anthesis mid-parent heterosis, 4 for days

to silk best-parent heterosis, and 1 for days to silk mid-parent heterosis). A larger number of QTL was mapped for best-parent heterosis than mid-parent heterosis.

A minority of the identified QTL were located in the pericentrometic regions (within 10cM from the centromere) with restricted recombination rate (Table 3.5). It should be noted that a number of the QTL here classified, as being located on the chromosomal arms were on the edge, in the regions between the arms and pericentromeric regions, depending on how the pericentromeric regions were defined.

Table 3.5. Results for joint linkage mapping for plant height, days to anthesis, and days to silk. Mapping was performed on data collected on inbred, hybrid, best-parent heterosis, and mid-parent heterosis. Table reports QTL location as chromosome and cM position. Gray boxes indicates QTL located in pericentromeric regions (within 10 cM from the centromere).

Plant Height

Inbred	Hybrid	Best-parent heterosis	Mid-parent heterosis
chr1 60.8cM	chr1 0.9cM	chr1 115cM	chr2 97.5cM
chr1 116.2cM	chr2 28.2cM	chr2 98.9cM	chr2 98.9cM
chr3 18.2cM	chr2 84.2cM	chr3 15cM	chr3 129.8cM
chr3 79.3cM	chr2 136.3cM	chr3 129.8cM	chr7 43.9cM
chr4 65.9cM	chr3 90cM	chr4 60.6cM	chr10 58.4cM
chr4 116.1cM	chr5 98.2cM	chr5 81.7cM	
chr5 58.3cM	chr7 69.8cM	chr9 53.1cM	
chr5 108.8cM	chr8 62.3cM		
chr6 22.1cM	chr9 64.5cM		
chr7 72.2cM	chr10 44.8cM		
chr8 57.5cM			
chr8 70.6cM			
chr9 56.7cM			
chr10 32.4cM			

Days to Anthesis

Inbred	Hybrid	Best-parent heterosis	Mid-parent heterosis
chr1 20.1cM	chr1 79.9cM	chr3 65.2cM	chr3 65.2cM
chr1 84.9cM	chr2 129.8cM	chr9 49.5cM	chr6 61.8cM
chr2 63cM	chr3 65.2cM		
chr2 67.9cM	chr10 41.9cM		
chr2 127.3cM			
chr3 56cM			
chr3 123.9cM			
chr4 118.4cM			
chr5 72.5cM			
chr6 96.4cM			
chr8 67.4cM			
chr9 53.1cM			
chr9 74cM			
chr10 41.9cM			

Days to Silk

Inbred	Hybrid	Best-parent heterosis	Mid-parent heterosis
chr1 17.4cM	chr1 87.9cM	chr2 17.8cM	chr2 17.8cM
chr1 84.6cM	chr2 129.8cM	chr2 19.5cM	
chr2 63cM	chr7 63.7cM	chr3 140.7cM	
chr2 127.3cM	chr7 71.2cM	chr9 49.1cM	
chr3 56cM	chr10 41.9cM		
chr3 115.8cM			
chr3 131.4cM			
chr4 55.8cM			
chr5 65.4cM			
chr7 75.3cM			
chr8 70.6cM			
chr9 62cM			
chr10 41.9cM			
chr10 91cM			

T-tests with two-tail distribution and two-sample unequal variance were used to examine the difference in recombination rate in the mapped QTL intervals for inbreds, hybrids, best-parent heterosis and mid-parent heterosis (Supplemental table 3.3). Only two combinations were marginally significant for difference in recombination rate: comparison days to anthesis best-parent heterosis and days to silk inbred gave *p*-value of 0.029, and days to anthesis best-parent heterosis and days to anthesis inbred gave *p*-value of 0.054.

The 7,389 marker intervals used to estimate effects across the genome were divided into two sets; low recombination rate and high recombination rate. F-test between the two datasets show that there was a highly significant difference of effect variance in low and high recombination regions, particularly for inbreds and hybrids (Table 3.6). Small differences in correlation between recombination rate and effects were observed in the two datasets. In inbreds r = 0.029 in regions with low recombination and r = 0.009 in regions with high recombination. The opposite pattern is true for the hybrids. The intervals with the largest effects were located in regions with low, but not the lowest, level or recombination (Figure 3.2).

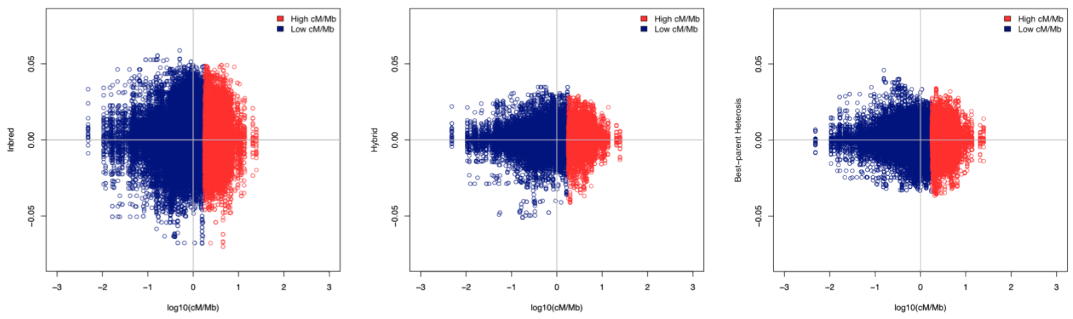


Figure 3.2. Genotypes divided into low and high recombination rate plotted against effect estimates for inbred, hybrid, and best-parent heterosis in plant height.

Table 3.6. F-test results from comparing distribution of effect estimates in regions with low and high recombination rate. Correlation between recombination rate and effect estimates in low and high recombination regions.

	Inbred	Hybrid	Heterosis
F-test	$2.6\ 10^{-273}$	$2.6\ 10^{-223}$	$2.6\ 10^{-36}$
Correlation (low cM/Mb)	0.029	0.008	-0.038
Correlation (high cM/Mb)	0.009	0.020	0.003

Four QTL were mapped for yield in the hybrids using joint linkage mapping (Table 3.7). The QTL on chromosomes 7 and 10 were located in the pericentromeric regions (10 cM within the centromere). The mapped loci on chromosome 3 and 5 were on the edge of the edge between the pericentromeric region and the chromosome arm. All of the four mapped QTL were in intervals with a recombination rate lower that the average across the genome, 1.382 cM/Mb. QTL for yield had in previous studies [17–19,38] been mapped to the same general regions as reposted in this study.

Table 3.7. Results for joint linkage mapping for yield. Result reports QTL location as chromosome and cM position. Gray boxes indicate QTL located in pericentromeric regions. cM/Mb is recombination rate in the QTL interval.

Chromosome	cM position	cM/Mb
3	33.7	1.293
5	50.8	0.595
7	49.1	0.198
10	41.1	0.165

Genomic prediction

This study generated reasonably high prediction accuracies for both plant height and flowering time in the hybrids taking into account the five-fold cross-validation and that the models were trained on a population size of around 50 lines. For plant height the accuracy was r = 0.57 across the NAM subfamilies with prediction estimated

within each subfamily (Figure 3.3). Equivalent estimations for flowering time gave an accuracy of r = 0.71.

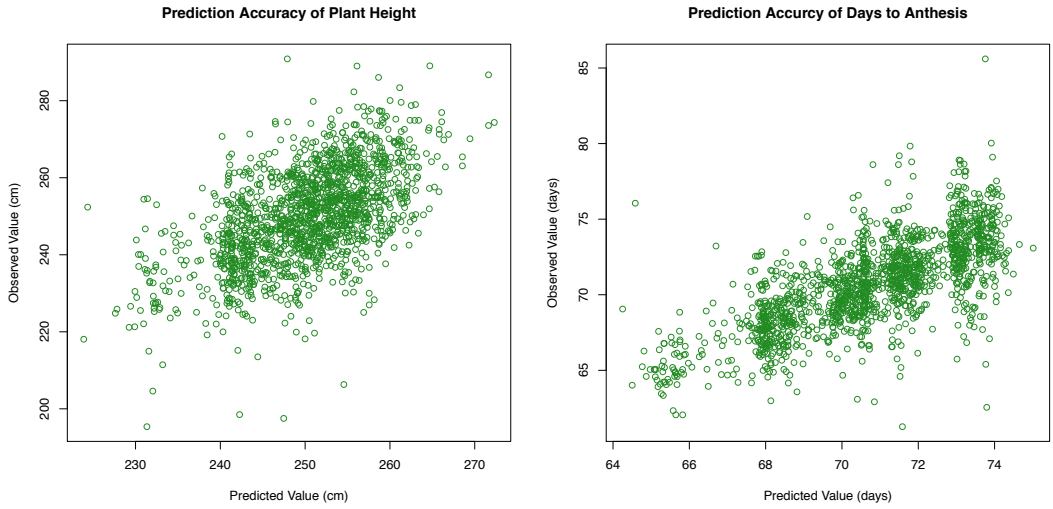


Figure 3.3. Prediction accuracies for plant height and days to anthesis in hybrids estimated using ridge regression within subfamilies.

Heterosis was estimated as both best-parent heterosis and mid-parent heterosis. Using ridge regression to predict breeding values for the traits resulted in accuracies comparable to predictions for the hybrid value. For plant height, best-parent heterosis was predicted with an accuracy of r = 0.6, estimated within subfamily. Prediction accuracy for mid-parent heterosis for plant height was estimated to r = 0.50 (Figure 3.4).

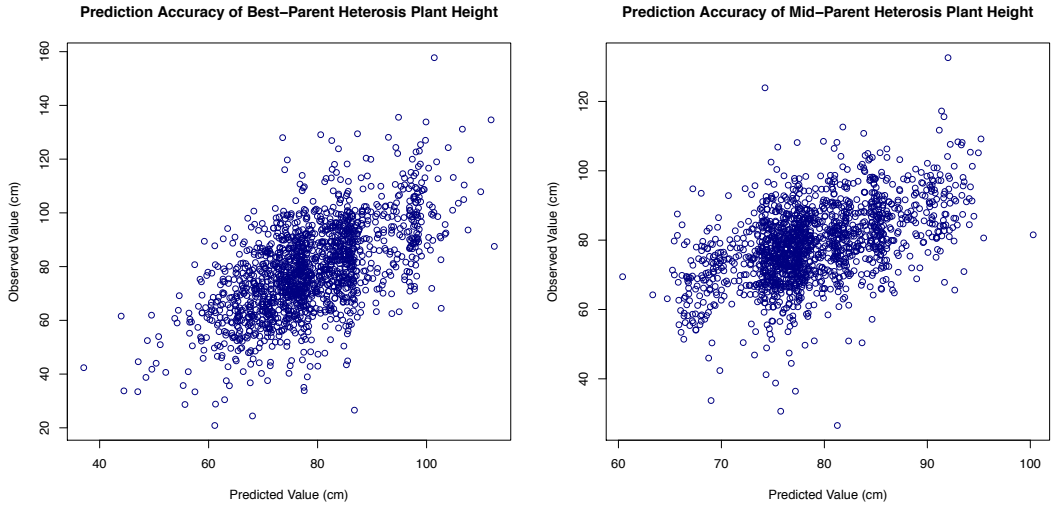


Figure 3.4. Prediction accuracies for best-parent heterosis and mid-parent heterosis for plant height estimated using ridge regression within subfamilies.

By using ridge regression we were able to predict the values of yield in the hybrids with an accuracy of r = 0.55, by performing the estimation within subfamily with a five-fold cross-validation (Figure 3.5).

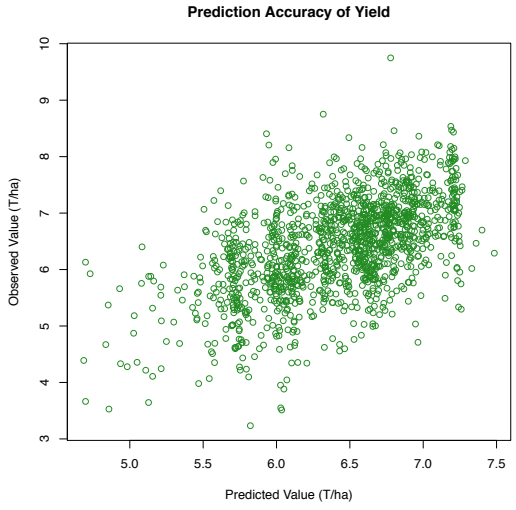


Figure 3.5. Prediction accuracy for yield in hybrids estimated using ridge regression within subfamilies.

Prediction accuracy for hybrid yield and plant height within individual subfamilies ranged from 0.01 to 0.61 (Table 3.8). In some cases, one trait was predicted with high accuracy in a subfamily and a second trait with a low accuracy in the same subfamily. We do also observe negative correlations in some of the subfamilies.

Table 3.8. Prediction accuracy for individual subfamily for hybrid yield and plant height within each subfamily.
* indicates a significance of <0.01 and
** indicates a significance of <0.001.

Subfamily	Yield	Plant Height
B97	0.03	0.45**
CML103	0.05	0.26
CML228	0.06	0.69**
CML247	0.33*	0.17
CML277	0.43**	0.68**
CML322	0.18	0.37*
CML333	0.01	0.11
CML52	-0.61**	0.19
CML69	0.08	0.49**
Il14h	NA	-0.13
Ki11	0.43*	0.25
Ki3	-0.25	-0.47**
Ky21	0.03	0.27
M162W	0.08	0.11
M37W	-0.28	0.38*
Mo18W	-0.03	0.17
MS71	-0.58**	0.03
NC350	-0.56**	0.28
NC358	-0.61**	0.46**
Oh43	-0.36*	0.39*
Oh7B	-0.61**	-0.08
P39	NA	-0.22
Tx303	-0.39*	0.35**
Tzi8	0.27	0.3

DISCUSSION

Heritability

Heritability estimated for the traits measured in the hybrids in this study were of moderate values (Table 3.4). The moderate heritability does have a negative effect on the analyses and limits our ability to map QTL as well as the accuracy of genomic prediction. These heritability estimates can be due to the nature of the traits. We expect flowering time and plant height to have high heritability. Traits such as yield, which are more influenced by the environment and most likely controlled by a large number of loci tend to have lower heritability. A second reason is the lodging damage that a number of the fields suffered. It is very possible that damage related to lodging events can be much more than just the visual breakage of the plants, such as insect and pathogen infestation, interruption in xylem and phloem and, unbalance in carbon relocation, resulting in phenotypic values that do not represent the genotype in a good location.

Trait Correlation

Across the NAM hybrids, taller plants flower in general later (Table 3.2), a pattern that has been seen in maize before (i.e. [39–41]). The negative correlation between flowering time and yield can be explained by the proportion of tropical material in the NAM population. These tropical lines are less adapted to the growing environment than the more temperate material. In US temperate material, flowering time and grain

yield are positively correlated. Hybrids with a longer season yield more and have higher grain moisture [42].

There are no correlations between plant height and yield across the NAM population, but within subfamilies the correlation varies from slightly negative to moderately positive. The positive correlation can be explained by the correlation between plant height and flowering time. Later flowering and more biomass result in higher carbon fixation, which can be used for grain fill. In conclusion, there is no general relationship between flowering time, plant height, and yield across the NAM population, except between days to anthesis and days to silk. To be able to use flowering time or plant height to predict yield it will be more accurate and efficient within subfamilies or at least using smaller groups of related subfamilies.

Correlation between female inbreds and hybrids across NAM for flowering and plant height is around 0.50, suggesting that inbreds can be used to predict the hybrid values to a certain extent. Level of correlation of the three traits varies by about 0.20 to 0.70 implying prediction based on inbred values can be much more successful in some subfamilies than others.

In general no strong correlations were observed for hybrid yield and 28 different traits measured in the female inbreds (Table 3.3). The strongest positive correlation was 20-kernel weight with a value of only 0.11. This suggests that a single trait evaluated in inbreds is probably not very successful to predict hybrid performance. Within subfamilies, all traits have very low or no impact on yield. All inbred traits also have negative affect in one subfamily and positive in another. To successfully be able to use data collected in inbreds to predict hybrid yield, a model

has to been developed including a number of traits, but the importance of the traits will vary depending on the population. In addition, the population cannot be too genetically diverse.

Joint linkage mapping

In this study QTL were mapped for yield in hybrids (Table 3.7) as well as plant height and flowering time for inbreds, hybrids and their heterotic effects (Table 3.5). Due to the relatively small sample size, this study identified about one third of the QTL that have been identified in previous studies of the NAM inbreds for flowering time [28] and plant height [29]. Here we used about one third of the full NAM population (5,000 RILs), which results in less power to detect QTL using this method. Additionally, a number of the field locations were damaged by environmental conditions, which had a negative impact on the quality of the collected data.

Over the years, a large number of studies have been performed to genetically map loci affecting yield. In this study, four loci were mapped for yield using joint linkage mapping. QTL in the same regions have previously been mapped for all four QTL: QTL on chromosome 3 [18], chromosome 5 [39,43], chromosome 7, and chromosome 10 [12,18].

Substantial portions of the identified QTL are located in, or near, the pericentromeric regions, regions with restricted recombination rate. These results agree with previous findings by other research groups using different experimental design and germplasm [18,19]. Findings suggest that loci controlling these traits are

more frequently located in regions with low recombination rate. As a result, this supports the pseudo-overdominance hypothesis to explain heterosis.

No significant differences were observed when comparing recombination rate in the QTL intervals mapped for inbreds, hybrids and heterosis. Though, it should be pointed out that nearly all of the QTL intervals had a recombination rate lower then the average value of the 1,106 marker intervals on the NAM map. When comparing the marker effects estimated using ridge regression and dividing the dataset into two halves (low and high recombination rate), F-tests between the two datasets were significant for inbred, hybrid, and heterosis (Table 3.6). The largest effect estimates are located in regions with restricted but not the lowest level of recombination rate. This agrees with the findings from the joint linkage mapping, where a number of the QTL are located in the pericentromeric regions. Additionally, an even larger proportion is located on the edge between the centromeric regions and the chromosomal arms. These regions have sufficient recombination for new haplotypes to form, but limited enough for linkage between loci leading to pseudo-overdominance. These regions may also have the lowest genic space to recombination ratio.

At this moment we can neither accept nor reject the pseudo-overdomiance hypothesis. The data in this study is analyzed on a haplotype level as the sequence for the tester, PHZ51, was not available at this time. This poses an important challenge as recombination rate, heterozygosity, and gene density are substantially correlated at these larger scales. Analysis on a nucleotide level is required to break up this relationship and to be able to get a better understanding of how these characteristics

are affecting the manifestation of hybrid vigor. To complement the findings in this study the genetic material needs to be sequenced with higher resolutions. Especially the tester, PHZ51, needs to be sequenced, which will enable us to model dominance on a SNP level.

Genomic prediction

Yield is the most important trait in maize breeding. However, it has a moderate heritability and is controlled by a large number of loci with small effects. These factors make improvement of yield a complex task. Furthermore, evaluating new hybrid combinations is an expensive and labor-intensive process. Statistical methods, first used in animal breeding [44] have been developed to perform genomic selection or prediction [45,46].

Here we were able to use ridge regression (rrBLUP) to predict hybrid breeding values in the NAM hybrid population as well as heterotic effect with reasonable accuracies (around r =0.55) (Figures 3-5), particularly, when taken into consideration that predictions were performed within each subfamily using a five-fold cross-validation resulting in a training population of around 50 lines. On the other hand, there are large variations of prediction accuracies within individual subfamilies (Table 3.8). The prediction accuracies for yield can be low, and high for plant height in the same subfamily.

Negative prediction accuracies were observed in some of the subfamilies in this study. Similar findings have been observed in other populations and species [47,48]. Prediction accuracy has been shown to be influenced by heritability of the

trait, number of genetic markers, model used, population size, and genetic distance (Heslot 2012). In this study, phenotypes had moderate heritability due to the nature of the trait and environmental effects (field damage), subfamilies were of small sample sizes, and the subfamilies captured a relatively large proportion of diversity. These are factors that potentially can explain the negative correlations. Additionally, the negative correlations may not have a biological explanation but rather be a result of the Simpson's paradox [49], whereby a trend in different groups of data disappears or changes direction when the data for the groups are joined. Further investigations have to be made to better understand the underlying factors of the negative prediction accuracies.

This study of the hybrid NAM population has enabled us to map QTL of yield as well as inbred, hybrid and heterotic effects for plant height and flowering time. The majority of these QTL are located in regions with below average recombination rate. In addition, marker effect estimates suggest that regions of the genome with larger effects on these traits are in regions with restricted recombination. This study has given us a better understanding of the relationship between yield and heterosis with recombination. However, there is more to learn, and further genomic data is needed to further examine the role of dominance across diverse maize germplasm.

REFERENCES

1. Duvick DN (2005) The contribution of breeding to yield advances in maize (*Zea mays* L.). Advances in Agronomy 86.

2. Tollenaar M, Lee E (2002) Yield potential, yield stability and stress tolerance in maize. Field Crops Research 75.

3. Duvick DN, Cassman KG (1999) Post–Green Revolution Trends in Yield Potential of Temperate Maize in the North-Central United States. Crop Science 39: 1622–1630.

4. Tollenaar M, McCullough D, LM D (1994) Physiological basis of the genetic improvement of corn. Slafer G, editor

5. Shull G (1908) The Composition of a Field of Maize. Journal of Heredity 4: 296–301.

6. Duvick DN (2001) Biotechnology in the 1930s: the development of hybrid maize. Nature reviews Genetics 2: 69–74.

7. Smith OS, Smith JSC, Bowen SL, Tenborg R a., Wall SJ (1990) Similarities among a group of elite maize inbreds as measured by pedigree, F1 grain yield, grain yield, heterosis, and RFLPs. Theoretical and Applied Genetics 80: 833–840.

8. Barbosa A, Geraldi I, Benchimol L, Garcia A, Souza C, et al. (2003) Relationship of intra- and interpopulation tropical maize single cross hybrid performance and genetic distances computed from AFLP and SSR markers. Euphytica 130: 87–99.

9. Garcia AAF, Wang S, Melchinger AE, Zeng Z-B (2008) Quantitative trait loci mapping and the genetic basis of heterosis in maize and rice. Genetics 180: 1707–1724.

10. Bruce A (1910) The mendelian theory of heridity and the arguments of vigor. Science 32: 627–628.

11. Keeble F, Pellew C (1910) The mode of inheritance of stature and of time of flowering in peas (Pisum sativum). Journal of Genetics 1: 47–56.

12. Stuber CW, Lincoln SE, Wolff DW, Helentjaris T, Lander ES (1992) Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132: 823–839.

13. Moll R, Lindsey M, Robinson H (1964) Estimates of genetic variances and level of dominance in maize. Genetics 49: 411–423.

14. Springer NM, Stupar RM (2007) Allelic variation and heterosis in maize: how do two halves make more than a whole? Genome research 17: 264–275.

15. Hill W, A R (1966) The effect of linkage on limits to artificial selection. Genetic Research 8: 269–294.

16. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, et al. (2009) Genetic properties of the maize nested association mapping population. Science 325: 737–740.

17. Graham GI, Wolff DW, Stuber CW (1997) Characterization of a Yield Quantitative Trait Locus on Chromosome Five of Maize by Fine Mapping. Crop Science 17: 1601-1610

18. Schön CC, Dhillon BS, Utz HF, Melchinger AE (2010) High congruency of QTL positions for heterosis of grain yield in three crosses of maize. TAG Theoretical and applied genetics 120: 321–332.

19. Lariepe A, Mangin B, Jasson S, Combes V, Dumas F, et al. (2011) The genetic basis of heterosis: multiparental QTL mapping reveals contrasted levels of apparent overdominance among traits of agronomical interest in maize (*Zea mays* L.). Genetics: 1–47.

20. Birchler J a, Auger D, Riddle N (2003) In Search of the Molecular Basis of Heterosis. The Plant Cell 15: 2236–2239.

21. Song R, Messing J (2003) Gene expression of a gene family in maize based on noncollinear haplotypes. Proceedings of the National Academy of Sciences of the United States of America 100: 9055–9060.

22. Auger DL, Gray AD, Ream TS, Kato A, Coe EH, et al. (2005) Nonadditive gene expression in diploid and triploid hybrids of maize. Genetics 169: 389–397.

23. Fengler K, Allen SM, Li B, Rafalski A (2007) Distribution of Genes, Recombination, and Repetitive Elements in the Maize Genome. Crop Science 47: S–83.

24. Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, et al. (2009) A first-generation haplotype map of maize. Science 326: 1115–1117.

25. Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nature reviews Genetics 10: 783–796.

26. Kaeppler S (2012) Heterosis: Many Genes, Many Mechanisms—End the Search for an Undiscovered Unifying Theory. ISRN Botany 2012: 1–12.

27. Kahler AL, Kahler JL, Thompson S a., Ferriss RS, Jones ES, et al. (2010) North American Study on Essential Derivation in Maize: II. Selection and Evaluation of a Panel of Simple Sequence Repeat Loci. Crop Science 50: 486.

28. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, et al. (2009) The genetic architecture of maize flowering time. Science 325: 714–718.

29. Peiffer J, et al. (in prep.) The Genetic Architecture of Plant Height. PloS one.

30. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, et al. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature genetics 43: 159–162.

31. Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, et al. (2011) Distinct genetic architectures for male and female inflorescence traits of maize. PLoS genetics 7: e1002383.

32. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6: e19379.

33. Bradbury PJ (in prep.)

34. SAS Institute (2004) SAS/STAT user's guide. Version 9.2. SAS Inst., Cary, NC

35. R D (2011) R: A language and environment for statistical computing.

36. Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal 4: 250.

37. Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2005) ASReml User Guide.

38. Austin D, Lee M (1996) Comparative mapping in F2:3 and F6:7 generations of quantitative trait loci for grain yield and yield components in maize. Theoretical and Applied Genetics 92: 817–826.

39. Veldboom LR, Lee M (1996) Genetic Mapping of Quantitative Trait Loci in Maize in Stress and Nonstress Environments: II. Plant Height and Flowering. Crop Science 36: 1320–1327.

40. Sari-Gorla M, Krajewski P, Di Fonzo N, Villa M, Frova C (1999) Genetic analysis of drought tolerance in maize by molecular markers. II. Plant height and flowering. TAG Theoretical and Applied Genetics 99: 289–295.

41. Khanal R, Earl H, Lee E a., Lukens L (2011) The Genetic Architecture of Flowering Time and Related Traits in Two Early Flowering Maize Lines. Crop Science 51: 146.

42. Lee E, Tracy W (2009) Modern maize breeding. Handbook of Maize II.

43. Messmer R, Fracheboud Y, Bänziger M, Vargas M, Stamp P, et al. (2009) Drought stress and tropical maize: QTL-by-environment interactions and stability of QTLs across environments for yield components and secondary traits. TAG Theoretical and applied genetics 119: 913–930.

44. Meuwissen T, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 49.

45. Heffner EL, Sorrells ME, Jannink J (2009) Genomic Selection for Crop Improvement. Crop Science 49: 1–12.

46. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, et al. (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nature genetics 44: 217–220.

47. Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic Selection in Plant Breeding: A Comparison of Models. Crop Science 52: 146.

48. Massman JM, Gordillo A, Lorenzana RE, Bernardo R (2012) Genomewide predictions from maize single-cross data. TAG Theoretical and applied genetics DOI 10.1007/s00122-012-1955-y

49. Freeman D, Pisani R, Purves R (2007) Statistics. 4th ed.

SUPPLEMENTAL MATERIAL

Supplement table 3.1. Average phenotypic value for each population within environment, and average BLUE for each population.

	days to anthesis								
	NC10	MO10	WL10	IA10	NC11	IA11	IN11	NY11	BLUE
Pop1	NA	52.7	60.9	81.5	61.2	71.7	66.2	70.4	68.6
Pop2	NA	53.9	62.5	83.0	64.2	75.4	69.7	72.4	70.2
Pop3	NA	54.9	63.5	84.8	64.5	75.7	70.2	75.6	73.5
Pop4	NA	55.3	64.4	85.1	64.8	76.4	68.5	75.0	71.9
Pop5	NA	55.4	64.7	84.9	65.0	75.9	69.8	76.3	69.1
Pop6	NA	54.7	63.0	84.0	62.3	73.7	67.6	73.6	70.6
Pop7	NA	55.9	63.8	84.6	64.6	75.5	68.5	74.4	71.6
Pop8	NA	55.7	65.9	86.0	65.5	77.8	69.8	76.0	73.1
Pop9	NA	55.3	63.9	84.5	65.0	75.6	68.8	74.5	74.0
Pop11	NA	NA	NA	NA	63.6	71.6	66.8	71.3	67.9
Pop12	NA	55.1	65.4	84.3	64.1	74.0	68.7	74.8	70.6
Pop13	NA	55.6	63.3	82.5	64.7	72.5	67.5	71.6	69.9
Pop14	NA	54.6	63.4	83.0	65.1	72.9	67.9	74.6	70.4
Pop15	NA	55.4	63.2	83.2	65.2	73.9	70.3	73.6	71.3
Pop16	NA	55.8	64.2	83.8	65.2	72.6	68.8	72.6	71.4
Pop18	NA	55.1	64.4	82.7	64.2	73.7	68.8	74.6	68.3
Pop19	NA	52.8	60.0	81.1	61.8	71.7	67.0	70.3	67.4
Pop20	NA	54.3	63.8	82.9	63.8	73.4	69.0	73.4	68.0
Pop21	NA	53.7	61.8	82.1	64.6	73.1	66.6	72.0	73.6
Pop22	NA	52.0	60.7	81.5	63.3	71.3	65.8	70.7	70.6
Pop23	NA	54.1	61.5	83.3	64.2	74.2	67.3	72.9	73.3
Pop24	NA	NA	NA	NA	63.5	70.6	63.9	70.5	65.5
Pop25	NA	54.2	62.5	83.5	63.3	74.0	69.5	72.7	72.9
Pop26	NA	54.3	63.8	83.7	63.5	73.9	68.1	73.0	72.1

	days to silk								
	NC10	MO10	WL10	IA10	NC11	IA11	IN11	NY11	BLUE
Pop1	NA	54.1	63.7	84.0	61.2	74.5	67.5	71.3	70.4
Pop2	NA	54.4	65.0	86.2	64.4	77.8	70.0	73.4	78.6
Pop3	NA	56.8	66.8	88.7	64.9	78.9	71.4	77.2	76.2
Pop4	NA	57.1	67.7	89.4	65.4	80.4	69.3	77.1	74.6
Pop5	NA	57.5	67.2	89.0	65.4	79.3	70.7	77.8	71.0
Pop6	NA	56.4	66.0	87.9	62.2	76.8	68.8	75.2	72.9
Pop7	NA	59.0	67.5	89.6	65.8	78.8	69.5	76.7	74.4
Pop8	NA	58.2	68.8	90.9	66.5	81.4	70.9	78.3	76.0
Pop9	NA	57.5	67.3	88.6	65.9	79.0	70.1	76.5	76.9
Pop11	NA	NA	NA	NA	64.0	74.1	68.8	72.3	69.5
Pop12	NA	57.7	68.7	89.7	64.5	78.7	70.3	77.4	73.5
Pop13	NA	57.5	66.4	87.0	65.7	75.9	69.1	72.9	72.8
Pop14	NA	57.0	67.1	87.0	65.8	76.2	68.8	76.1	72.8
Pop15	NA	57.1	65.7	86.4	65.5	76.0	70.8	75.0	73.0
Pop16	NA	58.2	67.4	88.2	66.1	75.8	70.1	73.9	74.2
Pop18	NA	58.2	67.9	87.9	64.9	79.5	70.4	77.0	71.1
Pop19	NA	54.7	62.3	83.1	62.2	74.1	68.4	71.2	68.9
Pop20	NA	56.3	67.1	87.2	64.3	76.9	70.6	75.5	70.4
Pop21	NA	55.2	64.9	84.7	65.3	75.9	67.2	72.8	75.7
Pop22	NA	53.3	63.4	83.5	63.6	73.9	66.9	72.2	72.3
Pop23	NA	55.9	65.1	87.2	64.7	77.3	68.5	74.5	76.0
Pop24	NA	NA	NA	NA	64.2	72.7	64.9	72.0	66.8
Pop25	NA	56.6	65.6	87.8	63.6	77.8	70.5	74.4	75.8
Pop26	Na	56.7	66.5	87.5	64.0	76.3	69.9	74.8	74.5

Supplement table 3.1. Continue.

	Plant height (cm)								
	NC10	MO10	WL10	IA10	NC11	IA11	IN11	NY11	BLUE
Pop1	197.2	272.2	249.6	260.7	NA	229.1	267.6	266.6	247.1
Pop2	208.9	276.4	249.1	263.5	242.1	239.6	261.5	252.7	253.4
Pop3	213.1	280.4	259.1	249.5	252.7	256.2	253.1	246.8	253.1
Pop4	186.3	277.5	247.1	257.0	247.3	255.7	266.8	223.3	244.8
Pop5	205.7	280.1	243.9	249.0	246.4	249.0	276.2	245.4	253.3
Pop6	225.8	264.9	258.2	258.9	266.8	258.5	277.2	235.9	258.9
Pop7	223.7	263.3	257.8	252.8	242.4	251.4	287.7	240.4	254.7
Pop8	206.3	272.5	261.9	262.8	254.1	258.5	282.0	240.3	252.6
Pop9	221.2	270.7	253.3	262.0	236.5	243.9	279.3	243.4	254.9
Pop11	NA	NA	NA	NA	235.8	238.1	255.7	265.0	249.2
Pop12	209.8	277.1	241.0	263.2	253.1	263.6	277.6	245.7	255.9
Pop13	205.2	248.1	237.4	253.0	209.3	247.3	266.8	248.5	241.4
Pop14	209.7	284.1	252.6	266.3	246.8	255.6	286.2	255.0	254.6
Pop15	193.8	269.2	259.4	260.6	234.7	239.4	259.4	241.2	243.0
Pop16	197.3	267.5	256.3	264.2	246.9	273.5	282.4	268.3	252.1
Pop18	177.8	269.3	240.4	250.5	244.7	252.3	271.0	249.6	249.6
Pop19	203.6	257.1	248.9	245.1	229.8	221.8	246.9	230.8	232.2
Pop20	206.9	276.3	244.6	251.8	257.3	257.5	270.6	248.0	253.7
Pop21	207.6	271.9	237.1	248.6	219.2	228.2	271.5	246.1	242.1
Pop22	201.4	278.0	229.5	255.9	218.6	241.4	268.5	245.3	248.7
Pop23	222.1	277.1	256.8	249.4	244.6	235.5	278.2	249.2	251.2
Pop24	NA	NA	NA	NA	219.7	221.3	261.4	240.2	243.5
Pop25	216.8	284.0	266.4	259.8	266.8	270.1	278.1	265.8	258.8
Pop26	198.0	270.2	248.0	256.0	246.8	244.1	277.9	234.7	251.1

	Yield (T/ha)								
	NC10	MO10	WL10	IA10	NC11	IA11	IN11	NY11	BLUE
Pop1	4.5	7.9	8.5	7.4	8.3	6.9	NA	NA	7.0
Pop2	5.2	6.1	8.6	6.7	7.4	5.5	NA	NA	6.5
Pop3	5.1	7.4	8.5	6.5	7.1	5.9	NA	NA	6.7
Pop4	4.6	6.3	8.5	5.8	6.8	5.4	NA	NA	6.0
Pop5	5.3	6.0	8.0	6.3	7.6	5.5	NA	NA	6.3
Pop6	4.6	6.5	8.3	6.9	8.0	6.7	NA	NA	6.9
Pop7	4.9	5.5	8.4	6.0	6.4	5.8	NA	NA	6.0
Pop8	5.1	6.5	7.9	4.9	6.2	4.9	NA	NA	5.7
Pop9	4.2	5.7	8.5	5.9	NA	5.0	NA	NA	5.7
Pop11	NA	NA	NA	NA	NA	6.4	NA	NA	NA
Pop12	4.8	5.8	7.0	5.1	NA	5.3	NA	NA	5.3
Pop13	5.5	6.8	8.6	7.5	5.7	7.5	NA	NA	6.8
Pop14	4.6	6.5	7.8	6.9	NA	6.6	NA	NA	6.7
Pop15	4.7	6.0	9.1	7.5	6.4	6.9	NA	NA	6.8
Pop16	4.3	5.5	8.2	7.1	NA	7.2	NA	NA	6.4
Pop18	4.6	6.2	8.2	6.7	8.2	5.2	NA	NA	6.4
Pop19	3.9	7.5	8.2	7.8	6.7	6.7	NA	NA	6.6
Pop20	5.3	6.6	8.7	6.7	8.4	6.5	NA	NA	7.2
Pop21	5.4	7.3	8.5	7.1	NA	5.6	NA	NA	6.6
Pop22	4.7	6.9	8.4	8.0	6.0	6.6	NA	NA	6.9
Pop23	4.0	6.9	8.4	7.1	NA	5.6	NA	NA	6.0
Pop24	NA	NA	NA	NA	NA	5.6	NA	NA	NA
Pop25	5.0	6.2	8.3	6.0	NA	6.1	NA	NA	6.1
Pop26	4.8	6.5	8.5	6.1	NA	6.2	NA	NA	6.3

Supplement table 3.2. Average phenotypic values across environments for female inbreds, male inbred, hybrid, mid-parent heterosis, and best-parent heterosis.

	Plant height (cm)				
	Female Inbred	Male Inbred	Hybrid	Mid-parent	Best-Parent
Pop1	173.0	171.0	247.1	75.1	74.0
Pop2	166.1	171.0	253.4	84.8	82.4
Pop3	170.0	171.0	253.1	82.6	82.1
Pop4	167.4	171.0	244.8	75.6	73.8
Pop5	176.1	171.0	253.3	79.8	77.2
Pop6	165.3	171.0	258.9	90.7	87.9
Pop7	187.7	171.0	254.7	75.3	67.0
Pop8	188.4	171.0	252.6	72.7	64.1
Pop9	182.9	171.0	254.9	78.0	72.0
Pop11	178.3	171.0	249.2	74.5	70.9
Pop12	168.7	171.0	255.9	86.0	84.9
Pop13	158.8	171.0	241.4	76.5	70.4
Pop14	180.3	171.0	254.6	78.9	74.3
Pop15	179.9	171.0	243.0	67.5	63.1
Pop16	184.2	171.0	252.1	74.5	67.9
Pop18	173.3	171.0	249.6	77.4	76.3
Pop19	155.1	171.0	232.2	69.1	61.2
Pop20	155.0	171.0	253.7	90.6	82.7
Pop21	157.3	171.0	242.1	78.0	71.1
Pop22	174.3	171.0	248.7	76.0	74.3
Pop23	167.1	171.0	251.2	82.2	80.2
Pop24	145.7	171.0	243.5	85.1	72.5
Pop25	190.3	171.0	258.8	78.2	68.6
Pop26	169.0	171.0	251.1	81.1	80.1

	Days to anthesis				
	Female Inbred	Male Inbred	Hybrid	Mid-parent	Best-Parent
Pop1	73.8	67.0	68.7	-1.8	-5.2
Pop2	78.3	67.0	70.2	-2.4	-8.1
Pop3	80.0	67.0	73.5	0.0	-6.5
Pop4	80.2	67.0	71.9	-1.8	-8.3
Pop5	80.1	67.0	69.1	-4.5	-11.1
Pop6	78.2	67.0	70.6	-2.0	-7.7
Pop7	78.0	67.0	71.6	-0.9	-6.4
Pop8	82.1	67.0	73.1	-1.4	-9.0
Pop9	79.0	67.0	74.0	0.7	-5.0
Pop11	72.4	67.0	67.9	-1.8	-4.5
Pop12	79.6	67.0	70.6	-2.7	-9.0
Pop13	77.7	67.0	69.9	-2.4	-7.7
Pop14	76.7	67.0	70.4	-1.7	-6.3
Pop15	78.3	67.0	71.3	-1.3	-7.0
Pop16	77.4	67.0	71.4	-0.8	-6.0
Pop18	79.4	67.0	68.3	-4.9	-11.1
Pop19	71.8	67.0	67.4	-2.1	-4.5
Pop20	78.3	67.0	68.0	-4.7	-10.3
Pop21	76.0	67.0	73.6	2.1	-2.4
Pop22	73.2	67.0	70.6	0.4	-2.6
Pop23	76.8	67.0	73.3	1.1	-3.5
Pop24	70.9	67.0	65.5	-3.7	-5.4
Pop25	77.3	67.0	72.9	0.8	-4.4
Pop26	78.3	67.0	72.1	-0.6	-6.2

	Days to silk				
	Female Inbred	Male Inbred	Hybrid	Mid-parent	Best-Parent
Pop1	76.0	72.0	70.4	-3.6	-5.6
Pop2	78.6	72.0	71.8	-3.5	-6.8
Pop3	82.0	72.0	76.2	-0.8	-5.8
Pop4	82.0	72.0	74.6	-2.4	-7.4
Pop5	81.9	72.0	71.0	-6.0	-10.9
Pop6	79.6	72.0	72.9	-2.9	-6.7
Pop7	80.3	72.0	74.4	-1.7	-5.8
Pop8	83.8	72.0	76.0	-1.9	-7.8
Pop9	81.0	72.0	76.9	0.4	-4.1
Pop11	74.0	72.0	69.5	-3.5	-4.6
Pop12	82.4	72.0	73.5	-3.7	-8.9
Pop13	80.0	72.0	72.8	-3.2	-7.2
Pop14	78.8	72.0	72.8	-2.6	-6.0
Pop15	79.5	72.0	73.0	-2.8	-6.5
Pop16	78.9	72.0	74.2	-1.2	-4.7
Pop18	82.1	72.0	71.1	-5.9	-10.9
Pop19	72.5	72.0	68.9	-3.4	-3.9
Pop20	80.3	72.0	70.4	-5.7	-9.9
Pop21	77.7	72.0	75.7	0.8	-2.1
Pop22	74.4	72.0	72.3	-0.9	-2.1
Pop23	79.2	72.0	75.9	0.3	-3.3
Pop24	72.4	72.0	66.8	-5.4	-6.2
Pop25	79.9	72.0	75.8	-0.2	-4.2
Pop26	81.3	72.0	74.5	-2.1	-6.8

Supplement table 3.3. T-test for recombination rate at QTL intervals for respective trait.

		Plant height (cm)			
		Inbred	**Hybrid**	**Best-parent**	**Mid-parent**
Plant Height	**Inbred**	1.000			
	Hybrid	0.878	1.000		
	Best-parent	0.301	0.332	1.000	
	Mid-parent	0.328	0.350	0.868	1.000
Days to anthesis	**Inbred**	0.764	0.676	0.262	0.298
	Hybrid	0.506	0.462	0.212	0.254
	Best-parent	0.060	0.104	0.145	0.200
	Mid-parent	0.996	0.927	0.345	0.348
Days to silk	**Inbred**	0.488	0.624	0.438	0.429
	Hybrid	0.617	0.558	0.236	0.276
	Best-parent	0.199	0.215	0.793	0.955
	Mid-parent	NA	NA	NA	NA
		Days to anthesis			
		Inbred	**Hybrid**	**Best-parent**	**Mid-parent**
Plant Height	**Inbred**				
	Hybrid				
	Best-parent				
	Mid-parent				
Days to anthesis	**Inbred**	1.000			
	Hybrid	0.638	1.000		
	Best-parent	**0.054**	0.513	1.000	
	Mid-parent	0.893	0.706	0.517	1.000
Days to silk	**Inbred**	0.322	0.237	**0.029**	0.692
	Hybrid	0.790	0.825	0.283	0.797
	Best-parent	0.178	0.145	0.115	0.229
	Mid-parent	NA	NA	NA	NA
		Days to silk			
		Inbred	**Hybrid**	**Best-parent**	**Mid-parent**
Plant Height	**Inbred**				
	Hybrid				
	Best-parent				
	Mid-parent				
Days to anthesis	**Inbred**				
	Hybrid				
	Best-parent				
	Mid-parent				
Days to silk	**Inbred**	1.000			
	Hybrid	0.275	1.000		
	Best-parent	0.282	0.161	1.000	
	Mid-parent	NA	NA	NA	NA

CHAPTER 4

GENETIC ANALYSIS OF LODGING IN DIVERSE MAIZE HYBRIDS

ABSTRACT

Damage caused by lodging is a significant problem in corn production that results in an estimated annual yield loss of 5-20%. Over the past 100 years, substantial maize breeding efforts have increased lodging resistance by artificial selection. However, less research has focused on understanding the genetic architecture underlying lodging. Lodging is a problematic trait to evaluate since it is greatly influenced by environmental factors such as wind, rain, and insect infestation, which make replication difficult. In this study over 1,723 diverse inbred maize genotypes were crossed to a common tester and evaluated in five environments over multiple years. Natural lodging due to severe weather conditions occurred in all five environments. By testing a large population of maize diversity in multiple field environments, we detected significant correlations for this highly environmentally influenced trait across genotypes grown in multiple environments and with other important agronomic traits such as yield and plant height. This study also permitted mapping, and the identification of QTL for lodging. A number of identified QTL overlapped with QTL previously mapped for stalk strength using nearly the same maize diversity measured in an inbred state. QTL intervals mapped in this study also overlapped candidate genes in the lignin and cellulose pathways.

INTRODUCTION

An important criterion necessary when advancing genotypes in modern maize breeding programs is that they are machine-harvestable, i.e. resistant to lodging. Lodging is a combination of a plant's inability to keep an upright position and weather conditions, such as rain and wind. Often, lodging is classified into occurring at the stalk and occurring at the root [1]. Extensive breeding efforts have been made to develop lines with increased lodging resistance [2]. However, due to higher planting densities, higher soil fertility levels, and ever changing environmental factors, lodging remains an important criterion in maize improvement. Lodging is a major problem in corn production causing harvest difficulties and resulting in annual yield losses of 5 to 20% [3–5].

During a plant's growing stage (V5-V8 and V12-R1), the rapid growth of the internodes weakens the cell walls increasing the probability for the stalk to break when exposed to strong winds [6]. When the plant has reached mature height stalk lodging risks are moderate as lignin and other structural material strengthen the cell walls and the stalk [7]. Stalk lodging can also occur later in the season near harvest when the ear is fully developed and heavier and the stalk cannot support it. The weakness of the stalk later in the season is affected by insect infestation (i.e. European corn borer, Ostrinia nubilalis Hubner), and stalk rot. In both cases stalk breakage occurs at the node below or above the primary ear, and most often results in no or very low yield, due to loss of the ear or lack of photosynthetic surface area [6].

Root lodging in maize is affected by root characteristics including the number of roots on upper internodes, total root volume, root angle from vertical, and diameter of roots [8]. With a weakened root system the plant is prone to wind damage resulting in snapping or buckling of the stalk at the base of the plant, or roots being pulled out of the soil. The risk for root lodging is highest during the mid-vegetative stage before brace roots are fully developed [8]. Root lodging early in the season is not devastating since plants can regain upright growing pattern due to its plasticity within a week with no negative effect on yield. This is not the case after the plant has reached full maturity [9].

Evaluating genotypes for lodging is a complex process, given the influences of environmental factors that are not easy to control or replicate. One method useful for stalk lodging is stalk crushing. The disadvantages to stalk crushing are its destructive nature and that it is time and labor expensive [10]. Stalk crushing is highly correlated with rind puncture resistance (RPR) [11]. RPR measures the kilograms of force required to penetrate the stalk rind using a spike connected to a force gauge [12]. A third strategy implemented to indirectly select for stalk lodging is stalk water content. It is an indicator of stay-green, which is a sign of increased photosynthetic activity in older vegetative parts [11]. Lastly, the use of near infrared (NIR) analysis is being used to measure cellulose and lignin [13].

Root lodging is most often evaluated as the proportion of lodged plants per plot. A plant is considered to be root lodged when it is tilting >30° (e.g. [14,15]). Alternative ways to phenotype susceptibility to root lodging is by vertical-pull resistance [8], measure of root volume by water replacement, or recording weight of the root clump

[16]. To evaluate both stalk and root lodging under controlled wind conditions DuPont Pioneer has developed a mobile wind machine that can generate winds up to 100 mph [17].

Given the apparent relationships between lodging and traits such as stay-green, it is a concern that selection for improved stalk strength may cause undesirable indirect selection on other agronomic traits. Relationships between stalk strength and other agronomic traits are unclear based on existing literature. Selection for increased RPR has reduced stalk lodging [18–20]. However, there exist disadvantages to selecting genotypes with increased rind thickness and higher stalk lodging resistance. Thicker rinds may divert limited carbohydrates from kernel fill. This has the potential to result in lower yields [21]. Studies have reported a negative correlation between increased stalk strength resistance and decrease in grain yield [22,23]. To the contrary, Colbert et al. (1984) [24] found a significant positive relationship, while still other studies have observed no significant correlation between increases in stalk strength and other morphological traits [11]. This suggests the relationships observed among traits may be strongly dependent upon both the germplasm surveyed and the environment in which it was observed.

A number of studies have been performed to better understand the genetic architecture underlying lodging. Stalk strength has for example been evaluated as RPR, NIR and mechanical strength and genetically mapped. Overall, these findings suggest stalk strength is a highly complex trait controlled by a large number of alleles, each with small effects, and loci are not necessarily shared among different populations [4,6,12,13,25].

Few QTL have been mapped for root lodging. One of the QTL identified to control root lodging is the root-ABA1 QTL on chromosome 2 [14]. Moreover, QTL have been mapped for a number of root traits correlated with root lodging (e.g. [26]). One explanation for the few studies to identify QTL can be the difficulties to evaluate the trait and to replicate experiments in multiple locations.

Lignin and cellulose content has been shown to influence stalk strength. Lignin influences stalk strength and stiffness [7], but also root lodging due to the function of lignin in cell elongation [9]. Lignin is structures like xylem (transport for water), sclerenchyman and bundle sheath cells. Natural variation in lignin has close to no influence on the phenotype [7]. However, the *brown midrib* genes (*bm1 and bm3*) alter the lignin composition resulting in weaker stalks [27]. These genes encode cinnamyl alcohol dehydrogenease (CAD) and a caffeic O-methyl transferase (COMT) [28]. Cellulose is located in the vascular bundles and is referred to as the rind. A number of the *CesA* genes in the cellulous pathway are involved in secondary wall formation, leading to stronger stalks [29].

In this study we examined hybrids from crosses of recombinant inbred lines (RILs) of the NAM population [30] to the male tester, PHZ51. These hybrids were grown in five different field environments and their progenitors were genotyped. All five fields were naturally exposed to unique environmental factors and each possessed substantial variation for lodging damage among its genotypes. To relate this lodging variation to genetic diversity, we employed joint linkage mapping. Range of QTL were identified for lodging, some that overlapped with previously identified loci for lodging, but also a number of new QTL.

MATERIAL AND METHOD

Germplasm

In this study we used hybrids of the maize nested association mapping (NAM) population developed by the Genetic Architecture of Maize and Teosinte Project consortium [30] crossed to the male tester PHZ51. NAM was created by selecting 25 inbreds to maximize diversity, and crossing them to the reference inbred, B73. From each of the 25 subfamilies, 200 progeny were chosen self pollinated for five generations and subsequently sib-mated. This resulted in a mapping population of about 5,000 RILs. From these RILs, we selected a subset of 60-70 lines of each subfamily (except the popcorn subfamily Hp301) and created hybrids by crossing to the male tester PHZ51, the non-stiff stalk line developed by DuPont Pioneer with an expired Plant Variety Protection (ex-PVP). The selection of the subset of NAM female inbreds for hybrid development was based upon flowering time. The earliest RILs from the late families and the latest RILs from the early families were selected to reduce flowering time variation and make hybrid production by isolation plots manageable.

Phenotypic Evaluation

Hybrids were grown and evaluated in Sandhills NC, Bradford MO, West Lafayette IN, and Slater IA in the summer of 2010, as well as Kinston NC, Bradford MO, West Lafayette IN, and Ames IA in the summer of 2011. All environments were cultivated in a conventional manner with respect to fertilization, weed, and pest management.

Hybrids were planted in two-row plots with a single replication per environment. The experiments were blocked by subfamily to avoid competition for space and light interception resulting from variation in height. Hybrids were randomized within blocks, and blocks were randomized within environments.

All phenotypic data were collected on a plot basis. Days from planting until half the plants in a plot shed pollen or had a visible silk was used as the criterion to measure days to anthesis and days to silk, respectively. All other traits were measured at full maturity after flowering. Plant height was measured as the distance from the soil line to the base of the flag leaf, and ear height as the distance from the soil line to the node of the primary ear. Leaf length and width was measure as the maximum length and width of the leaf below the primary ear. Numbers of nodes was divided into the number of nodes from the soil line to the node of the primary ear, and from the node above the primary ear to the tassel. Root lodging was determined as the fraction of lodged plants within a plot. A plant was determined as lodged when a tilt of 30 degree or greater was observed. Stalk lodging was measured as the proportion of plants in a plot with a broken stalk. Yield was measured using a two-row combine and moisture was measured automated on the combine. Yield was then adjusted to 15.5% moisture content and expressed in tons per hectare.

Genotypic Evaluation

Genotypic data for joint linkage mapping was collected as previously described [30,31]. In total, 1,106 markers were scored on an Illumina GoldenGate Assay across the NAM RILs.

The NAM RILs were genotyped at low coverage at about 80% using the GBS platform [32]. As a result of the relatively low coverage, about 60% of the data for individual markers was missing and about 80% of the heterozygous loci were called as homozygotes. An HMM algorithm was used to correct the heterozygote calls to about 99.8% accuracy [33]. Following, a set of markers to be used in fitting a regression model was imputed from the GBS data at 0.2 cM intervals based on flanking markers. The marker values identified the parent of origin, with B73 coded as 0 and the non-B73 parent coded as 2. Heterozygous loci were coded as 1. For the imputed sites, if the flanking markers were identical, the site was set to that value. If the flanking markers were not identical, an interpolated value was used based on the relative distance from each of the flanking markers. The 7,389 imputed marker intervals were used to estimate marker effects across the genome using ridge regression.

Statistical Analysis

Statistical analyses were performed using SAS [34] software, and R [35] scripts and libraries. Best Linear Unbiased Estimations (BLUEs) for all phenotypes across environments were calculated as LSmeans with range position as fixed effect for each genotype within block using SAS software.

After deriving BLUE, the base package in R was used to calculate Pearson correlation coefficients and to study relationships between phenotypes at the genotypic level across and within both environments and NAM families.

To better characterize genetic architecture, joint linkage mapping of BLUEs across and within environments was performed using proc GLMSelect in SAS. After

adjusting for differences between NAM subfamily, an imputed set of 1,106 markers were nested within each NAM subfamily and regressed against BLUEs for each of the phenotypes across and within environments in a stepwise manner, as described in Buckler et al. (2009) [31]. For the stepwise procedure, model inclusion and exclusion of family nested markers were discerned by comparison with a null distribution based on permutation testing. The *p*-value derived from the null distribution for model inclusion was 0.001 at an alpha level of 0.05. Joint linkage mapping was performed with both including and excluding flowering time in the model.

To assess our ability to predict breeding values from all assayed genetic diversity, we performed genomic prediction by ridge regression BLUP in R using the rrBLUP package [36]. Genomic relationship matrices were calculated using 7,389 marker intervals for the NAM population. To calculate genomic estimated breeding values (GEBVs) for each phenotype, genetic matrices were regressed against each phenotypic BLUE. Accuracy of the prediction was determined by Pearson correlation of the predicted GEBVs and the observed breeding values. This was accomplished by fitting all genotypes into a ridge regression model. To assess overfitting and the robustness of the modeling approach, the population was randomly divided into five mutually exclusive subpopulations (five folds), and one of the subpopulations was excluded and the other four were used to construct a model which was then fitted to all the subpopulations. Accuracy of the prediction was averaged across the five folds. This process was repeated with 20 different random partitions of 5 subpopulations for each trait. This allowed us to more robustly estimate the average prediction accuracy.

RESULTS

Phenotypic evaluation

In this study we examined five natural occurring lodging events by evaluating the same hybrid NAM population in five unique field environments. Due to weather conditions, environments possessed substantial variation for root and stalk lodging (Table 4.1). In 2010, Bradford MO and Slater IA environments underwent lodging at flowering. That same year, the Sandhills NC environment lodged after flowering. In 2011, the Bradford MO environment underwent lodging early in the season, before flowering and West Lafayette IN lodged late in the season after flowering. In analyses detailing lodging conditional on flowering stage, phenotypes from Bradford MO and Slater IA in 2010 were grouped to define lodging occurring at flowering. Using the same principal, data from Sandhills NC in 2010 and West Lafayette IN in 2011 were grouped to detail lodging occurring after flowering.

Table 4.1. Date of planting and storm events, and information on weather conditions (Steremberg 2012). As well as, percent of plots per environment damaged by lodging. GDDs are calculated with a base temperature of 10 C.

	IA10	IN11	MO10	MO11	NC10
planting date	4/22/10	5/10/11	5/27/10	5/10/11	4/21/10
date of lodging	7/18/10	8/13/11	7/18/10	7/3/11	late season
GDD	830	1291	785	651	NA
mean temp	75 F	76 F	80 F	78 F	NA
precipitation	0.50 in	0.84 in	1.02 in	1.29 in	NA
max gust speed	71 mph	59 mph	57.5 mph	52 mph	NA
% damaged plots root	96	78	75	6	7
% damaged plots stalk	11	87	67	99	83
% damaged plots total	96	98	91	99	85
% damaged plants root	21	11	18	0.2	12
% damaged plants stalk	0.6	48	10	41	0.3
% damaged plants total	21	59	28	41	13

All five environments experienced substantial proportion of lodging (Table 4.1). From 85-99 percent of the plots in the five environments had one or more plants lodged. In the MO11 environment, which was damaged by a storm early in the season, the majority of the lodging was stalk lodging. The same pattern was observed in the NC10 environment. IA10 on the other hand had a high proportion of root lodging and low proportion of stalk lodging. MO10 and IN11 had relative high percent of both root lodging.

Trait correlations:

Both root and stalk lodging are highly influenced by environmental factors such as weather conditions, especially wind and water. Correlations of root lodging among the five locations were strongest, 0.31, between the environments IA10 and MO10 (Figure 4.1). The WL11 environment shows correlation with MO10 and IA10. The NC10 environment that was exposed to late season lodging shows negative correlation for root lodging with the other four environments. Overall this environment did not possess much root lodging, since the probability for root lodging to occur is higher earlier in the growing season. For stalk lodging, the strongest correlations are found between IA10 and MO10 (0.17), NC10 and WL11 (0.25), as well as MO11 and WL11 (0.21). All correlations were statistically significant (p-value = <0.0001). For total lodging the two highest correlations are between MO10 and IA10, and MO11 and WL11. MO10 and IA10 are locations that were exposed to severe weather conditions at the same date, and at the around the same stage of development.

	RootLodg-MO10	RootLodg-IA10	RootLodg-NC10	RootLodg-WL11	RootLodg-MO11
RootLodg-MO10		0.31	-0.02	0.18	0.02
RootLodg-IA10	<0.0001		-0.04	0.23	0.00
RootLodg-NC10	0.5802	0.2170		-0.06	-0.01
RootLodg-WL11	<0.0001	<0.0001	0.0542		0.05
RootLodg-MO11	0.4893	0.9560	0.8674	0.0598	

	StalkLodg-MO10	StalkLodg-IA10	StalkLodg-NC10	StalkLodg-WL11	StalkLodg-MO11
StalkLodg-MO10		0.17	0.00	0.03	-0.05
StalkLodg-IA10	<0.0001		-0.07	0.02	-0.02
StalkLodg-NC10	0.9615	0.3410		0.25	0.07
StalkLodg-WL11	0.2950	0.6552	<0.0001		0.21
StalkLodg-MO11	0.1303	0.4865	0.017	<0.0001	

	TotalLodg-MO10	TotalLodg-IA10	TotalLodg-NC10	TotalLodg-WL11	TotalLodg-MO11
TotalLodg-MO10		0.28	0.08	0.16	0.00
TotalLodg-IA10	<0.0001		0.06	0.17	0.09
TotalLodg-NC10	0.0079	0.0720		0.20	0.08
TotalLodg-WL11	<0.0001	<0.0001	<0.0001		0.21
TotalLodg-MO11	0.9394	0.0066	0.0073	<0.0001	

Figure 4.1. Correlation between root lodging, stalk lodging, and total lodging across the five environments. Upper right half of the tables report r-value and lower left half reports *p*-values. Coloring indicates direction of the correlation where bright red = 1 and dark blue = -1.

Environments were grouped by the time of lodging with respect to flowering time, i.e., if the lodging event occurred before, at, or after time of flowering. The middle environments where lodging occurred at flowering (MO10 and IA10), and late environments where lodging occurred after flowering (NC10 and WL11) showed the expected patterns of correlation. High correlations were observed between flowering traits (days to silk and days to anthesis), and plant and ear height. For the middle environments, negative correlation between the traits and yield, especially the lodging traits, those have a higher negative correlation (Figure 4.2).

	D2A	D2S	Ear Height	Plant Height	Leaf Length	Leaf Width	Root Lodging	Stalk Lodging	Total Lodging	Yield
D2A		0.90	0.07	-0.15	0.11	0.03	0.18	0.06	0.18	-0.13
D2S	<0.0001		0.10	-0.14	0.17	0.06	0.18	0.07	0.18	-0.17
Ear Height	0.0167	0.0007		0.63	0.33	0.16	0.22	0.01	0.21	-0.21
Plant Height	<0.0001	<0.0001	<0.0001		0.28	0.07	0.18	-0.08	0.15	-0.14
Leaf Length	0.0003	<0.0001	<0.0001	<0.0001		0.16	0.14	0.07	0.15	-0.2
Leaf Width	0.3573	0.0721	<0.0001	0.0301	<0.0001		0.16	0.08	0.18	-0.12
Root Lodging	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001		0.06	0.90	-0.23
Stalk Lodging	0.0402	0.0185	0.7344	0.0047	0.0361	0.0149	0.0344		0.43	-0.21
Total Lodging	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001		-0.28
Yield	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0003	<0.0001	<0.0001	<0.0001	

Figure 4.2. Correlations between the three lodging traits and other developmental traits measured in the middle environments, where lodging occurred at flowering. Upper right half of the table reports r-value and lower left half reports *p*-values. Coloring indicates direction of the correlation where bright red = 1 and dark blue = -1.

The negative correlation between the lodging traits and yield can also been seen in the late environments (Figure 4.3). For the environments with a lodging event later in the season, there is also high correlation between the lodging traits and plant and ear height.

	DSA	D2S	Ear Height	Plant Height	Leaf Length	Leaf Width	Brace Roots	Nodes Ear To Tassel	Nodes Roots To Ear	Root Lodging	Stalk Lodging	Total Lodging	Yield
D2A		0.73	0.10	-0.06	0.22	-0.02	0.04	-0.01	0.17	-0.09	-0.11	-0.16	0.05
D2S	0.1798		0.16	-0.05	0.21	0.02	0.04	-0.06	0.23	-0.05	-0.12	-0.14	0.07
Ear Height	0.1462	<0.0001		0.75	0.15	0.09	0.01	0.10	0.50	0.29	0.40	0.50	0.09
Plant Height	0.7157	0.0005	<0.0001		0.05	-0.04	0.05	0.31	0.40	0.41	0.43	0.60	0.15
Leaf Length	0.0002	<0.0001	<0.0001	<0.0001		0.09	0.10	-0.02	0.09	-0.16	-0.04	-0.13	0.08
Leaf Width	0.4019	0.4810	0.5355	0.0004	0.0003		-0.02	-0.07	0.02	-0.13	-0.05	-0.12	0.09
Brace Roots	0.0034	0.8139	0.0404	0.0002	0.4572	0.0045		0.08	0.20	-0.08	-0.04	-0.08	0.04
Nodes Ear To Tassel	<0.0001	<0.0001	<0.0001	<0.0001	0.0007	0.4145	0.3058		-0.03	0.20	0.08	0.19	0.00
Nodes Roots To Ear	0.0645	0.0432	0.0788	<0.0001	0.0354	0.0941	<0.0001	<0.0001		0.18	0.21	0.28	0.05
Root Lodging	0.0029	0.0017	0.0838	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001		-0.05	0.51	0.03
Stalk Lodging	0.1348	0.0001	<0.0001	<0.0001	0.1433	0.0365	0.0019	<0.0001	<0.0001	0.0675		0.83	-0.19
Total Lodging	0.0039	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001		-0.15
Yield	0.1887	0.1427	0.0331	0.0031	0.0089	0.0017	0.8719	0.1047	<0.0001	0.3072	<0.0001	<0.0001	

Figure 4.3. Correlations between the three lodging traits and other developmental traits measured in the late environments, where lodging occurred after flowering. Upper right half of the table reports r-value and lower left half reports *p*-values. Coloring indicates direction of the correlation where bright red = 1 and dark blue = -1.

Genotypes within single environments were grouped by percentage of lodging damage. The larger proportion of damaged plants results in lower average yield among the grouped genotypes (Table 4.2). Genotypes with both low and high level of resistance to lodging have the ability to high yields in good season environments

(Figure 4.4). Correlation between percentage of total lodging and yield was -0.10 with a *p*-value of 0.0009.

Table 4.2. Average yield in T/ha of genotypes grouped according to percentage of lodging damage per plot for individual environments.

	Average yield		
Percent Lodging	MO10	IA10	NC10
0 - 10	6.65	6.72	4.82
11 - 20	6.01	6.50	4.51
21 - 30	5.34	5.96	4.27
31 - 40	4.04	6.07	3.73
41 - 50	1.30	5.90	3.57
51 - 60	1.12	4.48	3.13
61 - 70	0.00	1.96	1.81
71 - 80	0.00	Na	2.22
81 - 90	0.00	1.86	1.79
91 - 100	0.00	0.00	0.00

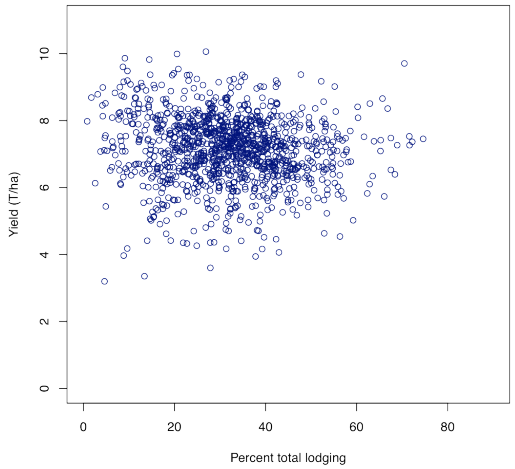


Figure 4. Total percentage of lodging from the five damaged environments regressed again yield evaluated in the three environments without significant lodging damage.

Joint linkage mapping

We performed joint linkage mapping for root lodging, stalk lodging, and total lodging using the 1,106 markers on the NAM map. Analyses were done for the grouped environments, middle and late, as well as for each single environment. Most of the QTL mapped within individual environments were shared across the grouped environments. For the two grouped environments, four QTL for stalk lodging, ten QTL for root lodging, and nine QTL for total lodging were identified (Figure 4.5, supplemental table 4.1). Joint linkage mapping was performed both excluding and including flowering time to account for stage of maturity at the time of the lodging event. Including flowering time in the model did not have a significant effect on the result of mapped QTL (data not shown).

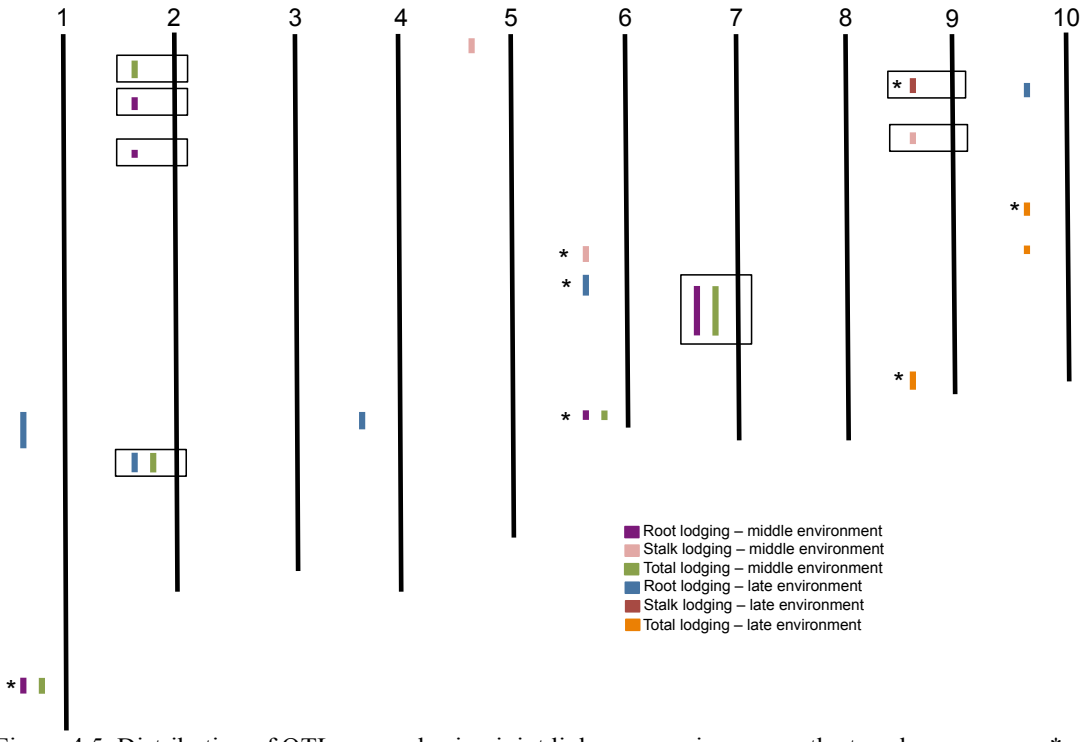


Figure 4.5. Distribution of QTL mapped using joint linkage mapping across the ten chromosomes. * indicates QTL intervals overlapping with results from stalk strength study in the NAM inbred population. Squares indicate QTL overlapping with candidate genes.

Genomic prediction

Genomic prediction using ridge regression was performed across the NAM population for both single environments and across grouped environments. For the middle and late environments we were able to predict the phenotype using all available data with an accuracy of 0.35-0.51 (Table 4.3). When randomly excluding a fifth of the phenotypic data and repeating this 20 times, the average accuracy decreases for each environment.

Table 4.3. Prediction accuracy for stalk lodging, root lodging, and total lodging within the early, middle and late environment. Top table presents accuracies calculated using all phenotypes into the model. Bottom table presents accuracies calculated with a five-fold cross validation.

Accuracy across all NAM using all genotypes in estimation set			
	Early (1 env)	Middle (2 env)	Late (2 env)
Stalk lodging	0.34	0.41	0.40
Root lodging	0.19	0.51	0.35
Total lodging	0.33	0.50	0.37

Accuracy across populations using 5-fold cross validation			
	Early (1 env)	Middle (2 env)	Late (2 env)
Stalk lodging	0.12	0.16	0.17
Root lodging	-0.06	0.34	0.14
Total lodging	0.12	0.33	0.16

DISCUSSION

Based on phenotypic data from five unique environments with natural lodging events we observed negative correlation between lodging traits and yield (Figure 4.2 and 4.3). It was also noticed that the higher proportion of damage the lower the yield (Table 4.2). A number of the genotypes with higher resistance perform more poorly than the susceptible genotypes in good seasons (Figure 4.4). This argues for a trade off

between breeding for higher yield or more resistant lines. It has been a concern for many years that breeding for lodging resistance by increased stalk strength will result in decreases in yield [22]. This reasoning is based on the sink-source relationship of available carbon in the plant. That is, if more carbon is used for stronger stalks there is less available carbon left overall in the plant for grain fill, which occurs later in the season [21].

In addition, in this study genotypes with less than 10 percent of the plot damaged had on average a lower yield than the corresponding genotypes evaluated in the environments with no significant damage. This suggests that damage related to lodging events can be much more than just the breakage of the plants, such as insect and pathogen infestation, interruption in xylem and phloem and, unbalance in carbon relocation. Proposing that a resistant line is higher yielding than a susceptible line in a damaged field, but this is not always true in a good season environment. Studies have to be performed to further investigate this relationship.

For the environments with late season lodging events, correlation between plant and ear height and lodging traits were observed (Figure 4.3). Similar relationship between height and lodging has been reported in previous studies (e.g. [4,25,37]). It is basic physics that higher ear placement and heavier ear in the late season in combination with either weaker roots or stalk will more likely result in lodging, compared to shorter plants with lower ear placement. In addition, there is a relationship between total plant height and yield [2] based on taller plants that have in general more biomass and thereby higher photosynthetic rate and more carbon fixation

that can be relocated to the ear as grain yield. Subsequently, it is not as simple to exclusively breed for shorter genotypes to avoid lodging.

In this study QTL for root, stalk, and total lodging were identified (Supplemental table 4.1). A smaller number of QTL were identified than previously mapped for stalk strength using RPR in the full set of NAM inbreds [12]. The inbred study identified 73 QTL clusters. The main reason for this is most like the difference in population size. This study only used about a third of the lines compared to the 5,000 RIL in the NAM population. Second, the traits in this study are caused by environmental conditions, which makes it more difficult to evaluate and replicate. Overall, it is suggested that the traits are controlled by a large number of loci with small effects. It is likely that we were only able to identify the QTL with the larger effects. However, we believe our results are robust.

Seven of the QTL mapped in this study are located in the same marker intervals on the NAM map as QTL identified by using RPR [12]. A large number of studies of lodging and stalk strength have been performed over time. We compared our results with a few of these studies representing different phenotyping strategies. Peiffer et al. (2012) measured stalk strength using RPR in the inbred population that was used as female in hybrid development for this study. Flint-Garcia et al. (2003) are two of the most extensive studies of stalk strength in maize using RPR. Ching et al. (2010) measured stalk strength using mechanical force, and Hu et al. (2012) used RPR and NIR. Overall, our results overlap with these previous studies (Table 4.4) [4,6,13,25]. In addition, the QTL for root lodging on chromosome 2 (97.2 – 98.9 cM)

is located in the same bin, 2.04, as the root-ABA QTL that has been mapped to influence root lodging [14].

Table 4.4. Mapped QTL and overlapping intervals with other lodging studies.

Trait	Chr	cM start	cM end	Peiffer et al. 2012	Ching et al. 2010	Flint-Garcia et al. 2003a	Flint-Garcia et al. 2003b	Hu et al. 2012
Root (late)	1	94.9	96.5			x		
Root / Total (middle)	1	175.7	176.9	x	x			x
Total (middle)	2	41.5	41.8	x				
Root (middle)	2	54.9	56.8					
Root (middle)	2	72.1	72.1					x
Root (late) / Total (middle)	2	97.5	98.9	x		x		
Root (late)	4	93.2	93.2					
Stalk (middle)	5	10.1	10.9					
Stalk (middle)	6	41.5	42.1	x				
Root (late)	6	43.4	43.7	x				
Root / Total (middle)	6	101	106.6	x			x	x
Root / Total (middle)	7	69.8	69.8					
Stalk (late)	9	44.5	45.2					
Stalk (middle)	9	47	47.2					
Total (late)	9	89.8	93.5				x	
Root (late)	10	40.1	40.6	x			x	
Total (late)	10	43.4	43.8				x	
Total (late)	10	44.7	44.8				x	

A list of 80 genes known to be involved in lignin synthesis, phenylpropanoid pathway, vegetative phase change and cellulose were compiled (Supplemental table 4.2). This list was compared to the mapped QTL in this study. Eleven of the 80 genes are located within 3 Mb from a lodging QTL. Six of the genes are located on chromosome 2, one on chromosome 4, two genes on chromosome 7, and two genes on chromosome 9. Four of the genes are involved in the lignin synthesis, one in the phenylpropanoid pathway, two are involved in the vegetative phase change, and three genes are in the cellulose synthesis pathway.

Using a very diverse population we predicted GEBVs for lodging with an accuracy of approximately 0.30 with a five-fold cross-validation (Table 4.3). This is a reasonable value considering the model is trained on data from only two environments. A prediction accuracy of 0.33 for total lodging in the middle environments is a good start of model development to predict lodging in maize. Particularly, considering that lodging is a highly environmentally impacted trait that was evaluated under multiple locations and years.

Since lodging is complex, both genetically and to evaluate, yet a highly valued trait, genomic prediction is believed to be very useful in breeding programs. Genomic prediction [38] allows selection from larger pools of diversity with reduced expenditure of phenotyping resources and field space. This will permit more rapid gains than would otherwise be feasible if the accuracy of selection were limited only to phenotyped individuals. Other industry studies in maize with data from a significantly larger number of more environments evaluating mostly bi-parental crosses of corn belt dent material have reported prediction accuracies for lodging of as high as 0.70 [39,40]. Hence, there is potential using genomic prediction in breeding programs to improve lodging traits in maize.

Here we have performed one of the largest studies in the public sector of natural lodging of diverse hybrids grown in five different environments. The diverse population and the greater LD decay have permitted mapping of QTL for root, stalk, and total lodging. A number of identified QTL overlap with previous studies on stalk strength. In addition, candidate genes influencing stalk and root composition are

located in the intervals. This study has given us a better understanding of natural lodging in maize, and a good beginning for genomic prediction of lodging.

REFERENCE

1. Lee E, Tracy W (2009) Modern maize breeding. Handbook of Maize II.

2. Duvick DN (2005) The contribution of breeding to yield advances in maize (Zea mays L .). Advances in Agronomy 86.

3. Zuber MS, MS K (1978) Corn lodging slowed by sturdier stalks. Crops and Soils 30.

4. Flint-Garcia S, McMullen MD, Darrah LL (2003) Genetic Relationship of Stalk Strength and Ear Height in Maize. Crop Science 43: 23.

5. National Agricultural Statistics Service (2012).

6. Ching A, Rafalski A, Luck S, Butruille MG (2010) Genetic loci associated with mechanical stalk strength in maize.

7. Pedersen JF, Vogel KP, Funnell DL (2005) Impact of Reduced Lignin on Plant Fitness. Crop Science 45: 812.

8. Ennos A (1993) The anchorage mechanics of maize, *Zea mays*. Journal of Experimental … 44: 147–153.

9. Zhang Q, Pettolino F a, Dhugga KS, Rafalski JA, Tingey S, et al. (2011) Cell wall modifications in maize pulvini in response to gravitational stress. Plant physiology 156: 2155–2171.

10. Zuber MS, Colbert TR, Darrah LL (1980) Effect of recurrent selection for crushing strength on several stalk componenets in maize. Crop Science 20.

11. Djordjevic JS, Ivanovic MR (1996) Genetic Analysis for Stalk Lodging Resistance in Narrow-Base Maize Synthetic Population ZPS14. Crop Science 36.

12. Peiffer JA et al. (in prep.) Genetic architecture of stalk strength. PloS one.

13. Hu H, Meng Y, Wang H, Liu H, Chen S (2012) Identifying quantitative trait loci and determining closely related stalk traits for rind penetrometer resistance in a high-oil maize population. TAG Theoretical and applied genetics 124: 1439–1447.

14. Landi P, Sanguineti MC, Liu C, Li Y, Wang TY, et al. (2007) Root-ABA1 QTL affects root lodging, grain yield, and other agronomic traits in maize grown under well-watered and water-stressed conditions. Journal of experimental botany 58: 319–326.

15. Bruce W, Desbons P, Crasta O, Folkerts O (2001) Gene expression profiling of two related maize inbred lines with contrasting root-lodging traits. Journal of experimental botany 52: 459–468.

16. Jenison J, Shank D, Penny L (1981) Root characteristics of 44 maize inbreds evaluated in four environments. Crop Science.

17. Barrerio R, Carrigan L, Ghaffarzadeh M, Goldman D, Michael H, et al. (2008) Device and method for screening a plant population for wind damage resistance traits.

18. Abedon BG, Darrah LL, Tracy WF (1999) Developmental Changes Associated with Divergent Selection for Rind Penetrometer Resistance in the MoSCSSS Maize Synthetic. Crop Science 39: 108–114.

19. Albrecht B, Dudley JW (1987) Divergent Selection for Stalk Quality and Grain Yield in an Adapted × Exotic Maize Population Cross. Crop Science 27: 487–494.

20. Dudley JW (1994) Selection for Rind Puncture Resistance in Two Maize Populations. Crop Science 34: 1458–1460.

21. Davis S, Crane PL (1976) Recurrent Selection for Rind Thickness in Maize and its Relationship With Yield, Lodging, and Other Plant Characteristtcs. Crop Science 16.

22. Rhen PN, Russell WA (1986) Indirect response in yield and harvest index to recurrent selection for stalk quality and corn-borer resistance in maize. Rev Brasil Genetics IX.

23. Martin MJ, Russell WA (1984) Correlated Responses of Yield and Other Agronomic Traits to Recurrent Selection for Stalk Quality in a Maize Synthetic. Crop Science 24: 746–750.

24. Colbert TR, Darrah LL, Zuber MS (1984) Effect of Recurrent Selection for Stalk Crushing Strength on Agronomic Characteristics and Soluble Stalk Solids in Maize. Crop Science 24.

25. Flint-Garcia S, Jampatong C, Darrah LL, McMullen MD (2003) Quantitative Trait Locus Analysis of Stalk Strength in Four Maize Populations. Crop Science 43: 13.

26. Hochholdinger F, Tuberosa R (2009) Genetic and genomic dissection of maize root development and architecture. Current opinion in plant biology 12: 172–177.

27. Vignols F, Rigau J, Miguel A, Capellades M, Puigdomènech P (1995) The brown midrib3 (bm3) Mutation in Maize Occurs in the Gene Encoding Caffeic Acid O-Methyltransferase. The Plant Cell 7: 407–416.

28. Sattler SE, Funnell-Harris DL, Pedersen JF (2010) Brown midrib mutations and their importance to the utilization of maize, sorghum, and pearl millet lignocellulosic tissues. Plant Science 178: 229–238.

29. Appenzeller L, Doblin M, Barreiro R, Wang H, Niu X, et al. (2004) Cellulose synthesis in maize: isolation and expression analysis of the cellulose synthase (CesA) gene family. Cellulose 11: 287–299.

30. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, et al. (2009) Genetic properties of the maize nested association mapping population. Science (New York, NY) 325: 737–740.

31. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, et al. (2009) The genetic architecture of maize flowering time. Science (New York, NY) 325: 714–718.

32. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6: e19379.

33. Bradbury PJ (in prep.)

34. SAS Institute (2004) SAS/STAT user's guide. Version 9.2. SAS Inst., Cary, NC

35. R D (2011) R: A language and environment for statistical computing.

36. Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal 4: 250.

37. Holthaus JF, Lamkey KR (1995) Population Means and Genetic Variances in Selected and Unselected Iowa Stiff Stalk Synthetic Maize Populations. Crop Science 35.

38. Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Science 50: 1681.

39. Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. TAG Theoretical and applied genetics 120: 151–161.

40. Massman JM, Gordillo A, Lorenzana RE, Bernardo R (2012) Genomewide predictions from maize single-cross data. TAG Theoretical and applied genetics DOI 10.1007/s00122-012-1955-y

SUPPLEMENTAL MATERIAL

Supplement table 4.1. Positions of QTL identified by joint linkage mapping for root, stalk and total lodging in middle and late environments.

	middle environments					late environments				
trait	Root	Root	Root	Root	Root	Root	Root	Root	Root	Root
Chr	1	2	2	6	7	1	2	4	6	10
cM start	175.7	54.9	72.1	101	69.8	94.9	97.5	93.2	43.4	40.1
cM end	176.9	56.8	72.1	106.6	69.8	96.5	98.9	93.2	43.7	40.6
RefGen2 start	281,709,021	22,275,514	51,816,033	164,891,527	134,099,896	175,642,801	189,447,059	181,223,123	118,087,791	84,096,310
RefGen2 end	282,796,367	22,999,224	51,820,222	166,592,347	134,137,353	179,989,018	190,159,942	181,229,893	118,604,616	86,424,768
F value	4.79	20.74	2.8	3.71	2.3	2.37	2.74	2.29	2.42	2.44
P-value	<0.0001	<0.0001	<0.0001	<0.0001	0.0006	0.0002	<0.0001	0.0004	0.0002	0.0001
Effect Pop1	-1.52	-0.22	2.13	-1.12	-1.05	-1.57	-1.87	1.63	2.33	0.78
Effect Pop2	5.10	13.36	2.81	8.39	-4.87	1.52	0.15	4.05	-0.29	-0.89
Effect Pop3	-0.63	0.85	-0.03	0.06	-2.42	-2.12	-1.80	-0.23	-0.63	-2.58
Effect Pop4	2.39	7.48	-1.26	2.31	-0.32	1.57	4.06	-2.43	5.74	-1.18
Effect Pop5	-0.37	0.18	-0.96	0.37	-0.38	-5.92	1.05	2.23	2.52	-4.02
Effect Pop6	-0.59	-0.90	0.60	4.45	-2.78	0.94	0.55	-1.99	1.00	-1.18
Effect Pop7	2.34	3.49	1.05	0.95	2.66	1.64	-0.86	1.13	1.19	0.98
Effect Pop8	10.40	5.21	2.31	2.52	-4.18	-0.85	-0.39	0.58	-0.39	4.79
Effect Pop9	6.50	0.22	2.71	7.31	-0.17	-0.01	3.09	-3.47	0.85	-1.88
Effect Pop11	-	-	-	-	-	1.08	9.31	-2.15	-3.88	2.74
Effect Pop12	2.98	5.37	2.61	2.44	-6.00	0.15	-1.47	0.08	-0.37	-2.94
Effect Pop13	-0.10	-1.08	2.50	1.59	-1.36	-2.33	3.09	-1.04	4.87	-4.64
Effect Pop14	-1.38	-0.27	3.70	0.08	-0.71	1.97	-0.38	-2.03	3.11	-4.67
Effect Pop15	2.12	-0.35	6.58	5.69	-4.82	2.19	2.72	3.13	-0.39	3.31
Effect Pop16	2.01	0.58	1.66	1.79	2.07	0.02	-1.97	-1.38	-0.38	0.11
Effect Pop18	-0.35	-7.06	10.45	-0.32	1.11	1.13	-0.18	-2.78	0.22	1.99
Effect Pop19	-3.69	-1.32	4.00	-2.95	-1.96	1.04	-0.65	-2.04	-0.65	-1.67
Effect Pop20	0.17	0.25	0.62	-1.42	-0.92	0.21	1.36	-2.59	3.41	-1.99
Effect Pop21	-2.29	3.44	-0.09	0.97	-1.13	2.06	0.47	-2.63	-0.39	3.92
Effect Pop22	-0.98	-0.40	-0.43	2.55	-0.11	1.46	0.73	-0.53	-0.82	-1.85
Effect Pop23	-1.74	0.20	9.86	-1.81	2.50	0.98	-2.96	6.73	0.97	2.03
Effect Pop24	-	-	-	-	-	9.30	-11.77	0.14	8.25	3.04
Effect Pop25	1.14	-0.93	5.23	0.40	-0.95	2.21	2.10	-4.07	-0.52	2.88
Effect Pop26	1.75	3.71	5.18	0.49	-1.62	4.23	1.12	1.37	-0.07	-0.29

	middle environments			late environments
trait	Stalk	Stalk	Stalk	Stalk
Chr	5	6	9	9
cM start	10.1	41.5	47	44.5
cM end	10.9	42.1	47.2	45.2
RefGen2 start	3,396,978	114,741,030	90,969,822	28,399,313

RefGen2 end	3,538,875	117,104,582	93,381,704	37,409,502
F value	2.69	4.76	3.5	2.57
P-value	<0.0001	<0.0001	<0.0001	<0.0001
Effect Pop1	0.41	-0.28	-0.58	-4.25
Effect Pop2	0.50	-1.26	0.84	-3.47
Effect Pop3	0.36	-0.24	0.10	4.27
Effect Pop4	0.05	0.61	-0.04	0.69
Effect Pop5	-1.57	-1.28	-1.95	0.98
Effect Pop6	0.23	-0.24	-2.04	-1.61
Effect Pop7	-2.12	-0.29	-5.46	2.23
Effect Pop8	-3.90	-7.74	-2.17	1.59
Effect Pop9	0.24	-1.23	-0.67	6.57
Effect Pop11	-	-	-	-2.10
Effect Pop12	0.98	-1.34	0.29	-2.92
Effect Pop13	-0.77	-0.60	-2.80	-2.63
Effect Pop14	-0.04	0.36	0.06	-2.25
Effect Pop15	-0.86	-2.76	-1.88	-1.78
Effect Pop16	-2.88	-3.72	-0.58	4.51
Effect Pop18	0.96	0.35	0.27	0.65
Effect Pop19	-0.35	0.23	0.39	-4.70
Effect Pop20	0.20	-0.01	0.05	0.28
Effect Pop21	0.75	-1.03	0.15	2.58
Effect Pop22	-0.77	-0.46	-1.78	-3.68
Effect Pop23	-0.10	0.02	-0.39	-1.36
Effect Pop24	-	-	-	12.76
Effect Pop25	-0.03	0.11	0.04	5.49
Effect Pop26	1.57	-1.32	0.12	-1.50

	middle environments					late environments		
trait	Total	Total	Total	Total	Total	Total	Total	Total
Chr	1	2	2	6	7	10	10	10
cM start	175.7	41.5	98.9	101	69.8	44.7	89.8	43.4
cM end	176.9	41.8	103.7	106.6	71.2	44.8	93.5	43.8
RefGen2 start	281,709,021	15,580,132	190,159,942	164,891,527	134,137,353	111,430,628	147,513,462	99,604,014
RefGen2 end	282,796,367	15,890,017	195523744	166,592,347	136,868,163	111,779,138	149,164,320	106,794,844
F value	5.64	2.56	2.99	3.73	2.76	2.46	2.44	2.24
P-value	<0.0001	0.0001	<0.0001	<0.0001	<0.0001	0.0001	0.0001	0.0006
Effect Pop1	-0.78	-0.30	1.78	-1.32	-1.73	8.04	4.42	-2.96
Effect Pop2	6.79	9.34	9.15	9.19	-6.39	0.15	4.79	0.85
Effect Pop3	-0.12	0.23	-0.66	0.59	-2.78	-10.23	5.97	4.18
Effect Pop4	1.63	5.58	2.92	1.19	-1.42	5.71	4.33	-9.39
Effect Pop5	0.92	-2.46	2.02	-1.45	0.45	3.05	4.52	-10.70
Effect Pop6	-0.42	-1.32	4.58	3.04	-3.04	2.86	0.04	-4.28
Effect Pop7	3.21	4.93	-1.41	4.14	5.67	4.19	3.38	-1.30
Effect Pop8	13.64	2.45	5.08	3.64	-2.17	4.61	0.26	-0.57

Effect Pop9	7.85	0.48	0.51	8.50	0.75	-13.12	1.88	13.28
Effect Pop11	-	-	-	-	-	-0.08	1.21	-0.40
Effect Pop12	0.84	8.21	5.92	2.06	-8.74	-20.60	2.40	13.40
Effect Pop13	-0.63	2.10	0.57	1.96	1.88	3.43	6.40	-7.23
Effect Pop14	-1.60	-0.74	3.96	0.19	-0.87	-2.84	-2.13	2.91
Effect Pop15	5.09	3.95	-0.78	6.96	-3.91	5.20	0.86	-1.92
Effect Pop16	1.76	7.26	2.64	1.14	3.47	-7.86	5.30	10.95
Effect Pop18	-0.27	-2.58	7.12	-0.31	0.72	-0.69	4.70	8.60
Effect Pop19	-3.72	1.32	0.19	-1.85	-2.01	-10.87	-3.64	9.37
Effect Pop20	1.32	-0.31	2.63	-2.16	-0.23	2.55	1.12	-4.74
Effect Pop21	-2.33	3.40	-3.72	3.03	1.43	8.67	5.67	-8.32
Effect Pop22	-0.67	3.00	0.27	0.92	-1.54	-29.33	2.60	29.37
Effect Pop23	-0.38	3.99	2.44	-3.33	4.72	11.53	-1.57	-14.88
Effect Pop24	-	-	-	-	-	12.95	6.48	-11.86
Effect Pop25	0.27	4.64	2.42	-1.27	-2.06	5.48	6.52	0.54
Effect Pop26	0.34	0.83	9.63	2.05	0.85	3.12	1.89	-4.87

Supplemental table 4.2. List of candidate genes for lodging. Genes located in QTL intervals are bold.

Lignin synthesis		chr	position (bp)
bm1	brown midrib1	5	98,993,016-98,997,371
bm3	brown midrib3	4	32,249,665-32,251,536
bm4	brown midrib4	9	154,600,921-154,752,762
cncr1	cinnamoyl-CoA reductase	1	212,986,619-213,227,291
cncr2	cinnamoyl-CoA reductase	7	19,497,483-47,934,542
pox3	peroxidase3	6	131,974,895-132,090,730
px1	**peroxidase1**	**2**	**198,482,067-199,166,751**
px3	peroxidase3	7	170,246,381-170,998,616
px13	peroxidase 13	5	28,335,249-31,688,205
px14	**peroxidase14**	**2**	**22,753,277-28,400,591**
cad1	cinnamyl alcohol dehydrogenase1	2	10,531,186-10,535,293
CCoA-OMT2	**Caffeoyl-CoA 3-O-methyltransferase 2**	**2**	**16,318,197-16,320,573**
CCoA-OMT1	Caffeoyl-CoA 3-O-methyltransferase 1	6	198,081,373-198,082,561
4CL2	**4-coumarate-CoA ligase 2**	**2**	**53,200,444-53,202,745**
Phenylpropanoid pathway			
pal1	phenylalanine ammonia lyase1	5	186,677,004-186,680,745
pal2	**phenylalanine ammonia lyase2**	**2**	**20,735,071-21,327,621**
pal3	phenylalanine ammonia lyase3	4	143,061,373-143,187,277
p1	pericarp color1	1	48,117,497-48,128,047
p2	pericarp color2	1	48,092,238-48,097,446
chi2	chalcone isomerase 2	2	206,127,228-206,130,402
whp1	white pollen1	2	223,888,706-223,892,691
c2	colorless2	4	192,580,450-192,583,847
Vegetative phase change			
tp1	**teopod1**	**7**	**131,009,736-131,984,320**
tp2	teopod2	10	127,514,254-128,401,699
gl1	glossy1	7	118,517,870-118,523,644
gl2	glossy2	2	10,624,501-10,627,540
gl3	**glossy3**	**4**	**185,653,178-185,654,761**
gl4	glossy4	4	160,279,269-163,130,361
gl5	glossy5	4	18,538,004-18,626,350
gl6	glossy	3	148,437,112-149,058,463
gl7	glossy7	4	31,272,143-36,545,436
gl8	glossy8	5	181,197,346-181,200,367
gl11	glossy11	2	40,487,829-41,096,833
gl14	glossy14	2	119,771,646-135,856,854
gl15	**glossy15**	**9**	**95,739,338-95,742,681**
gl17	glossy17	5	59,698,294-64,729,128
gl18	glossy18	8	110,709,532-112,836,130
epc1	early phase change1	8	34,593,113-60,338,399
Cellulose synthesis			
cesa1	cellulose synthase1	8	80,220,224-80,226,330
cesa2	cellulose synthase2	6	128,560,824-128,566,579
cesa3	cellulose synthase3	3	11,642,850-11,649,401
cesa4	cellulose synthase4	7	14,617,522-17,314,017

cesa5	cellulose synthase5	1	290,593,939-290,599,907
cesa6	cellulose synthase6	1	296,254,268-296,259,194
cesa7	cellulose synthase7	7	37,024,553-37,030,325
cesa8	cellulose synthase8	7	26,488,357-26,493,319
cesa9	cellulose synthase9	2	161,123,950-161,130108
cesa10	cellulose synthase10	1	233,557,986-235,493,684
cesa11	cellulose synthase11	3	198,366,610-198,371,864
cesa12	cellulose synthase12	7	117,513,371-117,517,542
GRMZM2G002523_P01		2	182,334,982-182,339,233
GRMZM2G011651_P01		5	49,637,565-49,641,824
GRMZM2G012044_P01		5	204,318,343-204,322,398
GRMZM2G014558_P01		**7**	**130,702,896-130,707,663**
GRMZM2G015886_P01		10	22,262,131-22,265,986
GRMZM2G027794_P01		8	172,971,672-172,975,677
GRMZM2G028353_P01		2	169,757,962-169,762,838
GRMZM2G044269_P01		1	221,785,864-221,789,385
GRMZM2G061764_P01		9	51,950,792-51,954,029
GRMZM2G074546_P02		**2**	**51,750,198-51,754,667**
GRMZM2G074792_P01		6	158,223,202-158,225,989
GRMZM2G082580_P01		2	170,771,739-170,775,663
GRMZM2G103972_P01		7	156,222,318-156,226,575
GRMZM2G104092_P01		9	64,924,031-64,927,122
GRMZM2G105631_P01		4	234,113,329-234,119,648
GRMZM2G110145_P01		10	77,269,832-77,275,336
GRMZM2G113432_P01		7	156,212,521-156,216,820
GRMZM2G122277_P01		4	31,242,535-31,248,167
GRMZM2G122431_P01		3	145,175,899-145,178,917
GRMZM2G150404_P01		2	161,135,365-161,137,790
GRMZM2G164761_P01		1	239,207,473-239,210,788
GRMZM2G173759_P01		7	118,581,737-118,584,207
GRMZM2G339645_P01		7	156,285,531-156,288,622
GRMZM2G349834_P01		6	102,678,170-102,681,460
GRMZM2G367267_P01		2	204,658,088-204,661,597
GRMZM2G405567_P02		5	206,843,832-206,846,794
GRMZM2G436299_P01		1	104,608,915-104,612,653
GRMZM2G445905_P03		1	234,254,443-234,264,575
GRMZM2G454081_P04		3	194,038,733-194,043,307
GRMZM5G870176_P01		**9**	**26,601,759-26,606,021**