ABSTRACT

BIAN, YANG. Genetic Mapping and Genomic Prediction of Maize Quantitative Traits. (Under the direction of James B. Holland).

Genomic data holds promise to help crop breeders to continue increasing crop production to keep pace with fast-growing global population while minimizing environmental impact from farming. Effective use of genetic variation for plant breeding requires an understanding of the genetic architecture of agronomic traits. In Chapter 2, I revisited quantitative trait locus (QTL) and genome-wide association (GWA) analyses of southern leaf blight (SLB) resistance in the maize Nested Association Mapping (NAM) populations to determine how sensitive the results of these studies are to changes in input data. I used an updated NAM linkage map that has a six-fold greater marker density and updated founder haplotype data set containing 17 times more (28.5 million) genomic variants than a previous study. Sensitivity analysis indicated that QTL were relatively stable to perturbations of data inputs. GWAS had lower power due to stringent thresholds designed to minimize false positive associations, resulting in variability of detection across studies. The updated higher density linkage map improved QTL estimation and, along with much denser SNP HapMaps, greatly increased the likelihood of detecting SNPs in linkage with causal variants. I recommend use of the updated genetic resources and results but emphasize the limited repeatability of small-effect associations.

QTL models can provide useful insights into trait genetic architecture because of straightforward interpretability, but their accuracy is limited due to difficulty in including numerous small-effect loci without overfitting. In Chapter 3, I developed a thinning and aggregating (TAGGING) method as a new ensemble learning approach to QTL mapping. The original linkage map marker sets were stratified into sub-maps, QTL mapping models
were trained upon the thinned marker sets as based learners in parallel, and aggregated by averaging the predictions from the base learners to predict the test data. TAGGING reduces collinearity problems by thinning dense linkage maps, maintains aspects of marker selection that characterize standard QTL mapping, and by ensembling, incorporates information from many more marker-trait associations than traditional QTL mapping. TAGGING was compared to standard QTL mapping using cross validation in the maize NAM populations. TAGGING-assisted QTL mapping substantially improved prediction ability for both biparental and multiparental populations, by reducing both the variance and bias in prediction. Furthermore, an ensemble model combining predictions from TAGGING-assisted QTL and infinitesimal genetic models improved prediction abilities over the component models, indicating some complementarity between model assumptions and suggesting that some trait genetic architectures involve a mixture of a few major QTL and polygenic effects.

GWA typically identifies a number of marker-trait associations, but top associations alone have low prediction ability for most traits. Genomic prediction (GP) can use the same genotypic and phenotypic data as GWA, but instead of identifying individual polymorphisms that control traits, GP models emphasize best possible predicted values. Ideally, GP will become indistinguishable from GWA if the identified SNPs are of importance. In Chapter 4, I developed two linear mixed models for GWA in maize NAM populations, based on SNPs having equal or variable effects across families, aiming for mapping reliable associations outstanding the polygenic background. By explicitly incorporating the GWA discoveries as fixed effects into the infinitesimal model, prediction of two disease resistance traits was significantly improved. Simulation studies revealed that the effectiveness of this joint approach depends on the extent of polygenicity of the traits. The refined GWA models can
provide guidelines for distinguishing the purely polygenicity from the mixture genetic architecture. The difference in genetic architecture of traits has important implications for breeding strategies that can benefit from improving the GP models. Some of the best opportunities to the future of crop improvement may lie in using genomic modification technologies to directly target discovery of variation derived from allele mining and comparative genomics.
Genetic Mapping and Genomic Prediction of Maize Quantitative Traits

by
Yang Bian

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

Crop Science

Raleigh, North Carolina

2016

APPROVED BY:

_______________________________  _________________________________
James B. Holland               Trudy F. C. Mackay
Committee Chair

_______________________________  _________________________________
Christian Maltecca              Jung-Ying Tzeng
DEDICATION

To my parents, family and friends, for your love and support!
Yang Bian was born in China and came to the U.S. for attending graduate school. Yang obtained his bachelor’s degree in Horticulture at Huazhong Agricultural University, and earned his master’s degrees in Horticultural Science and Analytics at North Carolina State University. Starting from summer 2013, he studied under Dr. Holland for his doctoral dissertation work. En route to this PhD degree, he would complete his master’s degree in Statistics.
ACKNOWLEDGMENTS

I’d like to especially thank my PhD advisor, Dr. Jim Holland, for all the knowledge, patience, and advice you gave me during my time being in the program. I owe many thanks to other advisory committee members, Drs. Trudy Mackay, Christian Maltecca, and Jung-Ying Tzeng. A lot of members of Holland Lab (Team Corn) have contributed their insights into this dissertation including: Drs. Bode Olukolu, Funda Ogut, Charlie Zila, Tiffany Jamann, Zhou Fang, Shilpa Sood, Luis Fernando Samayoa and Ryan Andres; technicians, Jason Brewer and Josie Bloom; and graduate students, Shang Xue, Thiago Marino and David Horne. I want to say Thank You to Drs. Charlie Stuber, Qin Yang, Peter Balint-Kurti, Randy Wisser, Peter Bradbury, Jason Peiffer, Ed Buckler, Dan Gianola, Brian Reich, Howard Bondell and Alison Motsinger-Reif for strengthening my research with your valuable grants, suggestions and collaborations.
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................... vii
LIST OF FIGURES .......................................................................................................... viii
CHAPTER 1: Literature Review ...................................................................................... 1
   Genetic mapping .......................................................................................................... 1
   Nested association mapping populations ................................................................. 3
   Genomic prediction research ..................................................................................... 6
   Missing pieces and perspectives ................................................................................. 8
   References .................................................................................................................. 10
CHAPTER 2: Limits on the Reproducibility of Marker Associations with Southern Leaf
   blight resistance in the maize nested association mapping population .................. 13
   Abstract ...................................................................................................................... 15
   Introduction ............................................................................................................... 17
   Results ....................................................................................................................... 20
   Discussion .................................................................................................................. 26
   Methods ...................................................................................................................... 31
   References .................................................................................................................. 50
CHAPTER 3: Ensemble Learning of QTL Models Improves Prediction of Complex Traits 56
   Abstract ...................................................................................................................... 58
   Introduction ............................................................................................................... 59
   Materials and Methods ............................................................................................. 64
   Results ....................................................................................................................... 74
   Discussion .................................................................................................................. 81
   References .................................................................................................................. 97
CHAPTER 4: Enhancing Genomic Prediction with Genome-wide Association Studies in
   Multiparental Maize Populations ................................................................................. 102
   Abstract ...................................................................................................................... 104
   Introduction ............................................................................................................... 105
   Materials and Methods ............................................................................................. 109
Results ............................................................................................................................................... 117
Discussion ....................................................................................................................................... 121
References ......................................................................................................................................... 129
APPENDICES .................................................................................................................................. 133
Appendix A: Supplemental Information for Chapter 2................................................................. 134
Appendix B: Supplemental Information for Chapter 3................................................................. 138
Appendix C: Supplemental Information for Chapter 4................................................................. 150
LIST OF TABLES

Table 2-1 Inputs for joint linkage mapping QTL analysis and GWAS models ............... 44

Table 2-2 Physical and genetic positions for QTL peaks and support intervals (SI) mapped using current updated phenotypes and linkage map (model 7) and comparison to QTL previously reported ................................................................. 45

Table 2-3 Input changes in joint linkage mapping QTL modeling affect the values of the chromosome-specific residuals more profoundly than the corresponding predicted phenotypic values ................................................................. 46

Table 2-4 Highly significant GWAS variants (RMIP ≥ 0.25) and their adjacent candidate genes ...................................................................................................................... 47

Table 2-5 Concordance of variants associated with SLB resistance at resample model inclusion probabilities (RMIPs) of 0.05 and 0.25 from five GWAS analyses ............ 49

Table 3-1 Description of prediction models compared ..................................................... 96
LIST OF FIGURES

Figure 2-1 Heat map of additive effect estimates of 25 founder parent alleles for QTLs of SLB resistance relative to B73 ........................................................................................................41

Figure 2-2 QTL positions on the ten maize chromosomes from seven joint linkage mapping models ........................................................................................................................................42

Figure 2-3 Manhattan plots from genome-wide association analysis for SLB across the ten chromosome pairs of maize .........................................................................................................................................43

Figure 3-1 Data sampling, map thinning, model building and ensemble schemes ..........91

Figure 3-2 Prediction $R^2$ for resistance to SLB in biparental and multiple-family prediction, comparing JF, SF and the TAGGING-assisted QTL analyses using multiple map densities and under multiple selection $p$ .........................................................................................................................................92

Figure 3-3 Prediction bias$^2$ and variance in TAGGING-assisted and single QTL models for resistance to SLB ........................................................................................................................................93

Figure 3-4 Marker-trait associations identified by JF and EJF models across ten chromosomes, using different map densities under selection $p$ at 0.001 for maize SLB resistance in NAM panel .........................................................................................................................................94

Figure 3-5 Ensemble between TAGGING-assisted QTL and GBLUP models showed their complementary effects on prediction ability ......................................................................................................................................95

Figure 4-1 Prediction $R^2$ in the simulated oligogenic and polygenic traits .................126

Figure 4-2 GWAS repeatability plot for the three agronomic traits .........................127

Figure 4-3 Within-family Prediction $R^2$ of the three agronomic traits using GBLUP and the new models ........................................................................................................................................128

Figure A-1 Histograms and box plots of prediction correlation coefficients from 100 random cross-validation analyses sets randomly subsampled from the common set of 4354 RILs ........................................................................................................................................135

Figure B-1 Within-family prediction $R^2$ for plant height in biparental and multiple-family prediction, comparing JF, SF and the TAGGING models using multiple map densities at multiple selection $p$ ........................................................................................................................................139
Figure B-2 Within-family prediction $R^2$ for days to anthesis in biparental and multiple-family prediction, comparing JF, SF and the TAGGING models using multiple map densities at multiple selection $p$ ..........................140

Figure B-3 Prediction bias$^2$ and variance in TAGGING and discrete QTL models for plant height (top) and days to anthesis (bottom) ..........................141

Figure B-4 Marker-trait associations identified by JF and EJF models across ten chromosomes, using different map densities under selection $p$ at 0.001 for plant height in NAM panel.................................142

Figure B-5 Marker-trait associations identified by JF and EJF models across ten chromosomes, using different map densities under selection $p$ at 0.001 for days to anthesis in NAM panel.................................143

Figure B-6 Resample Model Inclusion Probabilities (RMIP) distributions for genetic resistance to SLB using EJF models with 100 sets of 20 cM reduced maps at selection $p = 1e^{-5}, 1e^{-4}, 1e^{-3}$ and $1e^{-2}$.................................144

Figure B-7 Ensemble of QTL and GBLUP models showed complementary effects on prediction ability (Left, plant height, right, days to anthesis).............................145

Figure B-8 Resample Model Inclusion Probabilities (RMIP) of JF models that used different linkage map densities under selection $p = 1e^{-3}$ for genetic resistance to SLB..............146

Figure C-1 Power and false discovery rate in oligogenic trait simulation....................151

Figure C-2 Power and false discovery rate in polygenic trait simulation ....................152

Figure C-3 GWAS Manhattan plots for SLB using main effect models ....................153

Figure C-4 GWAS Manhattan plots for SLB using nested effect models ....................153

Figure C-5 GWAS Manhattan plots for GLS using main effect models ....................154

Figure C-6 GWAS Manhattan plots for GLS using nested effect models ....................154

Figure C-7 GWAS Manhattan plots for PHT using main effect models ....................155

Figure C-8 GWAS Manhattan plots for PHT using nested effect models ....................155

Figure C-9 Boxplots of average prediction $R^2$ for 25 NAM populations, comparing GBLUP model and main effect GWAS + REML-based linear mixed model ....................156
CHAPTER 1: Literature Review

Genetic mapping

Genetic mapping of quantitative traits aims to identify quantitative trait loci (QTL) underlying phenotypic variation. Two major approaches, linkage analysis (QTL mapping) and association mapping (linkage disequilibrium mapping), both aim to identify associations between phenotypes and genotypes by identifying sequence polymorphisms that are correlated with phenotypic variation through correlation with causal sequence variation. Adopted from human genetics studies in mid 2000s, genome-wide association study (GWAS) is widely applied to identify candidate regions associated with agronomic traits in maize and emerging as the method of choice in other crop species. The strategy of GWAS is to genotype enough DNA markers across the genome so that functional loci will likely be in linkage disequilibrium (LD) with the tested markers.

In maize, a haplotype map (HapMap) Project has generated millions of single nucleotide polymorphisms (SNPs) and other DNA structural variants (http://www.panzea.org/) (GORE et al. 2009; CHIA et al. 2012), increasingly satisfying the necessary amount to cope with the large genome size and fast-decay LD structure for diverse collections. While genotyping-by-sequencing (GBS) has been the economic high throughput genotyping technology thus far, the decreasing costs of next-generation sequencing (NGS) will allow future projects to cost-effectively generate whole sequence for a large sample size, unlocking all DNA polymorphisms in the mapping population under study.
After identifying quantitative trait nucleotides (QTNs), researchers hope to understand how the DNA sequence differences result in phenotypic differences. The dissection of the phenotypic variation due to genetics will not only provide guidance to the subsequent functional and pathway analysis, but also lead to more efficient breeding practice by introducing favorable alleles from new germplasm sources while eliminating deleterious alleles into or out of the elite variety, which is a particular fortunate task to agriculturalists (Tanksley and McCouch 1997).

Whole genome sequencing promises to provide information on all sequence variants in the sample under study, but complete sequence information alone does not result in identification of the variants causing phenotypic variation. Beyond coverage of all true sequence polymorphisms, the power and resolution of GWAS is limited by the extent of LD, allele frequencies, sample size and genetic architectures. Genetic architecture is the atlas of gene effects and networks that are responsible for heritable genetic variation in a quantitative trait, and can be seen as a map of genes locations, effects and relationships. Genetic architecture underlying quantitative trait is a critical factor affecting successful detection of genes, and the one factor that cannot be controlled by the investigator. The power to detect an association between a functional variant and a phenotype is a function of the allele frequency and the effect that the functional variants have on the phenotypic variation. Numerous QTL of small effects will be naturally difficult to detect at the gene level than a small number of QTL of large effects. The vast majority of maize QTL have small effects (Buckler et al. 2009; Kump et al. 2011; Tian et al. 2011; Olokolu et al. 2014; Wallace et al. 2014), but to what extent the genetic architectures differ among traits remains largely unknown. GWAS
models need to be improved for sufficient power to detect small-effect QTL, which can reveal a better picture of the common phenotypic variation.

Most GWAS analyses proceed first by filtering out the genetic markers that segregate at low frequency in the mapping population. However, most alleles have low frequency in diverse maize germplasm samples (ROMAY et al. 2013), and 30% of the polymorphisms in a diverse panel of 27 maize inbred lines are unique to a single line (Myles et al. 2009). If rare alleles collectively have substantial influence on phenotypic variation, we may be missing many functional variants in GWAS due to a combination of data filtering and low power of detection of rare allele effects. By making controlled crosses, however, the frequencies of alleles captured in the parents are balanced in the progeny, increasing power of detection of their effects.

**Nested association mapping populations**

Genetic analysis of quantitative traits using multiparental populations takes advantage of large sample size by constructing the populations from interconnected related small biparental families (Churchill et al. 2004; Kover et al. 2009; McMullen et al. 2009; Mackay et al. 2012; Bauer et al. 2013). By tracking identity-by-decent (IBD) of linkage markers, researchers can project extensive sequencing information from founder lines onto their progenies, based on lower density marker genotypes on the progenies. As a strategy to control for spurious marker-phenotype associations due to population structure, linear mixed models (LMM) were introduced into GWAS by taking population memberships, principal components as fixed-effect covariate and/or kinship relationship as random-effect correlation
structure (Yu et al. 2006; Zhang et al. 2010). GWAS sample size is largely increased by combining multiparental populations, and allele frequencies are pushed away from being rare. However, all previous GWAS applied biallelic genetic assumption, implying that the same allele always has the same effect independent of the source of multiparental populations. In theory, it is certainly possible for the tested SNPs to have differential effects across populations. The tested markers may have differential LD with the functional variants varying by populations, different populations may have different (or rare) functional variants, and QTN may have epistasis. In this dissertation, I therefore explored and evaluated a new GWAS model based on such population-dependent genetic assumption.

Maize nested association mapping (NAM) populations provided a community resource for studying quantitative traits (Buckler et al. 2009). The 25 diverse founder inbred lines selected to create biparental families in crosses with the common founder B73 were chosen to maximize the genetic diversity representing the global public maize germplasm (Liu et al. 2003; Buckler et al. 2009). The maize NAM populations were genotyped with a common set of thousands of single nucleotide polymorphism (SNP) markers (McMullen et al. 2009) and a consensus map was constructed to allow a joint family QTL analysis. Moreover, all NAM founder lines were whole-genome sequenced to produce maize HapMap, along with other important maize inbred lines. The first version of HapMap (HapMap V1) contained 1.6 million SNPs (Gore et al. 2009), HapMap V2 released 26.5 M SNPs and read-depth variants (RDVs) (Chia et al. 2012); currently HapMap V3 is under release, and the HapMap V4 is under construction. All NAM recombinant inbred lines (RILs) were grown and phenotyped together across a wide range of geographic and climate
regions for multiple years, and so far at least 41 agronomic traits have been thoroughly studied using maize NAM populations (for the list of traits, see (Wallace et al. 2014)). The NAM design itself has been applied to numerous other plant species, too.

As a genetic mapping system, maize NAM design is particularly advantageous at high level of genetic diversity, high-density linkage map, high mapping resolution due to fast LD delay, high heritability by repeated environment measures, known population structure, and non-rare allele frequencies (Yu et al. 2008; Buckler et al. 2009; McMullen et al. 2009). All those merits permit NAM a good detection power and high resolution in dissection of natural phenotypic variation in complex traits. Key findings from previous research on the maize NAM population include that the vast majority of maize QTL have small effects, little evidence of epistasis is found for most maize agronomic traits, and a significant proportion of the phenotypic variance is explained by the simple additive models of the detected SNPs. In addition, substantial phenotypic variation is controlled by sets of common QTL/genes with series of allele effects, indicating of clustered rare alleles contributing to the genetic variation. As part of this dissertation, I present a re-analysis of the maize NAM populations to gain further insights into the genetic architectures of southern leaf blight, flowering time, gray leaf spot and plant height by developing a few novel statistical genetic mapping models.

Questions about the optimal experimental design and statistical analysis remain. The best allocation of germplasm resource in multiparental populations is still a vigorous research area for experimental design; for NAM design, there has been speculation for doubling the common founder lines or increasing the alternate founder lines at the expense of reducing
each family sample size. More importantly, while the current NAM studies and other GWAS studies using the modern maize lines have restrictedly focused on common variants, most of the rare allele diversity has been left out in describing the basic genetic architecture of traits. For example, the current breeding germplasm in both public and commercial programs have a relatively narrow origin and are heavily recycled. Much of publically accessible protected germplasm was reported to originate from seven progenitor lines: B73, LH82, LH123, PH207, PH595, PHG39, and Mo17 (Mikel and Dudley 2006). While modern maize inbred lines developed from pedigree breeding have gone through a strong bottleneck on diversity, which drove many historical rare alleles to become (nearly) lost, the outcrossing maize landraces experienced less severe bottlenecks from selection during domestication and adaption. The maize landraces should have much more prevalent rare alleles across genome; and through proper design of mating and contrasting among modern maize and landraces, they make it a promising genetic modeling system to unravel the biology of rare alleles.

**Genomic prediction research**

Empirical and simulation evidence suggests that genomic prediction (GP) can produce greater genetic gain per unit time than phenotypic selection in plant breeding with use of year-round nurseries and high-throughput genotyping technology (Heffner et al. 2011; Massman et al. 2013). GP can utilize all available genomic markers simultaneously to prediction quantitative traits; many GP models assume that the performance of trait is a product of all loci in the genome. The implementation of GP in both public and commercial...
breeding programs can ultimately deliver higher yielding, more adapted, and most pest and disease resistant lines at a faster speed.

While both GWAS and GP can use the same data and similar statistical models, predictionists aim to accurately estimate breeding values without emphasizing genetic mapping. GP has shown great promise for animal breeding programs and is increasingly being adopted in plant breeding (HESLOT et al. 2012). Despite varying success in GWAS gene discovery due to LD structure, rare alleles and other constrains, GP for quantitative traits seems to work well as long as the marker coverage is extensive and training and test samples are closely related. For example, in genomic BLUP (GBLUP) model, if marker-based genetics relationships are a good approximation of the additive genetic relationships at QTL, predictions should be accurate. When the marker alleles have tight LD with causal alleles, GBLUP model promises to result in a good prediction performance, especially when the phenotypic variation is controlled by numerous QTL with small effect. A mixture of major and small-effect QTL would otherwise favor variable selection techniques from the Bayesian alphabet list (MEUWISSEN et al. 2001; PARK AND CASELLA 2008; GIANOLA et al. 2009; HABIER et al. 2011). The Bayesian models tend to require long computation time, so sometimes it is not practical to cope with the fast decision making needed in breeding program. The parameter estimates from Bayesian or non-parametric methods (kernel methods, random forest etc.) are sensitive to priors and unstable across studies, although good predicted values are of much interest in GP. In this dissertation, I present some novel ideas of merging QTL mapping model into GP model by ensemble learning. At the present, the efficacy of GP is limited by its accuracy and a continuing effect needs to be devoted into
improving methods. In this dissertation, I demonstrate that GP can be indistinguishable from GWAS, and GP can be further improved by the GWAS discoveries.

**Missing pieces and perspectives**

GP has proven useful for within-family prediction. Cross-population or cross-family prediction is arguably the most difficult hurdle to jump. So far, GP models tuned on one set of samples are often of little value in predicting unrelated populations. The limiting factors to transferability may include rare alleles, poor estimation, and epistasis, among others. There have been some GP studies investigating optimally allocating training set germplasms of multiple related or unrelated small biparental families to predict progeny from other crosses. However, further efforts needs to be put into studying the best practice of modeling in cross-population prediction: the statistics and analytics need to be further reformed to tailor the difference between training and test sets. Second, it is important to extend the GP framework into multi-trait joint selection (Jia and Jannink 2012), because only the comprehensive selections can realistically meet the real-world breeding practice. Additional layers of genomic information such as RNA profiles, transcriptomic data, epigenetic data, metabolomic data and other biomarkers may be crucial to broaden the current genetic variation source and genetic predictability (Riedelsheimer et al. 2012).

Finally, large scale genotypic and phenotypic data sets along with appropriate data analysis methods have the potential to assist continued gains in crop productivity. In the past several years, there has been increasing interest in high throughput phenotyping platforms (HTPPs). Aerial- and ground- based HTPPs collect continuous, detailed, and massive
information throughout the plant life cycle. Taking advantage of genetic, field, and climate data, precision agriculture is just around the corner. Last but not least, large-scale and cost-effective genomic editing in the future may pave the way to rapidly translate the gene discovery, evolution analysis and comparative genomics into targeted modification of plant genetic makeup. Extremely large numbers of causal variants and limited recombination hinder the use of conventional breeding techniques to eliminate deleterious alleles and introgress favorable alleles, but genomic editing could shift the paradigm to outpace the limitation of recombination beyond the biology of the species, permitting unprecedented advances in the improvement of crops.
References


Gore, M. A., J.-M. Chia, R. J. Elshire, Q. Sun, E. S. Ersoz et al., 2009 A first-generation haplotype map of maize. Science 326: 1115-1117.


Kover, P. X., W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich et al., 2009 A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. PLoS Genet. 5.

Liu, K., M. Goodman, S. Muse, J. S. Smith, E. Buckler et al., 2003 Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics. 165.


CHAPTER 2: Limits on the Reproducibility of Marker Associations with Southern Leaf blight resistance in the maize nested association mapping population

Citation
Limits on the reproducibility of marker associations with southern leaf blight resistance in the maize nested association mapping population

Yang Bian†
Email: ybian2@ncsu.edu

Qin Yang‡
Email: qyang6@ncsu.edu

Peter J Balint-Kurti‡,³
Email: Peter.Balint-Kurti@ars.usda.gov

Randall J Wisser⁴
Email: rjw@udel.edu

James B Holland¹,³*
* Corresponding author
Email: Jim.Holland@ars.usda.gov

¹ Department of Crop Science, North Carolina State University, Raleigh, NC 27695, USA

² Department of Plant Pathology, North Carolina State University, Raleigh, NC 27695, USA

³ U.S. Department of Agriculture-Agricultural Research Service, Plant Science Research Unit, Raleigh, NC 27695, USA

⁴ R.J. Wisser, Department of Plant and Soil Sciences, University of Delaware, Newark, DE 19716, USA

† Equal contributors.

Keywords

Quantitative trait loci, Nested association mapping, Disease resistance, Genome wide association study, Zea mays
Abstract

Background

A previous study reported a comprehensive quantitative trait locus (QTL) and genome wide association study (GWAS) of southern leaf blight (SLB) resistance in the maize Nested Association Mapping (NAM) panel. Since that time, the genomic resources available for such analyses have improved substantially. An updated NAM genetic linkage map has a nearly six-fold greater marker density than the previous map and the combined SNPs and read depth variants (RDVs) from maize HapMaps 1 and 2 provided 28.5 M genomic variants for association analysis, 17 fold more than HapMap 1. In addition, phenotypic values of the NAM RILs were re-estimated to account for environment-specific flowering time covariates and a small proportion of lines were dropped due to genotypic data quality problems. Comparisons of original and updated QTL and GWAS results confound the effects of linkage map density, GWAS marker density, population sample size, and phenotype estimates. Therefore, we evaluated the effects of changing each of these parameters individually and in combination to determine their relative impact on marker-trait associations in original and updated analyses.

Results

Of the four parameters varied, map density caused the largest changes in QTL and GWAS results. The updated QTL model had better cross-validation prediction accuracy than the previous model. Whereas joint linkage QTL positions were relatively stable to input changes, the residual values derived from those QTL models (used as inputs to GWAS) were
more sensitive, resulting in substantial differences between GWAS results. The updated NAM GWAS identified several candidate genes consistent with previous QTL fine-mapping results.

Conclusions

The highly polygenic nature of resistance to SLB complicates the identification of causal genes. Joint linkage QTL are relatively stable to perturbations of data inputs, but their resolution is generally on the order of tens or more Mbp. GWAS associations have higher resolution, but lower power due to stringent thresholds designed to minimize false positive associations, resulting in variability of detection across studies. The updated higher density linkage map improves QTL estimation and, along with a much denser SNP HapMap, greatly increases the likelihood of detecting SNPs in linkage with causal variants. We recommend use of the updated genetic resources and results but emphasize the limited repeatability of small-effect associations.
Introduction

Methods for elucidating the genetic architecture underlying quantitative variation in plants have evolved substantially over the last 25 years, following the first report of genome-wide quantitative trait locus (QTL) mapping [1]. The maize nested association mapping (NAM) population is composed of 5,000 recombinant inbred lines (RILs) derived from crosses between inbred line B73 and 25 other inbred lines of maize [2]. These parents were selected to capture a maximum amount of molecular genetic diversity present across the major subpopulations of public maize breeding germplasm [3,4]. The maize NAM population has been used to study genetic architectures for a number of quantitative traits of maize [5-11], including Southern leaf blight (SLB) resistance [8].

Southern leaf blight is a foliar disease of maize caused by the fungus Cochliobolus heterostrophus. The disease was responsible for a major epidemic in the U.S. in the 1970’s [12] and continues to limit or threaten maize production worldwide. Natural variation in resistance to SLB is polygenic and may involve a diverse array of functional genes and pathways [8,13]. Using joint linkage mapping (JLM) and genome-wide association study (GWAS) the genetic architecture of resistance to SLB in the NAM population was associated with more than 30 loci with small additive effects [8]. With the recent release of maize HapMap2 [14] and a denser linkage map based on genotyping-by-sequencing (GBS [15,16]) with markers positioned every 0.2 cM, QTL identified by JLM can be more precisely localized on the genetic and physical sequence maps. The denser linkage map is also expected to permit more accurate projection of the more than 28 M SNPs among parental
lines in maize HapMaps 1 and 2 onto NAM RILs, which should provide mapping precision to the limits dictated by linkage and disequilibrium in this population.

Two-stage regression analysis has been widely used to test SNPs for associations with quantitative diseases in human [17-19], and this approach has been adopted for GWAS in plants. In the first stage, observed phenotypes are regressed on covariates such as demographic, clinical, and/or environmental factors. In the second stage, the residual values from the first stage model (‘residual outcomes’ or adjusted phenotypic values) are regressed on genetic markers in a simple- or multiple-linear regression. Despite its convenience in computation, the two-stage method can result in a downwardly biased estimate of genotypic effects and loss of power in detection as a result of dependency between covariates and the tested SNP genotypes [20,21]. Two-stage approaches are also used to combine the complementary advantages of JLM and GWAS in NAM [7,8,10]. In the first stage, JLM is performed using a consensus linkage map to identify QTL across the genome. In the second stage, GWAS is performed chromosome-by-chromosome, using separate input values for each chromosome that are obtained as residuals from the first stage QTL model, built by excluding QTL on the chromosome to be tested for GWAS. The purpose of this is to adjust phenotype values used for association analysis for the effects of QTL on other chromosomes. This approach is expected to be largely free of the problem of dependency between covariates (QTL) fit in the first stage and SNPs tested in the second stage, since only QTL on different chromosomes than the test SNPs are fit as covariates.

We are working toward identifying the causal variants underlying quantitative resistance to SLB, relying, in part, on the information provided by NAM. The objective of
this study was to re-analyze resistance to SLB in the maize NAM panel using the updated genetic and haplotype maps, to compare the results with those of the previous analysis, and to determine which results are more reliable. The previous JLM analysis was based on SLB phenotypes measured on 4694 RILs and a linkage map of 1106 SNPs [2], and the previous GWAS analysis was based on 1.6 M SNPs of HapMap 1 [22]. Since that analysis, the mixed model used to produce the phenotypic inputs to the analysis was updated to better adjust for the effect of flowering time on SLB resistance phenotypes. The updated 7386-marker map has a uniform density of one marker every 0.2 cM, but the number (4413) of RILs phenotyped and genotyped with this map is smaller than previously (4694 RILs). Therefore, a second objective of this study was to measure the influence of each of the changes in the data used for analysis (genetic maps, RIL sample sizes, and phenotype values) on the current two-stage JLM-association analysis in the NAM panel, using SLB as an example. Finally, cross-validation was used to compare the prediction power of the original JLM model of Kump et al. [8] and the updated JLM model in the NAM panel.
Results

*Modeling the effect of flowering time on SLB resistance*

The relationship between flowering time and disease resistance was complex. Among the 135 of 156 possible combinations of rating × environment × NAM populations for which there were sufficient data for analysis, there was no significant relationship between flowering time and SLB resistance for 56 combinations, linear relationships for 75 combinations, and quadratic relationships for 4 combinations (Table 1 in File A-1). Only four populations exhibited a consistent relationship across ratings and environments (no effect for populations 11 and 22; linear effects for populations 8 and 26). Thus, the majority of disease ratings in every population exhibited significant but variable relationships with flowering time. The flowering time covariate effects were generally small (r$^2$ ranged from 1% to 22%), however, and the updated combined mixed model incorporating variation in the flowering time covariate effect only slightly altered the BLUPs: the original and updated BLUPs were highly correlated between entries in each population (min = 0.980; max =0.997), and there were only subtle differences in the rankings of population mean effects (r$^2$ = 0.983).

*Precision of QTL localization and improved QTL prediction power*

JLM analysis with the 7386-marker map, updated BLUPs of 4413 RILs, and iterative optimization (model 7 in Table 2-1), identified 33 QTL (referred to as ‘model 7 QTL’) associated with variation in resistance to SLB in the NAM panel, with support intervals averaging 4.6 cM and ranging from 1.8 cM to 14.0 cM (Table 2-2). Combined, the 33 QTL
from model 7 were associated with 84% of the phenotypic and 98% of the genotypic variation for resistance to SLB. All model 7 QTL had small effects; absolute values of significant (p <0.05) allele effects averaged 0.14 (range: 0.07 to 0.35) on the 1–9 scale used for quantifying resistance [8] (Figure 2-1; Additional file 1: Tables S2 and S3). The two model 7 QTL with the largest resistance effects across RIL families mapped to 43.4 cM and 54.4 cM on chromosome 3.

The JLM results are generally similar to those reported by Kump et al. [8], who identified 32 QTL (here, model 1), each of which had relatively small allelic effects of similar magnitude to model 7 QTL. To directly compare the positions of QTL from models 1 and 7, we interpolated model 1 QTL peak positions onto the 7386-marker map according to the AGP v2 physical positions of the SNPs identified as model 1 QTL peaks (Table 2-2). The median distance between the closest matching QTL peaks of model 1 and 7 was 5.6 cM (Table 2-2). Smaller-effect QTL tended to have larger discrepancies in position between the models.

Prediction accuracy of JLM QTL models developed from original and updated inputs were compared by cross-validation. A small number of RILs used in model 7 were not used in model 1 because of missing data in the original linkage map, so we identified a set of 4354 RILs in common between the original and updated data sets. QTL positions from models 1 and 7 were fit to random subsets of these RILs to re-estimate the allele effects and predict phenotypes in the validation sets. On average, across 100 randomly sampled training and validation sets, model 7 had a significantly (p <0.0001) greater prediction correlation coefficient (r =0.86 ± 0.01) than model 1 (r = 0.83 ± 0.01; Figure A-1).
Sensitivity analysis

Seven different QTL models (including models 1 and 7 previously described) were generated using different combinations of model inputs. The inputs that varied included the genetic map (1106-marker map vs. 7386-marker map), phenotypes (“original BLUPs” vs. “updated BLUPs”), and RIL sample sizes (4354 vs. 4431 vs. 4694 RILs; Table 2-1). QTL peak locations were generally concordant among the seven models tested (Figure 2-2).

Predicted phenotypic values of RILs based on JLM QTL models were also similar among the models, with all correlation coefficients between model predictions greater than 0.94.

In contrast to the general stability of QTL localization and predicted phenotypic values observed among models varying for different inputs, we observed substantially lower correlations between chromosome-specific residual outcomes from these JLM QTL models (Table 2-3). Genetic map density (1106-marker map vs. 7386-marker map) had the greatest impact on the correlation between residual outcomes, followed by sample size (4694 vs. 4354 RILs), and then the different methods for handling the flowering time covariate in generating the RIL phenotypes (original vs. updated BLUPs). Using identical algorithms and phenotype inputs but different marker densities (1106 vs. 7386) produced a correlation of 0.79 (models 2 and 5, models 4 and 6; Table 2-3). Dropping 340 RILs (4694 vs. 4354 RILs) produced a correlation of 0.85 (models 1 and 2, models 3 and 4; Table 2-3). The decrease in correlation due to sample size was attributed to total sample size per se and not likely to representation of particular families, since the proportional representation of each family in the total NAM family did not change due to dropping lines (Table 4 in File A-1). Although original and updated BLUPs were highly correlated as input values (r = 0.99), their small
differences resulted in QTL model residuals with much greater differences, reflected in correlations of 0.80 to 0.91 between residual outcomes differing only for original vs. updated BLUPs (Model comparisons 1 vs 3, 2 vs 4, and 5 vs 6; **Table 2-3**). When multiple inputs were changed simultaneously, the correlation between residuals diminished more (r ranging from 0.77 to 0.85, **Table 2-3**).

**GWAS for SLB resistance in the NAM panel**

The updated GWAS (model E, **Table 2-1**) was performed using the 28.5 M combined HapMap 1 and 2 SNPs and RDVs with phenotype values adjusted for unlinked QTL from model 7. A total of 192 variants were significantly associated with SLB resistance at RMIP ≥0.05. ([Figure 2-3](#) and Table 5 in [File A-1](#)). Model 7 QTL support intervals were highly enriched for significant associations: whereas only 17% of all variants tested localized within the QTL support intervals, 98 of 192 (51%) significantly associated variants were in QTL intervals, and 32 out of 33 model 7 QTL support intervals contained one or more of the significant associations. Genes containing or adjacent to the 26 significantly associated variants were identified (**Table 2-4**). Twenty-four candidate genes underlying 25 variants were identified from the B73 reference genome, but no gene was found within 100 kb of SNP S10_64647379 (**Table 2-4**). Eighteen of 24 candidate genes were located in model 7 QTL support intervals ([Figure 2-3](#)).
Comparison of multiple NAM GWAS results for SLB resistance

The original NAM GWAS for SLB resistance used 1106-marker map, model 1 QTL residuals, and 1.6 M HapMap 1 SNPs (GWAS model A, Table 2-1), and identified 245 significant SNP loci at p <1e-04 with RMIP ≥0.05 [8]. Comparing the positions of the 245 SNP associations identified with RMIP ≥0.05 in model A to the 192 variant associations identified in updated model E, only 6% of the total set colocalized within 10-kb windows between the two analyses. The three-fold enrichment of associations within QTL intervals compared to all tested variants observed in the updated GWAS model E (51% of associations vs 17% of all tested) was greater than the two-fold enrichment observed in the original GWAS model A (31% of associations vs 15% of all tested).

To evaluate NAM GWAS sensitivity to different GWAS inputs i.e. genetic map, residual inputs, and GWAS marker density, four separate GWAS models were compared (Table 2-1). The comparisons of GWAS analyses based on different input data sets indicate that both linkage map density (1106- vs. 7386-marker map) and GWAS marker density (1.6 M SNPs vs. 28.5 M SNPs and RDVs) influence the GWAS results dramatically (Table 2-5; Table 6 in File A-1). When both linkage map and HapMap marker densities were changed simultaneously, less than 25% of associations at RMIP ≥0.05 and 10% of associations at RIMP ≥0.25 localized within 200 kb of each other across analyses (Table 2-5).

RMIP values for each variant are determined based on the proportion of data subsamples in which the variant was selected in a multiple regression model at a given p-value, so RMIP values are subject to stochastic variation in the random sampling of data sets. Therefore, some of the inconsistency among analyses may be due simply to the process of
randomly sampling the complete data set 100 times for each analysis. We estimated the consistency of RMIP values from NAM GWAS (based on re-sampling 80% of RILs) by conducting five separate model E GWAS analyses for all variants on chromosomes 3 and 10, chosen to represent chromosomes with different numbers of QTL. Each of the five analyses included a unique set of 100 random data samples of 80% of RILs to calculate RMIP values from independent runs. Pairwise comparisons of association analyses indicate that 72% - 81% of variants with RMIP ≥0.05, and 88% - 100% of SNPs with RMIP ≥0.25 overlapped within 10-kb windows matched between each sampling procedure (Table 7 in File A-1). To estimate the consistency of RMIP values with more subsamples, five separate model E GWAS analyses were conducted by analyzing sets of 200 samples and 500 samples on chromosomes 3 and 10. Pairwise comparisons showed that 81% - 89% of associations with RMIP ≥0.05, and 89% - 100% of SNPs with RMIP ≥0.25 overlapped within 10-kb windows when 200 subsamples were used to compute RMIP (Table 7 in File A-1). About 93% - 96% of variants with RMIP ≥0.05, and 100% of SNPs with RMIP ≥0.25 overlapped within 10-kb windows when 500 subsamples were used (Table 7 in File A-1).
Discussion

The maize NAM panel is a community genetic resource for dissecting the genetic architecture of quantitative traits. It allows for the combination of high power in conventional QTL linkage mapping and high resolution in genome-wide association mapping [23,24]. We identified 33 QTL with small additive effects across 25 NAM families. The 7386-marker map has 6.7 times the marker density of the original 1106-marker map, and, importantly, it has uniform marker spacing at 0.2 cM without gaps. The uniform spacing and denser map improved the power and precision of QTL mapping in our analysis. One of the strongest effect QTL had a discrepancy in position between QTL models 1 and 7 because it localized to a >10 cM gap in the 1106-marker map. The current study resolved what was previously mapped as a single QTL detected at 50.0 cM of chromosome 3 (at the edge of the map gap) into two separate QTL at 43.4 and 54.4 cM. Evidence from fine mapping and high-resolution bi-parental QTL studies [25-27] supports the existence of the two rather than the one QTL. Thus, it appears that the increased map density improved precision of QTL position estimates despite the loss of sample size that occurred by dropping 281 RILs in model 7. This study only evaluated relatively small changes in sample size, we expect much larger effects on results for more substantial sample size changes, as shown by simulation [28].

The high and uniform density of the 7386-marker map eliminates the need for interval mapping, but increases the risk of selecting collinear markers and overfitting the QTL models. In this study, we recognized and corrected collinearity problems that occurred during automated stepwise selection by inspecting results for some diagnostic signatures of collinearity: inflated allele effect estimates at marker pairs within 10 cM of each other, and
inflated standard errors of the allele effects. Linkage disequilibrium is extensive within mapping families, such that the increasing power and resolution of QTL mapping plateaus at some point with increasing numbers of markers [29]. We believe that there would be diminishing returns from a more dense linkage map than the current, 0.2 cM dense linkage map (7386-marker map), since we would not expect further QTL resolution and predictive accuracy from more markers, while the collinearity issues and computational burden would continue to increase [30].

Joint linkage–association mapping has been applied in a two-stage process. In the first stage, the phenotype is regressed on genetic markers to identify QTL and estimate their effects. In the second stage, the residual values adjusted by unlinked QTL are then regressed on dense HapMap SNP genotypes, one chromosome at a time. The two-stage approach has several practical advantages in that it is convenient to implement as well as computationally efficient. However, the sensitivity of JLM residual outcomes to first-stage inputs contributes to variation in second-stage GWAS outputs. In this study, changes in inputs (genotypes, phenotypes, sample size, or combinations of them) to the JLM QTL modeling had relatively minor effects on the QTL position estimates (Figure 2-3). The resulting predicted phenotypic values from the different models had average correlations of $r = 0.95$, but the average correlations between corresponding residual values were somewhat lower: $r = 0.82$. After removing QTL effects from 9 out of 10 chromosomes, the chromosome-specific residuals are composed of genetic effects (‘signal’) from just one chromosome plus error effects (‘noise’) from all chromosomes. The chromosome-specific residual values are convenient for SNP testing, because they remove the effects of QTL on other chromosomes, but as a
consequence, the residuals represent a lower signal to noise ratio compared to the original phenotypic values. This is unavoidable because each chromosome contributes only a fraction of the total genetic effects to a complex trait. The sensitivity of the residual outcome values to first-stage inputs highlights the difficulty of identifying individual variant effects that account for only a small proportion of the total heritability.

GWAS results were unstable due to changes in the initial inputs to the QTL analysis as well as to the marker set used for association testing. Only about 35% of the associated variants with RMIP \( \geq 0.05 \) localized to common 100-kb windows between analyses when the HapMap marker set was changed (from 1.6 M SNPs to 28.5 M variants). The proportion of overlapping significant variants in 100-kb windows was even lower (20-30%) when using different genetic maps but the same GWAS markers. Changing both genetic map and GWAS marker inputs reduced the proportion of overlapping significant SNPs to between 13 and 21% (Table 2-5). Only four candidate genes contained variants that were significant (RMIP \( \geq 0.05 \)) across all four GWAS analyses in a 10-kb window (Table 5 in File A-1). The generally poor correspondence between GWAS results of the four analyses may be due in part to the highly polygenic nature of the trait. If many sequence variants with small effects control the trait, but only a small proportion of the SNP associations pass stringent thresholds, then relatively small perturbations in analysis inputs could cause substantial differences in the particular SNPs declared as significant.

Eighteen of the 24 candidate genes identified with GWAS model E were in QTL intervals (Table 2-4). Most of the candidate gene homologs have been reported to be involved in disease resistance (Table 8 in File A-1). Leucine-rich repeat transmembrane
protein kinases (LRR-PK) regulate a wide range of developmental and defense related processes, such as hormone perception, host specific and non-host specific defense response, and wounding response [31]. The well-studied LRR-PK genes include rice Xa21 (Xanthomonas resistance 21) [32,33], Arabidopsis FLS2 (flagellin sensitive 2) [34], and the Arabidopsis elongation factor Tu receptor (EFR) [35]. The Arabidopsis Cytochrome P450 gene was previously identified as associated with resistance to necrotrophic fungi and aphids [36-38]. The plant U-box 13 (spotted leaf1) mutant confers enhanced non race-specific resistance to fungal and bacterial pathogens in rice [39,40]. The lysine histidine transporter 1 (LHT1) mutant of Arabidopsis affects resistance to a broad spectrum of pathogens [41].

The SLB-associated variants identified here by JLM-GWAS were found residing within some QTLs reported in previous studies. Maize genome bins 3.04, 6.01, and 9.02/03 had been identified from different studies contributing to major effects on SLB resistance [42-48]. The association with the highest RMIP (0.9) was localized to 32,885,733 bp of chromosome 3, within a QTL region identified in other populations [42,48]. The nearest annotated gene to the associated SNP is ~60 kb downstream and encodes an RNA recognition motif (RRM)-containing protein (GRMZM2G132936). Another strong variant association (RMIP = 0.4) was in the LHT1 gene (GRMZM2G127342) on chromosome 6; this gene was previously suggested to be the causal factor for the classically defined rhml1 locus based on QTL fine-mapping [47]. One of the most significant SNPs (RMIP =0.8) is 332 bp downstream of GRMZM2G099363 encoding a caffeoyl-CoA O-methyltransferase (CCoAOMT), within a QTL region on chromosome 9 identified in other populations [42,49]. CCoAOMT has been reported to participate in lignin biosynthesis in plants [50-52]. Lignin
has a particular role as a physical barrier against external pathogens, limiting the penetration of pathogens into host cells. In conclusion, we recommend use of the updated JLM QTL (model 7) and GWAS (model E) results in the search for candidate genes controlling resistance to Southern leaf blight. The updated QTL model had better prediction accuracy than the original model, and the updated GWAS provided substantially higher marker density, which is expected to provide a better chance of identifying variants in linkage disequilibrium with causal variants. Further work will attempt to validate biologically the effects of candidate genes with the strongest statistical evidence to provide more detailed insight into the genetic basis of SLB resistance. Finally, our results highlight the difficulties and contingencies of reliably identifying genomic variants with small effects on quantitative traits.
Methods

Analysis of phenotype data

Kump et al. [8] fit a multivariate (repeated measurements) mixed model to phenotype data collected on resistance of NAM RILs to SLB in three environments. The data and model described by Kump et al. (S1 supplement in [8]) were updated for this study, correcting errors identified in the data file and applying a different approach to modeling the flowering time covariate.

An error was found in the coding levels or dummy values for incomplete blocks nested in population blocks to which entries had been assigned in the 07NC trial. Correcting this led to a significant effect of incomplete blocks in 07NC, which was therefore included in the updated model (BLUPs from the previous analysis are expected to be less precise, because the random effects of incomplete blocks were confounded with the random effect of RILs). There were two additional minor errors that were corrected: i) one plot in 06NC was associated with an incorrect incomplete block; ii) two plots with SLB data from 07NC had their entry information swapped. Instead of fitting the effect of flowering time on disease resistance as a quadratic function for the entire dataset [8], the new model considered the relationship in a more specific manner. In a pre-analysis step of the data for each environment, ANOVA was used to compare the fit of linear, quadratic and cubic functions relating flowering time (measured as days to anthesis) to disease resistance for each rating x environment x population-specific combination. The final multivariate mixed model was as follows:
\[ y = X_\beta + X_\xi + X_{e+f+\phi} + Z_{h(R+S)} b + Z_{r(R+S)} i + Z_{c(S)} c + Z_{r(S)} r + Z_{p} p + Z_{s} p \xi* p + Z_{e(R)} e + Z_{s} e*\xi* e + \epsilon \]

where \( y = [y_1', y_2']' \) corresponding to the SLB disease ratings 1 and 2, \( \beta \) corresponding to the overall mean fixed effects for ratings 1 and 2, and so on for the rest of the terms in the model where \( \xi \) = environment fixed effects; \( \phi \) = rating \times environment \times population level flowering time fixed effects (linear or linear + quadratic effects [cubic was never significant in the pre-analysis]); \( b \) = population block nested in replication \times environment random effects; \( i \) = incomplete block nested in population block \times replication \times environment random effects; \( c \) = columns nested in environment random effects; \( r \) = rows nested in environment random effects; \( p \) = population random effects; \( \xi* p \) = environment \times population random effects; \( e \) = entry (RIL) nested in population random effects; \( \xi* e \) = environment \times entry (RIL) nested in population random effects; and \( \epsilon \) = residual random effects. As in [8], only those random factors significant (LRT, \( p < 0.10 \)) in single environment analyses were retained in the multi-environment model. Modeling of variance-covariance structures was also the same as [8], whereby a bivariate/unstructured covariance was modeled on all random terms in an environment-specific manner for nested design factors (blocking effects, row and column effects, and residual effects) and across all environments for the cross-classified factors (population, population \times environment, and entries nested in populations [fitted per population]). For example, the covariance assumed for entries nested in population 1 is as follows, where \( i \) indicates the specific cohort (\( i = 1 \) to 28; 1–26 corresponds to the biparental subpopulations of NAM):
The multivariate mixed model was used to estimate BLUPs or phenotypic values as an equally weighted index of the two scores for the first stage analysis of the JLM-GWA procedure. We refer to the BLUPs used in [8] as “original BLUPs” and those calculated for this study as “updated BLUPs.”

Genotyping and genetic linkage map

NAM RILs were genotyped with the GBS approach by the Institute of Genomic Diversity and the Buckler Lab at Cornell University [15, 16, 53]. A consensus genetic map was constructed based on 7386 SNPs segregating in the NAM RILs. RILs with high levels of homozygosity and reliable genotype scores were used in the construction of the updated genetic map and further genetic analysis. A small proportion of RILs included in the previous analysis [8] were excluded from the current analysis because their GBS data were of insufficient quality to permit reliable genotype calling. A set of markers representing the linkage map at a uniform distance of 0.2 cM was retained from the larger set of GBS SNPs scored, with missing marker data imputed based on linkage intensities and flanking non-missing markers using the Full Sib Family Haplotype Imputation algorithm [54].
Joint linkage mapping was performed using stepwise selection implemented in Proc GLMSelect in SAS v9.3. Thresholds for markers to enter and stay in the model at each step were set at $\alpha = 0.0001$, as used previously [5,8]. Family main effects were always included and marker effects were nested in families. Kump et al. [8] reported no significant epistatic interactions after accounting for additive effects. We thus modelled the genetic architecture with a pure additive model:

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

where $Y$ is an $N \times 1$ column vector of the updated SLB resistance best linear unbiased prediction (BLUP) values; $A$ is an $N \times P$ incidence matrix relating each individual RIL to its corresponding family, $\mu$ is a $P \times 1$ column vector of family main effects; $X_i$ is a $N \times P$ matrix relating each RIL’s genotype score at locus $i$ to its corresponding family-specific allele effect, the elements of $X_i$ are coded 0 for lines homozygous for the B73 reference allele, and 2 for homozygotes with the alternate parental allele, 1 for heterozygotes, and a non-integer between 0 and 2 for the imputed recombinants as described above; $\beta_i$ is a $P \times 1$ column vector of the family-specific additive effects associated with locus $i$ relative to B73, $k$ is the number of significant loci retained in the final model; and $\varepsilon$ is a $N \times 1$ column vector of errors.

High collinearity hinders the selection of markers closest to true QTL positions in linear regression and may result in the selection of pairs of tightly linked loci, with biased effect estimates. Collinearity between tightly linked markers selected in the model was
diagnosed based on inflation of standard errors associated with QTL effects and suspiciously large magnitudes of QTL effects of opposite signs for markers located within 10 cM of each other [55]. When obvious collinearity between a pair of markers was detected, one of the problematic predictors was removed from the model, and further selection was implemented with the remaining predictors retained in model. The diagnostic process was repeated until all predictors were free of collinearity.

After initial model selection, the model was further optimized through an iterative process in which one candidate marker was dropped from the full model and replaced with an adjacent marker, and the process was repeated sequentially, fitting each marker within 10 cM of the original peak QTL position one at a time in place of the original marker. The position that resulted in the maximal R2 was recorded. This process was then repeated for each QTL, and then the entire process was iterated until the model stabilized (no marker positions changed). Allele (nested) effects for each QTL within family were estimated in the final optimized QTL model.

To construct support intervals associated with each QTL, a marker immediately adjacent to the QTL on the left side was added to the full model, and the p for the QTL peak itself was noted. Typically, addition of an adjacent collinear marker reduced the Type III sum of squares for the tested peak QTL marker, resulting in a p > 0.05 for the peak marker. An iterative process was followed by moving the position of the added marker sequentially along the linkage map to the left side of the QTL peak marker until the tested QTL peak marker became significant at the p < 0.05. The point at which the QTL peak regained significance signifies the limit of that QTL effect, so the position of the added marker was considered the
left boundary of the QTL support interval. The right support interval boundary was identified the same way.

**Internal cross-validation of QTL model prediction accuracy**

To compare the prediction accuracies of the models 1 and 7 QTL in the NAM panel, we estimated prediction accuracies by a cross validation scheme of random subsampling. The QTL were fixed at the positions selected in the two final models, but we re-estimated the allele effects for the two models in each subsample. Models were compared using a subset of 4354 RILs common between the 4694 RILs used by Kump et al. [8] and the 4413 RILs used in this analysis. Within each replication of the cross-validation, 80% of RILs (3484) of each family were randomly sampled without replacement from the 4354 common RILs to use as a training set. The QTL positions from models 1 or 7 were fit to the data and the QTL allele effects were estimated for each of the two models using their own linkage maps. The remaining 20% of the RILs (870) were held out as a validation set, and the prediction accuracies were estimated as the within-family Pearson’s correlation between predicted values and actual SLB resistance BLUP values. Prediction accuracies were then averaged over 100 random validation sets.

**Correlation analysis of JLM residual outcomes and predicted phenotypic values**

GWAS is conducted in the NAM panel on a chromosome-by-chromosome basis; tests for association of SNPs on a particular chromosome are conducted using residuals values from the regression of the SLB phenotypes on the final JLM QTL model after dropping any
QTL markers on the chromosome in question (a reduced JLM QTL model). To evaluate the sensitivity of JLM residual outcomes to changes in three JLM inputs, we designed a series of computational experiments controlling for differences in genetic maps, RIL sample sizes and phenotypic data, and combinations of them. A series of six QTL models were constructed with those different inputs. Correlations between the chromosome-specific residual sets were computed for all 15 pairs of model comparisons, as well as for the corresponding predicted phenotypic value sets. Average correlations over 10 chromosomes were computed and compared between residual sets and predicted BLUP values. The QTL peaks and support intervals for those six models involving the sensitivity analysis and model 7 were positioned on the 7386-marker map to examine concordance of the QTL mapping.

**GWAS and identification of candidate genes**

The maize HapMap 1 data set includes 1.6 M SNPs polymorphic between B73 and at least one other founder line. We also used an updated version of the maize HapMap 2 data set (dated March 28, 2012), which includes 27.3 M SNPs. About 60% of HapMap 1 SNPs are not included in HapMap2. Therefore, we combined the two SNP data sets and retained 28.2 M unique SNPs. In a few cases, the founder allele calls for the same SNP in the two data sets differed; we retained the HapMap 1 allele calls in such cases, to permit direct comparison to HapMap 1 results. We also included 228,212 read depth variant (RDV) calls reported by Chia et al. [14], resulting in a total of 28.5 M variants tested for association with SLB resistance. Each read depth variant represents either an increase or decrease of log2 or more in read depth along a 10 kb window with respect to the reference B73 genome.
The RIL residuals from reduced JLM models (dropping QTL from one particular chromosome) represent the phenotype values for which most of the genetic effects are due to sequence variation on the chromosome considered. GWAS was performed by randomly sampling 80% of the RILs from each family and analyzing chromosome by chromosome with forward stepwise selection of the combined HapMap 1, 2 SNPs and RDVs at \( p < 1 \times 10^{-7} \). An exception to this was the original GWAS model A, which used a threshold of \( p < 1 \times 10^{-4} \) for inclusion in the chromosome-specific model [8]. The much higher number of variants available for testing in the HapMap 2 data set requires a more stringent threshold to prevent overfitting of GWAS models. This subsampling and analysis procedure was repeated 100 times. The resample model inclusion probability (RMIP) was calculated for each variant as the proportion of 100 data samples in which the variant was selected in the regression model. Candidate genes encompassing or near (within \( \sim 100 \) kb) the variants with strong association signals (RMIP \( \geq 0.25 \)) were identified in the maize B73 genome using the genome browser at http://www.maizegdb.org [56].

**GWAS result comparison**

SNPs associated with variation in SLB resistance at two RMIP thresholds (RMIP \( \geq 0.05 \) or RMIP \( \geq 0.25 \)) using GWAS model A were compared to those identified with model E. Significant variants from the two GWAS were compared on the basis of their positions on AGP version 2 maize B73 reference genome. Positions of variants identified in different analyses were compared using each of 10-kb, 100-kb and 200-kb windows. The variant match rate was calculated as the ratio of variants from different analyses found within
common windows to all significant variants for a pair of analyses. To determine the relative importance of changes in different inputs to GWAS, each combination of genetic map, GWAS marker density, and residual inputs were used to perform separate GWAS analyses, and results were compared on the basis of position matches in windows.

To determine the effect of the number of data on the stability of RMIP estimates, we analyzed 1000 random samples of 80% of all NAM families and performed model E GWAS analyses on chromosomes 3 and 10. We chose chromosomes 3 and 10, as they represent with the range of numbers of QTL mapped on each chromosome while greatly reducing the computational burden compared to whole genome analyses. We then compared the consistency of RMIP values from five disjoint samples of 100 resamples each, five disjoint samples of 200 each, and five random partitionings of the 1000 analyses into pairs of 500 resamples. The rate of matching associations at each of 10-kb, 100-kb, and 200-kb windows with RMIP ≥0.05 and RMIP ≥0.25 was computed for all pairwise comparisons of the five association analyses at each sample size.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

YB designed the sensitivity and cross-validation studies and performed joint linkage analyses and helped to draft the manuscript. QY conducted genome-wide association studies and helped to draft the manuscript. PBK conceived the study. RW performed mixed models
analyses and helped to draft the manuscript. JH participated in study design and analysis and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Sharon Mitchell, Jeff Glaubitz, Peter Bradbury, and Edward Buckler for the updated NAM linkage map. Research was supported by United States National Science Foundation grants IOS-1127076, IOS-0820619, and IOS-1238014. YB is a graduate student at NCSU supported by NSF grant IOS-0820619 and a Monsanto Co. fellowship. QY is a post-doctoral researcher at NCSU supported by NSF grant IOS-1127076. PBK and JH are research geneticists with the USDA-ARS. RW is faculty of University of Delaware. NSF grants IOS-1127076 supports research in the laboratories of PBK, RW, and JH. NSF grants IOS-0820619 and IOS-1238014 support research in the laboratory of JH.
Figure 2-1 Heat map of additive effect estimates of 25 founder parent alleles for QTLs of SLB resistance relative to B73. QTL are identified by their genetic locations in the consensus genetic map (7386-marker map); effect estimates for each parental allele are indicated by color blocks. Negative cM values for markers indicate that they are distal to the first marker from the original NAM linkage map on that chromosome.
Figure 2-2 QTL positions on the ten maize chromosomes from seven joint linkage mapping models. Positions in cM are based on 7386-marker map. QTL bar heights are proportional to their partial R2, blue-colored margins denote the QTL support intervals, and black spikes denote the QTL peak positions. For the definitions of model inputs, see Table 2-1.
Figure 2-3 Manhattan plots from genome-wide association analysis for SLB across the ten chromosome pairs of maize. The dashed horizontal line in red depicts the resample model inclusion probability (RMIP) thresholds of 0.05 and 0.25. Eighteen candidate genes underling the most robust GWAS hits (RMIP ≥0.25) located in QTL support intervals are indicated (Table 2-4).
## Tables

### Table 2-1 Inputs for joint linkage mapping QTL analysis and GWAS models

<table>
<thead>
<tr>
<th>QTL Model</th>
<th>Phenotype</th>
<th>No. markers in linkage map</th>
<th>No. NAM RILs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>original&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1106</td>
<td>4694</td>
</tr>
<tr>
<td>2</td>
<td>original</td>
<td>1106</td>
<td>4354</td>
</tr>
<tr>
<td>3</td>
<td>updated</td>
<td>1106</td>
<td>4694</td>
</tr>
<tr>
<td>4</td>
<td>updated</td>
<td>1106</td>
<td>4354</td>
</tr>
<tr>
<td>5</td>
<td>original</td>
<td>7386</td>
<td>4354</td>
</tr>
<tr>
<td>6</td>
<td>updated</td>
<td>7386</td>
<td>4354</td>
</tr>
<tr>
<td>7</td>
<td>updated</td>
<td>7386</td>
<td>4413</td>
</tr>
</tbody>
</table>

**GWAS models**

<table>
<thead>
<tr>
<th>GWAS Model</th>
<th>QTL model used to adjust phenotypes</th>
<th>Variants tested for association</th>
<th>p value threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>1.6 M HM1 SNPs</td>
<td>1E-4</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1.6 M HM1 SNPs</td>
<td>1E-7</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>28.2 M HM1/2 SNPs +0.2 M RDVs</td>
<td>1E-7</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
<td>1.6 M HM1 SNPs</td>
<td>1E-7</td>
</tr>
<tr>
<td>E</td>
<td>7</td>
<td>28.2 M HM1/2 SNPs +0.2 M RDVs</td>
<td>1E-7</td>
</tr>
</tbody>
</table>

<sup>a</sup> QTL model 1 and GWAS model A results were previously published [8].

<sup>b</sup> Original phenotype inputs were RIL BLUPs across 3 environments, based on a model with a common flowering time covariate; updated phenotype inputs were RIL BLUPs across 3 environments based on a model with environment-specific flowering time covariates.
Table 2-2 Physical and genetic positions for QTL peaks and support intervals (SI) mapped using current updated phenotypes and linkage map (model 7) and comparison to QTL previously reported (model 1).
Step included denotes the regression model building step in which each QTL was selected for inclusion in model 7. SI overlapped indicates if models 1 and 7 QTL SI overlapped: y for overlapped, n for not overlapped and c for very close (within 1.5 cM) but not overlapped. Distance between QTL indicates the cM distance between the peaks of nearest QTL from models 1 and 7.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Peak position (AGP v2 bp)</th>
<th>Peak position (cM)</th>
<th>SI Map Position (cM)</th>
<th>Step Included</th>
<th>SI Overlapped</th>
<th>Distance between QTL peaks (cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90,443,174</td>
<td>85.2</td>
<td>84.2 - 86.2</td>
<td>3</td>
<td>y</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>218,082,692</td>
<td>127.6</td>
<td>125.6 - 129.6</td>
<td>8</td>
<td>y</td>
<td>1.9</td>
</tr>
<tr>
<td>1</td>
<td>251,723,948</td>
<td>146.8</td>
<td>144.8 - 147.8</td>
<td>18</td>
<td>y</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>283,549,061</td>
<td>177</td>
<td>175 - 179</td>
<td>22</td>
<td>n</td>
<td>29.2</td>
</tr>
<tr>
<td>2</td>
<td>7,180,393</td>
<td>21.4</td>
<td>19.4 - 23.4</td>
<td>21</td>
<td>n</td>
<td>20.2</td>
</tr>
<tr>
<td>2</td>
<td>36,838,070</td>
<td>65.4</td>
<td>64.4 - 66.4</td>
<td>5</td>
<td>y</td>
<td>2.3</td>
</tr>
<tr>
<td>2</td>
<td>206,275,294</td>
<td>111.2</td>
<td>110.2 - 113.2</td>
<td>23</td>
<td>n</td>
<td>30.8</td>
</tr>
<tr>
<td>3</td>
<td>5,348,237</td>
<td>19.6</td>
<td>18.6 - 22.6</td>
<td>13</td>
<td>y</td>
<td>5.9</td>
</tr>
<tr>
<td>3</td>
<td>16,246,035</td>
<td>43.4</td>
<td>41.4 - 44.4</td>
<td>1</td>
<td>n</td>
<td>6.6</td>
</tr>
<tr>
<td>3</td>
<td>31,533,927</td>
<td>54.4</td>
<td>53.4 - 55.4</td>
<td>10</td>
<td>c</td>
<td>4.4</td>
</tr>
<tr>
<td>3</td>
<td>170,374,260</td>
<td>79.4</td>
<td>78.4 - 84.4</td>
<td>14</td>
<td>n</td>
<td>12.9</td>
</tr>
<tr>
<td>3</td>
<td>214,568,867</td>
<td>124.4</td>
<td>123.4 - 126.4</td>
<td>26</td>
<td>c</td>
<td>6.8</td>
</tr>
<tr>
<td>3</td>
<td>219,498,075</td>
<td>134</td>
<td>132 - 137</td>
<td>6</td>
<td>y</td>
<td>2.8</td>
</tr>
<tr>
<td>4</td>
<td>1,892,489</td>
<td>2.4</td>
<td>-3.6 - 7.4</td>
<td>29</td>
<td>y</td>
<td>4.6</td>
</tr>
<tr>
<td>4</td>
<td>141,422,921</td>
<td>59.8</td>
<td>54.8 - 60.8</td>
<td>20</td>
<td>c</td>
<td>7.9</td>
</tr>
<tr>
<td>4</td>
<td>181,935,681</td>
<td>94</td>
<td>91 - 102</td>
<td>17</td>
<td>n</td>
<td>24.2</td>
</tr>
<tr>
<td>5</td>
<td>15,138,119</td>
<td>45.2</td>
<td>44.2 - 49.2</td>
<td>33</td>
<td>y</td>
<td>8.2</td>
</tr>
<tr>
<td>5</td>
<td>36,905,989</td>
<td>58</td>
<td>55 - 59</td>
<td>11</td>
<td>y</td>
<td>4.6</td>
</tr>
<tr>
<td>5</td>
<td>158,046,075</td>
<td>74.2</td>
<td>72.2 - 77.2</td>
<td>19</td>
<td>y</td>
<td>3.0</td>
</tr>
<tr>
<td>5</td>
<td>200,161,433</td>
<td>106.6</td>
<td>104.6 - 109.6</td>
<td>27</td>
<td>n</td>
<td>18.1</td>
</tr>
<tr>
<td>5</td>
<td>214,137,041</td>
<td>144.4</td>
<td>140.4 - 154.4</td>
<td>31</td>
<td>n</td>
<td>19.7</td>
</tr>
<tr>
<td>6</td>
<td>6,929,855</td>
<td>-0.6</td>
<td>-1.4 - 0.4</td>
<td>9</td>
<td>y</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>144,806,691</td>
<td>56.6</td>
<td>53.6 - 58.6</td>
<td>28</td>
<td>n</td>
<td>56.8</td>
</tr>
<tr>
<td>7</td>
<td>8,270,900</td>
<td>31.2</td>
<td>28.2 - 34.2</td>
<td>30</td>
<td>y</td>
<td>3.0</td>
</tr>
<tr>
<td>7</td>
<td>142,429,440</td>
<td>74.8</td>
<td>73.8 - 76.8</td>
<td>25</td>
<td>c</td>
<td>5.2</td>
</tr>
<tr>
<td>8</td>
<td>36,978,572</td>
<td>51.2</td>
<td>50.2 - 53.2</td>
<td>4</td>
<td>y</td>
<td>1.4</td>
</tr>
<tr>
<td>8</td>
<td>118,435,453</td>
<td>63.2</td>
<td>61.2 - 67.2</td>
<td>24</td>
<td>n</td>
<td>10.6</td>
</tr>
<tr>
<td>8</td>
<td>166,705,312</td>
<td>102.2</td>
<td>101.2 - 104.2</td>
<td>15</td>
<td>y</td>
<td>1.3</td>
</tr>
<tr>
<td>9</td>
<td>16,361,287</td>
<td>28.4</td>
<td>27.4 - 29.4</td>
<td>7</td>
<td>y</td>
<td>0.1</td>
</tr>
<tr>
<td>9</td>
<td>109,899,486</td>
<td>52.6</td>
<td>51.6 - 54.6</td>
<td>2</td>
<td>n</td>
<td>5.6</td>
</tr>
<tr>
<td>10</td>
<td>2,040,278</td>
<td>-1.6</td>
<td>-2.6 - 3.4</td>
<td>32</td>
<td>n</td>
<td>38.3</td>
</tr>
<tr>
<td>10</td>
<td>75,958,884</td>
<td>37.8</td>
<td>34.8 - 38.8</td>
<td>12</td>
<td>y</td>
<td>1.1</td>
</tr>
<tr>
<td>10</td>
<td>135,354,773</td>
<td>59.8</td>
<td>58.8 - 60.8</td>
<td>16</td>
<td>n</td>
<td>23.1</td>
</tr>
</tbody>
</table>
Table 2-3 Input changes in joint linkage mapping QTL modeling affect the values of the chromosome-specific residuals more profoundly than the corresponding predicted phenotypic values

<table>
<thead>
<tr>
<th>Model</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4694 RILs, original BLUPs, 1106 map</td>
<td>0.88</td>
<td>0.91</td>
<td>0.81</td>
<td>0.81</td>
<td>0.77</td>
</tr>
<tr>
<td>2</td>
<td>SZ</td>
<td>4354 RILs, original BLUPs, 1106 map</td>
<td>0.85</td>
<td>0.87</td>
<td>0.82</td>
<td>0.79</td>
</tr>
<tr>
<td>3</td>
<td>PH</td>
<td>PH, SZ</td>
<td>4694 RILs, updated BLUPs, 1106 map</td>
<td>0.82</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>4</td>
<td>PH, SZ</td>
<td>PH</td>
<td>SZ</td>
<td>4354 RILs, updated BLUPs, 1106 map</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>5</td>
<td>GN, SZ</td>
<td>GN</td>
<td>GN, PH, SZ</td>
<td>GN, PH</td>
<td>4354 RILs, original BLUPs, 7386 map</td>
<td>0.80</td>
</tr>
<tr>
<td>6</td>
<td>GN, PH, SZ</td>
<td>GN, PH</td>
<td>GN, SZ</td>
<td>GN</td>
<td>PH</td>
<td>4354 RILs, original BLUPs, 7386 map</td>
</tr>
</tbody>
</table>

Upper diagonal shows average correlation of residual outcome for each chromosome for 15 pairwise model comparisons. Lower diagonal shows the input(s) that differed in each pair of model comparison. “GN”, “PH”, and “SZ” denote the different genotype inputs: GN, linkage map (1106-vs. 7386-marker map); PH, phenotype inputs (original vs. updated BLUPs); and SZ, sample size (4354 vs. 4694 RILs), respectively. Diagonal shows the three inputs for each model.
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Physical position (AGPv2)</th>
<th>Allele</th>
<th>Effect</th>
<th>P-value</th>
<th>RMIP</th>
<th>Genic position</th>
<th>Nearest gene ID</th>
<th>Position of nearest gene (AGPv2)</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59772038</td>
<td>T/C</td>
<td>0.09</td>
<td>9.4E-10</td>
<td>0.35</td>
<td>Intergenic</td>
<td>N</td>
<td>GRMZM2G134671</td>
<td>59764838</td>
</tr>
<tr>
<td>1</td>
<td>90059294</td>
<td>C/T</td>
<td>0.08</td>
<td>8.5E-14</td>
<td>0.35</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM5G806839</td>
<td>90108398</td>
</tr>
<tr>
<td>2</td>
<td>34246559</td>
<td>A/C</td>
<td>0.11</td>
<td>1.8E-16</td>
<td>0.39</td>
<td>Intergenic</td>
<td>N</td>
<td>GRMZM5G876621</td>
<td>34361966</td>
</tr>
<tr>
<td>2</td>
<td>6938074</td>
<td>G/A</td>
<td>0.10</td>
<td>1.4E-22</td>
<td>0.53</td>
<td>3'UTR</td>
<td>Y</td>
<td>GRMZM2G022627</td>
<td>36936010</td>
</tr>
<tr>
<td>2</td>
<td>204757213</td>
<td>G/A</td>
<td>0.11</td>
<td>4.5E-09</td>
<td>0.41</td>
<td>Intergenic</td>
<td>N</td>
<td>GRMZM2G142932</td>
<td>204771189</td>
</tr>
<tr>
<td>2</td>
<td>207520001</td>
<td>CNV</td>
<td>0.11</td>
<td>4.4E-09</td>
<td>0.28</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G062156</td>
<td>207635000</td>
</tr>
<tr>
<td>2</td>
<td>16574125</td>
<td>T/C</td>
<td>0.12</td>
<td>1.8E-26</td>
<td>0.43</td>
<td>Non-synonymous-coding</td>
<td>Y</td>
<td>GRMZM2G463580</td>
<td>16575869</td>
</tr>
<tr>
<td>3</td>
<td>16575842</td>
<td>C/G</td>
<td>0.15</td>
<td>9.7E-14</td>
<td>0.46</td>
<td>5'UTR</td>
<td>Y</td>
<td>GRMZM2G132936</td>
<td>32957454</td>
</tr>
<tr>
<td>3</td>
<td>32885733</td>
<td>C/T</td>
<td>0.19</td>
<td>9.1E-16</td>
<td>0.9</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G132936</td>
<td>32957454</td>
</tr>
<tr>
<td>3</td>
<td>219885527</td>
<td>T/C</td>
<td>0.12</td>
<td>3.1E-12</td>
<td>0.28</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G2074572</td>
<td>219900229</td>
</tr>
<tr>
<td>3</td>
<td>2046350</td>
<td>CGG/---</td>
<td>0.08</td>
<td>3.8E-09</td>
<td>0.26</td>
<td>Non-synonymous-coding</td>
<td>Y</td>
<td>GRMZM2G131442</td>
<td>20440167</td>
</tr>
<tr>
<td>4</td>
<td>164110001</td>
<td>CNV</td>
<td>0.10</td>
<td>1.5E-11</td>
<td>0.42</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G111872</td>
<td>164120892</td>
</tr>
<tr>
<td>7</td>
<td>7001531</td>
<td>A/C</td>
<td>0.08</td>
<td>1.1E-15</td>
<td>0.4</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G127342</td>
<td>7036171</td>
</tr>
<tr>
<td>7</td>
<td>7958250</td>
<td>+/--</td>
<td>0.08</td>
<td>1.9E-08</td>
<td>0.3</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G021149</td>
<td>7954180</td>
</tr>
<tr>
<td>7</td>
<td>142997161</td>
<td>G/A</td>
<td>0.09</td>
<td>7.3E-09</td>
<td>0.28</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G901919</td>
<td>143003673</td>
</tr>
<tr>
<td>8</td>
<td>22229116</td>
<td>G/A</td>
<td>0.13</td>
<td>1.3E-08</td>
<td>0.35</td>
<td>Intergenic</td>
<td>N</td>
<td>GRMZM2G325612</td>
<td>22249167</td>
</tr>
<tr>
<td>8</td>
<td>62217745</td>
<td>T/A</td>
<td>0.10</td>
<td>6.3E-14</td>
<td>0.34</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G130101</td>
<td>62267220</td>
</tr>
<tr>
<td>8</td>
<td>171334544</td>
<td>A/G</td>
<td>0.06</td>
<td>1.2E-09</td>
<td>0.26</td>
<td>Intergenic</td>
<td>N</td>
<td>GRMZM2G476009</td>
<td>171332654</td>
</tr>
</tbody>
</table>

- **Table 2-4 Highly significant GWAS variants (RMIP ≥ 0.25) and their adjacent candidate genes**

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>59764838</td>
<td>59769739</td>
</tr>
<tr>
<td>90108398</td>
<td>90107598</td>
</tr>
<tr>
<td>34361966</td>
<td>34343251</td>
</tr>
<tr>
<td>36936010</td>
<td>36938218</td>
</tr>
<tr>
<td>204771189</td>
<td>204767728</td>
</tr>
<tr>
<td>207635000</td>
<td>207637178</td>
</tr>
<tr>
<td>16575869</td>
<td>16572429</td>
</tr>
<tr>
<td>32957454</td>
<td>32949608</td>
</tr>
<tr>
<td>219900229</td>
<td>219884844</td>
</tr>
<tr>
<td>220651805</td>
<td>220659525</td>
</tr>
<tr>
<td>143003673</td>
<td>143001781</td>
</tr>
<tr>
<td>164120892</td>
<td>164124546</td>
</tr>
<tr>
<td>7036171</td>
<td>7039824</td>
</tr>
<tr>
<td>7954180</td>
<td>7955128</td>
</tr>
<tr>
<td>143003673</td>
<td>143001781</td>
</tr>
<tr>
<td>22249167</td>
<td>22245141</td>
</tr>
<tr>
<td>62267220</td>
<td>62273955</td>
</tr>
<tr>
<td>171332654</td>
<td>171335508</td>
</tr>
</tbody>
</table>

**Annotation:**
- **CCT/B-box zinc finger protein**
- **AP2 domain containing protein**
- **IBR domain-containing protein**
- **Transducin/WD40 repeat-like superfamily protein**
- **DNA-binding protein transporter family protein**
- **leucine-rich repeat transmembrane protein kinase**
- **RNA recognition motif (RRM)-containing protein stomatal cytokinesis defective / SC1 protein**
- **beta galactosidase 1**
- **MYB domain protein 112**
- **heavy metal-associated domain containing protein lysine histidine transporter 1 (LHT1)**
- **LSM domain containing protein**
- **protein kinase superfamily protein**
- **cytokinin oxidase 5**
- **Unknown**
- **DNA-directed RNA polymerases subunit RPABC1**
Table 2-4 continued

<table>
<thead>
<tr>
<th>Allele</th>
<th>Position</th>
<th>Effect</th>
<th>p-value</th>
<th>MAF</th>
<th>Side</th>
<th>allele ID</th>
<th>Start</th>
<th>End</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>16263773</td>
<td>A/G</td>
<td>0.17</td>
<td>9.9E-09</td>
<td>0.33</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G301860</td>
<td>16253405</td>
</tr>
<tr>
<td>9</td>
<td>16317865</td>
<td>G/A</td>
<td>0.10</td>
<td>1.4E-15</td>
<td>0.59</td>
<td>Down_stream</td>
<td>Y</td>
<td>GRMZM2G099363</td>
<td>16320573</td>
</tr>
<tr>
<td>9</td>
<td>106051436</td>
<td>G/A</td>
<td>0.08</td>
<td>1.7E-05</td>
<td>0.45</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G107886</td>
<td>106201175</td>
</tr>
<tr>
<td>9</td>
<td>139547728</td>
<td>C/G</td>
<td>−0.11</td>
<td>3.4E-08</td>
<td>0.32</td>
<td>Intergenic</td>
<td>N</td>
<td>GRMZM2G402015</td>
<td>139591514</td>
</tr>
<tr>
<td>10</td>
<td>3090835</td>
<td>G/A</td>
<td>−0.13</td>
<td>6.2E-10</td>
<td>0.47</td>
<td>Intergenic</td>
<td>Y</td>
<td>GRMZM2G068465</td>
<td>3066458</td>
</tr>
<tr>
<td>10</td>
<td>64647379</td>
<td>G/A</td>
<td>0.06</td>
<td>5.7E-09</td>
<td>0.53</td>
<td>Y</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>134501142</td>
<td>C/T</td>
<td>0.06</td>
<td>4.4E-08</td>
<td>0.77</td>
<td>Intronic</td>
<td>Y</td>
<td>GRMZM2G063972</td>
<td>134495450</td>
</tr>
</tbody>
</table>

*a* Alleles reported as: “B73 allele/alternate allele”. “CNV-” represents the read depth of a line is significantly lower than B73.  
*b* The mean effect of each significant SNP across data subsamples.  
*c* SNPs located within model 7 QTL SI are indicated as “Y” and “N” otherwise.  
*d* Variants more than 500 bp away from an annotated gene.  
*e* Variation lies in the coding region and results in an amino acid change.
<table>
<thead>
<tr>
<th>Comparisons</th>
<th>GWAS models&lt;sup&gt;a&lt;/sup&gt;</th>
<th>RMIP ≥ 0.05</th>
<th>Proportion of overlapped SNPs (%)&lt;sup&gt;c&lt;/sup&gt;</th>
<th>RMIP ≥ 0.25</th>
<th>Proportion of overlapped SNPs (%)&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No. of all significant SNPs&lt;sup&gt;b&lt;/sup&gt;</td>
<td>10-&lt;i&gt;k&lt;/i&gt;b window</td>
<td>100-&lt;i&gt;k&lt;/i&gt;b window</td>
<td>200-&lt;i&gt;k&lt;/i&gt;b window</td>
</tr>
<tr>
<td>GWAS marker density</td>
<td>B vs. C</td>
<td>326 (151,175)</td>
<td>26</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>D vs. E</td>
<td>393 (201,192)</td>
<td>22</td>
<td>34</td>
<td>39</td>
</tr>
<tr>
<td>Genetic map density</td>
<td>C vs. E</td>
<td>367 (175,192)</td>
<td>14</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>B vs. D</td>
<td>352 (151,201)</td>
<td>25</td>
<td>29</td>
<td>35</td>
</tr>
<tr>
<td>GWAS and genetic marker</td>
<td>C vs. D</td>
<td>376 (175,201)</td>
<td>13</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>density and genetic marker</td>
<td>B vs. E</td>
<td>343 (151,192)</td>
<td>7</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>density and &lt;i&gt;P&lt;/i&gt;-value</td>
<td>A vs. E</td>
<td>437 (245,192)</td>
<td>6</td>
<td>12</td>
<td>15</td>
</tr>
</tbody>
</table>

<sup>a</sup> See Table 2-1 for details of each GWAS model.

<sup>b</sup> Total number of all the significant SNPs from a pair of analyses is shown outside of parentheses and numbers of significant SNPs from each analysis considered separately are inside the parenthesis “.”

<sup>c</sup> Proportion of overlapped SNPs was estimated as the total number of overlapped SNPs from a pair of analyses/total number of all the significant SNPs from the two analyses.

Comparisons show the input(s) that differed in each pair of GWAS comparisons. No. of significant GWAS markers denotes the total number of significant unique variants between the two analyses (with the number of significant variants within each analysis in parenthesis).
References


CHAPTER 3: Ensemble Learning of QTL Models Improves Prediction of Complex Traits

Citation

Ensemble Learning of QTL Models Improves Prediction of Complex Traits

Yang Bian\textsuperscript{1} and James B. Holland\textsuperscript{1,2,*}

\textsuperscript{1} Department of Crop Science, North Carolina State University, Raleigh, NC 27695
\textsuperscript{2} U.S. Department of Agriculture-Agricultural Research Service, Plant Science Research Unit, Raleigh, NC 27695

Keywords: Quantitative trait loci; thinning and aggregating; ensemble modelling; \textit{Zea mays}

*corresponding author (\texttt{james\_holland@ncsu.edu})
Abstract

Quantitative trait locus (QTL) models can provide useful insights into trait genetic architecture because of their straightforward interpretability, but are less useful for genetic prediction due to difficulty in including the effects of numerous small effect loci without overfitting. Tight linkage between markers introduces near collinearity among marker genotypes, complicating detection of QTL and estimation of QTL effects in linkage mapping, and this problem is exacerbated by very high density linkage maps. Here we developed a thinning and aggregating (TAGGING) method as a new ensemble learning approach to QTL mapping. TAGGING reduces collinearity problems by thinning dense linkage maps, maintains aspects of marker selection that characterize standard QTL mapping, and by ensembling, incorporates information from many more markers-trait associations than traditional QTL mapping. The objective of TAGGING was to improve prediction power compared to QTL mapping while also providing more specific insights into genetic architecture than genome-wide prediction models. TAGGING was compared to standard QTL mapping using cross validation of empirical data from the maize (Zea mays L.) nested association mapping population. TAGGING-assisted QTL mapping substantially improved prediction ability for both biparental and multi-family populations, by reducing both the variance and bias in prediction. Furthermore, an ensemble model combining predictions from TAGGING-assisted QTL and infinitesimal models improved prediction abilities over the component models, indicating some complementarity between model assumptions and suggesting that some trait genetic architectures involve a mixture of a few major QTL and polygenic effects.
Introduction

Massive numbers of molecular markers can now be readily provided by next generation sequencing and high through-put genotyping (DAVEY et al. 2011; ELSHIRE et al. 2011; BIAN et al. 2014a; GLAUBITZ et al. 2014). High density marker maps (with marker intervals smaller than 1 cM in mapping populations) have been available to an increasing number of plant species, such as in Arabidopsis thaliana (KOVER et al. 2009; HUANG et al. 2011), maize (SWARTS et al. 2014), rice (YU et al. 2011), and sorghum (ZOU et al. 2012). Unfortunately, the dramatic increase of marker availability may adversely affect QTL mapping as collinearity among markers may complicate detection and estimation of QTL. Statistical tests to declare a QTL or make QTL inference are conditional on other QTL in multi-QTL models (ZENG 1994). The value of a test statistic depends on other QTL in the model if they share some proportion of genetic variation. Strong covariance between tightly linked markers may increase the risk of selecting collinear markers, bias the QTL estimates and possibly overfit the predictive model, especially when a relaxed selection threshold is applied (BIAN et al. 2014b; OGUT et al. 2015).

Ensemble learning is an alternative approach that could improve QTL-based prediction ability. Ensemble learning involves estimating multiple “learner” models on a training data set and predicting unseen observations by a vote or weighted average among the multiple learners (DIETTERICH 2000). For example, suppose one has available a vector of some input variables, \( \mathbf{x} \), associated with a numerical outcome vector \( \mathbf{y} \). The estimation process involves finding a function of \( F(\mathbf{x}) \) that maps values in the space of \( \mathbf{x} \) to values in the corresponding space of \( \mathbf{y} \). \( F(\mathbf{x}) \) can be estimated in various ways, one of which is by
ensemble learning. The generic ensemble estimator takes the form: \( \hat{F}(x) = h_s(\{f_s(x)\}_{1}^{S}) \), where \( S \) is the number of ensemble members (or “base learners”), the base learners \( \{f_s\}_{1}^{S} \) are functions of \( x \) derived from the training data, and \( h_s(\cdot) \) is an ensemble learning function. As one example, the ensemble learner can be estimated simply by model averaging the base learners, \( \{f_s(x)\}_{1}^{S} \). The base learners can be trained based on random samples from the original data set or a subset of \( x \). Common ensemble methods include bootstrap aggregation of predictions (bagging) and random forests, and several features characterize them, such as loss function, ensemble dispersions, and memory of base learners (Hastie et al. 2009). Bagging of regression models is trained by paralleling base regressions with no influence (zero memory) among them and minimizing squared-error loss function on bootstrap samples (Breiman 1996a). If the bootstrap estimation is roughly correct, then aggregating would reduce variance without increasing bias. Random forests (Breiman 2001) increase ensemble dispersion over bagging by additionally using a randomly chosen subset of the predictors rather than using all of them, and this method typically solves for paralleled decision or regression trees (non-linear base learners). The ensembles of base learners produced by bagging and random forests can be conducted in a parallel manner, meaning that each individual learner is trained independent of the results from others.

Here, we developed a new ensemble learning approach used in QTL analyses, referred to as a thinning and aggregating (TAGGING) method. We thinned linkage map marker sets into sub-maps with equally reduced inter-marker density, built QTL mapping models upon the thinned marker sets as based learners in parallel, and aggregated by averaging the predictions from the base learners to predict the test data (Figure 3-1). The
TAGGING method shares some aspects of model averaging with bagging and random forest, as they all generate multiple predicted values and aggregate prediction by averaging. In contrast to random forests and bagging, however, TAGGING uses stratified sampling of the linkage map to create disjoint marker sets to generate independent discrete QTL models upon the unchanged set of observed mapping lines so that prediction variance is expected to decay faster and bias is not affected by bootstrapping samples. By comparison, bagging uses the original marker set to build prediction models on bootstrapped samples of lines, and random forests uses randomly sampled linkage markers and bootstrapped samples of lines.

QTL mapping has proven useful for detecting major QTL with relatively large effects, but may lack of power in accurately modeling small QTL effects or polygenic background effects (HEFFNER et al. 2009). QTL models typically underestimate the number and overestimate the effects of QTL, reducing the accuracy of QTL-based predictions (BEAVIS 1998; SCHÖN et al. 2004). Predictability and interpretability are two competing goals and usually compromise each other for predictive machines. The parameters in some genomic prediction (GP) models are not easily interpretable in terms of genetic theory, and this problem is compounded in non-parametric models (GIANOLA AND VAN KAAM 2008; CROSSA et al. 2010). Such models may serve as good prediction machines, but may provide no insights into the locations of important QTL or their gene actions. The TAGGING method presented here is a compromise method that aims to improve the prediction ability of QTL models, while maintaining some of their interpretability in terms of QTL locations and effect sizes. Furthermore, the ensembling concept is fully general, so predictions can be obtained by ensembles of models that capture distinct components of the true genetic architecture, for
example, ensembles of QTL models that are best at identifying genes with larger effects and polygenic models that capture the ‘background’ effects of many genes of small effect distributed throughout the genome. Models that can account for two hypotheses may better approximate the true biological mechanism better than either of the two models individually.

In this study, two QTL mapping models, joint-family (JF) linkage analysis for multi-family mapping populations and single-family (SF) QTL analysis for biparental mapping populations (Ogut et al. 2015), were tested within the TAGGING framework for maize complex traits. The new ensemble QTL models were examined for the number of QTL detected and proportion of genotypic variance explained by the detected QTL. We expect that for a high-bias base learner (e.g., QTL models based on sparse linkage maps), subset aggregation ensures model flexibility and therefore protects against high biasness in the ensemble predictions. For high-variance base learners (e.g., models in which QTL are included at low selection stringency with dense maps), ensembling might provide a reduction in variance of predictions.

The objectives of this study were to: (1) develop a thinning and aggregating (TAGGING) method as a new ensemble learning approach for QTL analysis, (2) compare prediction abilities of TAGGING, QTL models, and standard GP GBLUP models, and (3) test whether useful complementarity occurs between QTL-based and polygenic models in their ensemble learning for GP. We tested these models using a cross-validation strategy on data for three complex traits previously reported for the maize nested association mapping population (Buckler et al 2009; McMullen et al 2009). At all QTL selection stringency thresholds and thinning intensities we tested, the TAGGING strategy resulted in a better
prediction result than conventional QTL methods, implying the robustness of this new method. Results indicate that TAGGING provides information on the position and effect sizes of QTL (including their precision), while improving prediction ability compared to traditional QTL mapping procedures. With very high density linkage maps becoming increasingly available, we expect the TAGGING method will find use in genetic investigations and genetic prediction.
Materials and Methods

Plant phenotypes and genotypes

The maize nested association mapping (NAM) population comprises a set of ~5000 recombinant inbred lines (RILs) derived from crosses between a reference parent, inbred line B73, and 25 other diverse founder inbred lines of maize (McMullen et al. 2009). Three complex traits: resistance to southern leaf blight (SLB), plant height (PHT), and days to anthesis (DA) were previously studied for genetic architectures, and the predicted mean values for NAM RILs across multiple environments were previously reported (BUCKLER et al. 2009; KUMP et al. 2011; YANG et al. 2013; BIAN et al. 2014b; PEIFFER et al. 2014) (File B-1, B-2, B-3, B-4, B-5, and B-6). A second generation NAM linkage map consisting of 7386 imputed pseudo-markers with a uniform 0.2 cM inter-marker distance generated from genotyping-by-sequencing followed by full-sib family haplotype imputation (SWARTS et al. 2014) was used for linkage analysis.

Realized genomic relationship matrices (G matrices) were developed for each NAM family separately using the linkage map markers based on the first method described in (VANRADEN 2008), which track within-family identity-by-descent (IBD) for genome segments (Table 3-1). The linkage map marker scores do not reflect IBD between families, however, so a separate relationship matrix for the whole NAM panel developed from the maize HapMap v1 of 1.6 million SNPs (PEIFFER et al. 2014) was also used here (Table 3-1; File B-7). This G matrix measured the pairwise relationships among all lines with reference to the hypothetical population that would result from intermating all of the mapping population F1’s. To maintain identical numbers of markers for both the within-family
relationship matrices and the matrix for the whole panel, we extracted 7386 HapMap v2 markers polymorphic in NAM and closest to the 7386 linkage marker in terms of the AGP v2 coordinates (File B-8, B-9, and B-10) to construct an identity-in-state (IIS) $G$ matrix for the whole NAM panel. A total of 4354, 4359 and 4359 RILs with both genotypic and phenotypic data were available for analyzing SLB, PHT and DA, respectively.

Prediction accuracy calculation

Prediction abilities of models were evaluated both for the whole NAM panel and for within each of the 25 biparental mapping population using a cross-validation procedure. Training sets were created by randomly sampling 80% of RILs from each of the 25 NAM families. The remaining lines constituted the validation sample for the single family and joint family QTL models created using the training data set. Within-family predictions were evaluated, and prediction abilities were measured as the proportion of total trait variance explained by the model ($R^2$) in the validation set. The $R^2$ was used to enable direct comparison to heritabilities and was estimated by averaging squared Pearson’s correlations ($R^2$) between the observed and predicted BLUP values. We converted the value of $R^2$ to have a negative sign for a few cases of negative correlations between predicted and observed line values.
**QTL Models**

Multiple linear regression (MLR) models were first fit within each family independently (single-family “SF” models; **Table 3-1**). For a given linkage maps $s$, the SF model for family $f$ is:

$$ Y^f = 1 \mu^f_s + X^f_s \beta^f_s + \epsilon^f_s, $$

where $Y^f$ is a vector of $N^f$ length referring to the trait phenotypic values in a given family, 1 is an $N^f \times 1$ intercept vector of 1’s, $\mu^f_s$ is the intercept, $X^f_s$ is a $N^f \times K^f_s$ matrix of marker genotypes, $\beta^f_s$ is a $K^f_s \times 1$ column vector of the additive effects relative to B73, $K^f_s$ is the number of significant loci in stepwise selection, and $\epsilon^f_s$ is a $N^f \times 1$ column vector of errors. The stepwise incorporation and exclusion of markers were based on pre-defined alpha threshold ($p$). After including as many markers as possible based on their $p$-values, the model was then reduced by split-sample cross-validation. Specifically, the model selection step with minimum prediction error variance was chosen in five-fold split-sample cross-validation within the training set, using the ‘choose = CV cvMethod = split(5)’ option in SAS Proc GLMSSelect (INSTITUTE 2013). The RIL phenotype values were predicted from the linear model.

Joint family linkage (JF) models were trained in all 25 families of NAM populations. For a given linkage maps $s$, JF analysis for the multi-family connected populations can be described as:
\[ Y = A\mu_s + \sum_{i=1}^{K_s} X_{si}\beta_{si} + \varepsilon_s , \]

where \( Y \) is a vector of \( N \) length referring to the trait phenotypic values for all RILs, \( A \) is an \( N \times P \) incidence matrix relating RILs to their corresponding family, \( \mu_s \) is a \( P \times 1 \) column vector for the family effects, \( X_{si} \) is a \( N \times P \) matrix relating each RIL’s genotype at locus \( i \) to its corresponding family-specific allele effect, \( \beta_{si} \) is a \( P \times 1 \) column vector of the family-specific additive effects associated with locus \( i \), \( K_s \) is the number of significant loci in stepwise selection based on pre-defined \( p \), \( \varepsilon_s \) is a \( N \times 1 \) column vector of errors. The prediction ability was then evaluated for each family separately to make a comparable comparison with biparental scenarios.

*Thinining and Aggregating method for QTL Analyses*

The TAGGING QTL analysis method can be summarized as thinning of dense marker maps into a set of disjoint maps of lower density, and then aggregating predictions from paralleled QTL models on the same training data sets (Figure 3-1). The method begins by conducting a stratified sampling of the markers on the linkage map, which is simplified in the case of the maize NAM linkage map which has a uniform density of 0.2 cM between each pair of adjacent linked markers. The original map is thinned into \( s \) disjoint sets of markers, maintaining the linkage map position information for the markers, starting with the first marker on the first linkage group. This procedure was then repeated by starting selection at the 2\(^{nd} \) marker to create a new sample map, and continuing to initiate selections at subsequent markers until the \( s^{th} \) marker. The result is \( s \) disjoint maps, each with \( s \times 0.2 \) cM distance between adjacent markers. Thinning is expected to reduce the extent of covariance
and collinearity that otherwise occur in base QTL analyses with the dense maps and/or relaxed selection stringencies.

JF and SF models for multi-family and biparental mapping populations were fit using each thinned map separately, and phenotype values for the validation set lines were predicted separately for each reduced map (File B-11 and B-12). The two types of QTL base learners were constructed with each subset of markers for the same training samples to estimate the regression coefficients, and then the $s$ sets of SF and JF predicted values for the test sets were aggregated to form the ensemble prediction (Figure 3-1). The ensembles of QTL models proposed here were named “ensemble joint family linkage” (EJF) analyses for multi-family mapping populations and “ensemble single family” (ESF) analyses for biparental mapping populations, respectively. The ensemble learner $\hat{F}(x)$ is formed as some linear combination of predictions of each base learner: $\hat{F}(x) = a_0 + \sum_{s=1}^{s} a_s f_s(x)$, with $\{a_s\}_{s=1}^{s}$ being the coefficients for $\{f_s(x)\}_{1}^{S}$. For both EJF and ESF modeling, the coefficients were set to $a_0 = 0$ and $\{a_s = 1/s\}_{1}^{S}$. Essentially, we averaged predictions over $s$ map subsets for each training set, and the ensemble predictions were an arithmetic mean of those base learners.

To estimate precision of QTL localization, the frequencies of QTL positions detected in JF and EJF across the resampled training sets were summarized to elucidate the QTL architectures for the three complex traits. Resample Model Inclusion Probability (RMIP) (Valdar et al. 2009) was computed to measure the power of detection for the trait-marker associations across NAM panel. The RMIP was calculated for each marker as the proportion of data samples in which the marker was tested and selected in the model of interest at the given selection $p$. 
To compare the efficiency of the TAGGING ensemble learning method with existing ones, we implemented subsample aggregating (subagging) of both JF and SF analyses using the same base learning algorithms as in TAGGING. Subagging is a sobriquet for subsample aggregating where simple random sampling (SRS) is used instead of aggregation of bootstrapped samples implemented by bagging. The basic difference between TAGGING and subagging is that TAGGING averages over subsamples of markers on a fixed set of RILs, while subagging (subsample aggregating) averages over samples of RILs for a fixed set of markers. Subagging instead of bagging was used here because subbagging is more efficient in computation, due to reduction in sample size compared to bagging, and it avoids collinearity problems induced or aggravated from the increased relatedness among bootstrapped samples, which is a concern with the dense linkage map used here. We formed ten SRS subsamples with sample size equal to 80% of each family from each original training set. We implemented subbagging of both SF and JF analyses upon the same training and test sets as used by all other analyses.

SF and ESF models were constructed using selection thresholds ranging from $p = 1e^{-4}$ to $p = 0.05$, and $p = 1e^{-5}$ to 0.01 for JF and EJF. Map resolutions used ranged from 0.2 to 20 cM inter-marker distances for both. Since $p = 1e^{-05}$ was too stringent for SF, and $p = 0.01$ or greater were too relaxed for JF using 0.2-cM density map (OGUT et al. 2015), we did not include them in this study.
Prediction error decomposition

The cross-validation experiments allow us to evaluate the mean squared prediction error (MSPE) and to decompose this into terms due to the bias and variance in prediction. Suppose there is some underlying true function \( f(x) \) estimated from ensemble or discrete QTL models on the training set \( T \) by \( h(x, T) \). Given a new data point in test set \( x^* \), \( y^* = f(x^*) + \varepsilon \) where \( \varepsilon \) is normally distributed with zero mean and variance \( \sigma^2 \), a function of 1-heritability. The expected pointwise prediction error can be decomposed in a familiar form:

\[
E_T[(y^* - h(x, T))^2] = E_T[(y^* - E_T[h(x, T)])^2] + E_T[(h(x, T) - E_T[h(x, T)])^2],
\]

where \( E_T \) denotes expectation over resampled training sets drawn from the same distribution to estimate \( y^* \). While the second part of the right hand side of (1) is prediction variance, the first part of the right hand side of (1) is the sum of bias\(^2\) and irreducible error variance and decomposed as:

\[
E_T[(y^* - E_T[h(x, T)])^2] = E_T[(f(x^*) - E_T[h(x, T)])^2] + E_T[(y^* - f(x^*))^2] = E_T[(f(x^*) - E_T[h(x, T)])^2] + \sigma^2,
\]

The TAGGING method estimated \( h(x, T) \) by averaging over \( s \) predictions from \( s \) sets of thinned marker maps based on a training set \( T \) used, and estimated \( E_T[h(x, T)] \) by averaging over all test sets that contain the data point in question. Finally, averaging over all points predicted throughout all test sets for their bias and variance gives the more accurate estimates. We refer to \( E_T[(y^* - E_T[h(x, T)])^2] \) in (1) as “bias\(^2\)” in our results, since the true underlying genetic mechanism is unknown and thus the irreducible error variance could not be disentangled from the bias\(^2\) in (2). The bias\(^2\) and variance in prediction were compared.
between single and TAGGING-relied models all of which used the same 7386 linkage marker genotypes. In addition, a few predicted values representing the most severe outliers in prediction ability were excluded from prediction error calculation in order to guard against artifactual inflation of the prediction error variance simply due to high collinearity that occasionally occurred in few cross-validations. The cutoff for excluding an outlier was the mean predicted value ± 100 times the standard error of the mean.

**GBLUP models using IIS and IBD genotypes**

The $G$ matrix derived from 1.6 million SNPs in HapMap v1 was used to create a GBLUP model for whole NAM families (HGBLUP models, Table 3-1) in the same cross validation setup as in QTL analyses. Single family GBLUP models (SGBLUP model, Table 3-1) were developed for each NAM biparental mapping family using the 25 disjoint $G$ matrices developed from the original linkage map of 7386 markers. We also extracted the same number of allele calls as close as possible to the physical positions of the 7386 linkage map pseudo-markers, and with the IIS $G$ matrix across whole NAM panel derived from that, we conducted the joint GBLUP models (JGBLUP model, Table 3-1) to compare the efficiencies of using the same amount of genotypic information in different models.

**Ensemble learning of TAGGING-assisted QTL models and GP models**

To capture the complementary strengths of single- and joint-family analyses, we averaged predictions from EJF and ESF, and referred to this ensemble as the EJF+ESF model. All pairs of QTL and GP models using the same number of 7386 loci were then
combined in ensemble learning in a linear fashion. We explored ensemble learning by averaging predictions from either EJF, ESF or EJF+ESF QTL models with predictions from SGBLUP or JGBLUP models in order to test the hypothesis that mixture of models with contrasting genetic assumptions can improve genetic prediction by better mimicking genetic architectures (File B-13). For simplicity, the coefficients in ensembles of two component models were set to 0.5. Finally, we attempted ensemble learning that combined QTL and GBLUP models derived from both 7386 marker genotypes and 1.6 M HapMap1 SNP genotypes. We evaluated the importance of the four participating models to their ensemble prediction $R^2$, EJF+ESF, SGBLUP, JGBLUP and HGBLUP, by analysis of variance in a $2^4 - 1$ factorial experiment, in which 15 possible combinations of the four models’ presence or absence were included.

Pseudo optimal ensemble coefficients

We tested whether additional tweaking of the coefficients of the base leaners in the ensemble model can improve the ensemble prediction. The base learners can be considered as input variables to a multiple regression model aimed at predicting the unknown trait values. Here we introduced pseudo optimal coefficients for giving relative weights to each component model, which we estimated from using cross-validation with respect to the true phenotypic data of test sets to identify a fixed set of ensemble coefficients to minimize the least-square errors for the test data. The ensemble coefficients were tuned using the Nelder–Mead optimization technique (NELDER AND MEAD 1965), in which average prediction $R^2$ values were dynamically recorded for tested sets until they (as objective functions)
converged to a plateau. The sum of coefficients was scaled to one, under non-negativity constraints, as recommended by (Breiman 1996b). The pseudo optimal accuracies should be considered as an idealistic upper limit and not likely to occur in a real prediction scenario, as those predictions were obtained using the phenotypic values in test sets to optimize the ensemble coefficients. Finally, two sets of natural coefficients were attempted in ensembles of the above three component models, EJF, ESF, and SGBLUP: equal weights of 1/3 for each, and equal weights between QTL-based models (0.25 for EJF and ESF) and SGBLUP models (0.5).
Results

TAGGING of QTL models improves QTL model prediction

EJF models had substantially better prediction R2 compared to the individual JF models using either the original or reduced maps for the three traits (Figure 2, B-1, and B-2, Table B-1, B-2 and B-3). Using the thinned maps of 10 ~ 15 cM inter-marker distances at selection p = 0.01 generated the best prediction R2 in EJF for SLB (R2 = 0.475), PHT (0.476) and DA (0.445); those best EJF models outperformed the best JF model substantially [1 cM maps under p = 0.0001 for SLB (R2 = 0.385) and PHT (0.371), and 1 cM maps under p = 0.001 for DA (0.332)]. The improved prediction abilities of EJF over JF indicated that the genetic architecture can be better depicted by averaging multiple independent QTL models developed from disjoint subset maps as achieved by TAGGING. TAGGING of SF analyses also improved the predictions for the biparental mapping populations substantially compared to the regular SF analyses. The prediction R2 values generated by the optimal ESF models outperformed those from the best SF models by an average of 0.118, 0.123, and 0.116 across 25 families for SLB, PHT and DA, respectively (Figure 3-2, B-1 and B-2; Table B-1, B-2 and B-3). Among the pure linkage models (SF, JF, ESF, and EJF) and across all traits, map densities, and selection p thresholds tested, EJF models had the highest prediction abilities, whereas SF models always had lowest prediction ability (Table B-1, B-2 and B-3).

The performances of both discrete and TAGGING QTL models varied with the densities of maps and selection p. In general, for EJF, sparse maps and relaxed p were favorable in prediction, and for JF, dense maps and stringent p were favorable. Prediction
abilities from EJF increased with decreasing p thresholds and decreasing map densities until the map density reduced down to 15 to 20 cM inter-marker distances (Figure 3-2, B-1 and B-2). In contrast, for biparental mapping populations, moderate map densities along with relaxed p were advantageous in prediction using both SF and ESF analyses. In addition, using linkage maps of 1 cM or 0.2 cM density made little difference in prediction abilities for both JF and SF QTL models (Figure 3-2, B-1 and B-2).

The efficiency of TAGGING was compared to the previously proposed ensemble learning method subbagging, both of which were applied in the same inputs and cross validation schemes. The best prediction R2 values were 0.452, 0.440, and 0.405 using subbagging of JF models under selection p = 0.001, and 0.396, 0.378 and 0.391 using subbagging of SF models under selection p = 0.01 for SLB, PHT and DA, respectively (Table B-1, B-2 and B-3). In all cases, the optimal prediction abilities of TAGGING-assisted QTL models were superior to the corresponding subbagging predictions (Table B-1, B-2 and B-3).

Bias and variance

For all traits and under all selection p thresholds, TAGGING substantially reduced prediction variance in both JF and SF models (Figure 3-3 and B-3). Variance in prediction was basically eliminated when a large (≥ 10) number of thinned maps were averaged in TAGGING. TAGGING of JF models always reduced the bias compared to the single JF models, and the magnitude of bias in EJF models was roughly equal across a large range of thinning intensities, even when the map density was as low as 100 markers per map. The
prediction bias in TAGGING of SF analyses can fluctuate compared to that of the individual SF models, depending on the thinned map densities (Figure 3-3 and B-3). When SF models suffered from severe collinearity at the relaxed selection stringency ($p = 0.01$), thinning reduced covariance among markers and alleviated model bias. When SF models were based on high stringency of marker selection ($p = 1e-04$), aggregating of the appropriately thinned maps (2 ~ 5 cM inter-marker distances) seemed to slightly improve model fit by incorporating more predictors in the ensemble. In general, TAGGING seemed to be more protective against high bias for JF than SF models (Figure 3-3 and B-3). By examining the error compositions, we found that the prediction advantages of TAGGING of JF models over that of SF models were related to the lower level of bias before using TAGGING and more stable reduction of bias after using TAGGING (Figure 3-3 and B-3). In addition, when comparing across $p$-value thresholds, prediction bias in both TAGGING models decreased with more relaxed $p$-values because of higher model flexibility.

Genetic architecture revealed by marker-trait associations

To better understand the trait QTL architectures, the probability of inclusion of a marker in a JF or EJF models was estimated across the resampled training sets (resample model inclusion probability, RMIP; Valdar et al. 2009). RMIP plots visualized the enrichment of marker-trait associations within particular genomic regions (Figure 3-4, B-4, and B-5). In general, with the decrease in the thinned map densities used in EJF (which also reflects the increase in the number of JF models combined by TAGGING), RMIP values increased substantially and the regions containing marker associations expanded (Figure 3-4,
B-4, and B-5). As the selection $p$ relaxed, the RMIP values in EJF increased and the enriched regions expanded, especially for sparse maps (Figure B-6).

**GBLUP model prediction performance**

The same marker data used for SF and ESF models were also used to construct within-family realized genomic relationship matrices. These relationship matrices were used to implement SF GP models (SGBLUP model) (Table 3-1). Average within-family prediction $R^2$ values were 0.460, 0.465, and 0.450 for SLB, PHT and DA, respectively, using the SGBLUP model. The JGBLUP model, which used an integrated relationship matrix calculated based on IIS information of the same number of marker genotypes that were most adjacent the linkage marker physical positions, generated best prediction results among EJF, ESF, SGBLUP and JGBLUP models (Table B-4). Similarly, a previous report showed that joint analyses of several half-sib families in GP models can increase prediction abilities over family-specific GP models (LEHERMEIER et al. 2014). In addition, the NAM panel was also analyzed using HGBLUP model for which the relationship matrix was calculated based on the 1.6 M HapMap v1 SNPs. HGBLUP generated the best prediction abilities of all component models (Table B-4) probably because of the enormous amount of genotype information used.
The SF and JF QTL models were based on different genetic assumptions of genetic heterogeneity and allele effects, and may be considered complementary in describing QTL architectures (OGUT et al. 2015). To test the hypothesis that the combined QTL mapping results can improve prediction, we ensembled by model averaging to combine results from EJF and ESF. The EJF+ESF models did not noticeably improve prediction abilities over the EJF models (Figure 3-2, B-1 and B-2). The EJF+ESF results indicated that the selection \( p = 0.001 \) and 0.01 in EJF and ESF analyses resulted in the optimal prediction \( R^2 \) when 1 cM reduced maps were employed, and \( p = 0.01 \) and 0.05 for 20 cM reduced maps (Figure 3-2, B-1 and B-2, Table B-1, B-2, and B-3). We then restricted our modeling to using those \( p \) thresholds for the most thinned 20-cM maps in further ensemble learning.

For ensemble learning between TAGGING-assisted QTL and GBLUP models, first, we attempted pairwise combinations of the three QTL models and two GBLUP models that used the same underlying 7386 marker genotypic information. Adding EJF or EJF+ESF models substantially increased within family prediction \( R^2 \) and generated better prediction results for at least 21 (\( p \)-value < 0.0005) out of 25 NAM families, compared to their ensemble partner SGBLUP or JGBLUP models (Figure 3-5 and B-7, Table B-4). Second, the best ensemble combinations of QTL and GBLUP models using the same limited genotypic information improved prediction abilities, resulting in comparable performance relative to GBLUP models using much larger scale genotypic information. Specifically, (EJF+ESF)+JGBLUP models outperformed HGBLUP models (based on > 200 times more
marker information) by 0.01 ~ 0.02 in within-family prediction $R^2$ for SLB and PHT (Figure 3-5 and B-7, Table B-4).

The high consistency of model prediction abilities between individual component models and their ensemble models suggested that the ensemble learning performance is likely predictable. Indeed, linear models involving terms of HGBLUP and JGBLUP or (EJF+ESF)*JGBLUP explained ~72 % to ~84% (Table B-5) of the variance in within family prediction $R^2$ among all combinations of the four tested component models (EJF+ESF, SGBLUP, JGBLUP, and HGBLUP). The (EJF+ESF)+HGBLUP+JGBLUP model (based on equal weights of 1/3 for the three components) resulted in the best prediction $R^2$ (0.516 for SLB, 0.515 for PHT and 0.512 for DA), which was 0.02 better for SLB and PHT, and marginally better for DA than the HGBLUP model; it also predicted better in significantly more families than the HGBLUP model (Table B-4). The magnitude of the increase in prediction $R^2$ from ensembling TAGGING-assisted QTL models and GBLUP models was small but consistent across families. The traits studied are highly multigenic, so the results may be considered as a conservative case for evaluating the utility of the additional QTL information in GP models.

**Pseudo optimal ensemble coefficients**

We further searched for a fixed set of optimal coefficients that can maximize the accuracy from the EJF, ESF and SGBLUP predictions. The prediction accuracy reached plateau after several rounds of Nelder–Mead optimizations. The resulting average within
family prediction $R^2$ based on the “pseudo optimal” ensemble coefficients differed by less than 0.01 from the two sets of simple coefficients $(1/4)\text{EJF}+(1/4)\text{ESF}+(1/2)\text{GBLUP}$ and $(1/3)\text{EJF}+(1/3)\text{ESF}+(1/3)\text{GBLUP}$ we developed (Table B-6). The results implied that the current ensemble learning to combine QTL-based and GP models was efficient in capitalizing the model complementarities.
**Discussion**

*Optimizing QTL mapping parameters*

Determining the optimal prediction model for a collection of multiple biparental families, as is commonly encountered in both academic genetic studies and commercial breeding programs, is important for maximizing precision of QTL mapping and accuracy of genome-enabled prediction. Our results demonstrate that there is an interaction between map density and QTL significance threshold on model prediction performance. The optimal threshold for a given analysis can vary according to map density. In earlier studies of prediction accuracy within biparental families from QTL models, accuracy was optimal at quite relaxed $p$-value thresholds (HOSPITAL et al. 1997; BERNARDO 2004). Those studies were conducted using linkage maps with substantially lower marker density than the current study. As genomics technology has made increasingly dense linkage maps available, however, marker collinearity becomes problematic for QTL mapping and the relaxed threshold is no longer optimal. For example, when $p = 0.05$ were adopted in SF analysis at the highest map density, average prediction ability was 0.21, and this improved to 0.31 by increasing the selection stringency to $p = 0.01$ (the first column of Figure 3-2). Using appropriately chosen selection $p$ reduced the propensity to overfitting or underfitting resulting from the improper marker densities. Nevertheless, an overly sparse map used in JF or SF models may fail to capture the QTL information or compromise the precision of QTL positioning because the markers are not in reasonable linkage to many QTL, and therefore lead to omitted-variable bias (Figure B-8). Our JF and SF model results suggest that 1 cM
density is appropriate for the level of LD that exists within a biparental RIL family in maize. By comparison, previous research reported that marker densities increasing from 5 cM to 1 cM in a doubled haploid maize population improved neither the overall QTL detection power nor the proportion of genotypic variance explained by the detected QTL (STANGE et al. 2013). Finally, our results confirmed that the optimal selection threshold also differs by QTL mapping design. SF models are more prone to underfitting (high bias) problems than JF models, and more stringent selection threshold should be used in multiparental analysis (BLANC et al. 2008). To conclude, it is important to consider the dynamics among map density, QTL analysis method and QTL detection stringency, the optimization is important to gain the best performance in QTL analysis based prediction (Figure 3-2, B-1 and B-2). Performances of both traditional and TAGGING-assisted QTL models varied with the densities of maps and selection $p$. However, the prediction ability of TAGGING models was greater than the respective traditional QTL mapping methods for all selection $p$ and thinning intensities tested.

Bayesian genomic prediction models represent an important class of models to consider, but the computational difficulty of conducting on many repeated samples with the millions of markers used in this study (for the HGBLUP models) precluded their inclusion in the model comparisons in this study. In addition, Bayesian models have shown little or no advantage over GBLUP models in many plant breeding scenarios (HESLOT et al. 2012; GUO et al. 2012).
Marker-trait association and genetic architecture

The EJF models may reflect the true genetic architecture better than discrete models concerning only a few strong associations. Indeed, the greater number of marker associations contributing to ensemble predictions is one of the bases for the improved prediction ability. By using the TAGGING method, association architectures in the EJF models were equipped with the polygenic feature that can include many more markers linked to QTL, while it also involved marker selection and differential weighting of marker information. The precision of QTL mapping, however, can be compromised in ensemble mapping. For example, previous data had indicated with high confidence the presence of two linked QTL on chromosome 3 for genetic resistance to SLB (Bian et al. 2014b), but the blurring of QTL positions in EJF resulted in a merging of the two QTL into a single broad peak in the RMIP visualizations (Figure 3-4). The larger number of marker effects that are ensembled in the TAGGING method may offer a useful compromise between QTL detection and predictions. We focused on cross-validation comparisons based on real data here, but simulation studies would be required to determine the accuracy of QTL position and effect estimates from the TAGGING method.

Ogut et al. (2015) suggested that JF models (which assume common QTL positions among families) and SF models (which assume independent QTL positions among families) can complement one another by capturing different aspects of the overall genetic architecture. We tested this hypothesis by ensembling predictions from the two QTL models in a single EJF+ESF model. The ESF+EJF model appeared to offer little advantage over EJF: the relatively poor predictive ability of ESF negated any advantage of complementarity
between ESF and EJF models when they were ensembled. A single integrated base learner model that more flexibly fits allele effects only within families where they are significant may more effectively achieve the goal of taking advantages of both models in depiction of genetic architectures. Development of such a model is underway.

Marker-trait associations are found highly enriched in some genomic regions, as indicated by the disjoint TAGGING scans. For a particular region of interest \textit{a priori}, the RMIP information surrounding that region indicates its importance and the resolution of QTL information. The association-enriched regions might represent probable intervals of QTL effects, that is, QTL or the cluster of QTL may reside in the multi-locus regions or in linkage with the loci in the regions. Leveraging association-enriched regions may better explain the underlying genetic architectures compared to the traditional point estimates of single QTL peaks, although further research remains needed to effectively define and set boundaries for the regions or factor in their kernel density. We suspect that the association-enriched regions might be important to reveal hidden genetic variation. Further research will be required to develop this method and test the model efficiency.

\textit{Bias and variance in TAGGING prediction}

The prediction ability of ensemble learning is usually stronger than that of a single learner. The first reason is that the training sample size might not be sufficient for determining a single best base model. In TAGGING, the reduced maps provide similar information as they only differ by consecutive markers. The base learners should perform
similarly well on searching those slightly different hypothesis spaces to fit the same training data sets. Thus, combining these learners can join the marker information that would otherwise not be obtained by searching in the original hypothesis space with a single algorithm (ROKACH 2010). The second reason is that, the search processes of the learning algorithms might be imperfect, especially in the “small N, large P” situations that have become common in genomics. Even if there exists a unique best set of predictors (actual functions of genes underlying the studied traits), the high dimensionality of dense linkage maps may hinder this set from being selected by efficient search algorithms. Thus, ensembles can compensate for such imperfect search processes by reducing the hypothesis space. The third reason is that, the model specification (linear models here) being tested might not contain the true target function, but ensembles can nevertheless provide better approximations to the true function than a single base learner function can.

The proposed TAGGING framework was successful at strengthening “weak learners” (two types of QTL models) by first reducing the hypothesis space and then aggregating by averaging the base models. A rule of thumb for optimizing TAGGING in QTL-based prediction is to conduct EJF based on heavily thinned maps (more sets to aggregate) under relaxed selection $p$ thresholds. Thinning alleviates the collinearity within each marker set, and this allows the selection threshold to be relaxed without overfitting. In addition to thinning, aggregating across predictions from multiple models also decreases the prediction variance. For example, averaging $N$ identical independent model predictions would reduce the resulting prediction variance by a factor of $1/N$. In most ensemble learning, including TAGGING, reduction is obviously less than $1/N$ because of dependent base predictions.
Thinning by stratified sampling of markers from the linkage map takes advantage of the consistent ‘spatial’ pattern of correlation among markers, such that the subset hypothesis spaces defined by thinning represented disjoint representations of the linkage map. By decomposing the prediction error variances, we showed that the contribution of variance among prediction sets was reduced to almost zero with large samples of sparser subset maps (Figure 3-3). Finally, the TAGGING strategy tracks the linkage structure exactly in our case of a perfectly uniform marker density. In more typical QTL mapping situations in which the markers are not evenly spaced in the linkage map, it is important to stratify maps accounting for the original marker correlations, instead of just even sampling across chromosomes.

Similarly, in genome-wide association studies for which the correlation structure of marker set is always not constant, the LD between marker genotypes does not decays monotonically as their physical distances extend. Special attention needs to be paid to appropriate marker stratification schemes for these more general situations of uneven marker spacing for QTL mapping and complex LD structure for association mapping when TAGGING.

The minimum density required for TAGGING to work was around 20 cM in the thinned maps, and further thinning caused decreased prediction ability in TAGGING. Since dense maps are expected to be more easily available for many species, the applicability of TAGGING will only increase over time. Furthermore, most crop plants regularly deal with maps with density greater than one marker per 20 cM, so there is already general applicability at this time.

In addition to reducing the variance by model averaging, another motivation of TAGGING was to alleviate bias in prediction. First, a large collection of disjoint predictors
can be considered and weighted when ensembling them, and therefore more genotypic information (Figure 3-4, B-4 and B-5) contributes to the ensemble prediction, resulting in reduced bias in prediction. Moreover, when a few dominating predictors consistently perform better than their correlated competitors (for example, markers within the same LD block), they will tend to be selected in prediction models at the expense of those weaker competing predictors, some of which may provide information about local features of the data. TAGGING thins the map into equally spaced disjoint maps, providing more opportunity for predictors to be considered without competition from the dominant predictors, possibly increasing the chance that weak local features will be represented by some of the base learners.

We expect that bootstrapping of data samples in addition to TAGGING to generate even greater numbers of base learners will not result in a substantial decrease in the prediction error, as the variance is already nearly eliminated by TAGGING, whereas bias may increase due to the use of smaller effective training samples for each base learner. As observed here (Figure 3-3 and B-3) and found in many other ensemble studies, the dominating error source turns out to be the bias² and irreducible error (BAUER AND KOHAVI 1999), which suggests necessity of bias reduction in ensemble learning. A better approach may be to implement TAGGING upon higher variable (lower bias) base learners. Moreover, bias corrected estimators (EFRON 1987) can be further incorporated into TAGGING, which may help reduce the prediction bias due to finite-sample bias in base learner estimators (HOROWITZ 2001; ZHANG AND LU 2012). Similar to bagging and random forest, TAGGING applied a natural model averaging weight to combine base learners and did not require a
tuning process. Another direction of future work could be related to exploring sophisticated
ensemble learning algorithms. Regularized linear regression on the base learners can be
easily implemented (FRIEDMAN AND POPESCU 2008). Alternatively, meta-learning strategies
such as stack regressions (BREIMAN 1996b) are approaches based upon the parallel training
of multiple learning programs, followed by a meta-learning stage to stack them in a
principled fashion.

The prediction ability of TAGGING models is not sufficient to outperform current
standard genomic prediction methods, such as GBLUP. Our results suggest, however, that
TAGGING can simultaneously match the prediction ability of the GBLUP model (with a bit
of complementarity that can be exploited in additional ensembling) while also providing
information on important genomic regions, which can be utilized in gene discovery. The
improved prediction ability of the ensemble models over conventional QTL mapping imply
they can better model the true QTL architecture, for example, by highlighting important
genomic regions instead of relying on point estimates of QTL effects and by ameliorating
collinearity in dense genetic maps.

*Oligogenic and polygenic model complementation*

The high heritabilities (over 85%, line mean based) of the three traits implied a great
proportion of trait variation should be attributable to differences between maize lines after
accounting for known environmental effects (Buckler et al. 2009; Kump et al. 2011; Peiffer
et al. 2014). Traditional additive QTL models nevertheless provide only moderate prediction
ability. This indicates that the genetic factors underlying these high heritable traits are complex or might not be approximated well without considering more complex model hypothesis. No strong specific digenic interaction was found in NAM populations for the studied traits, although it is still possible that polygenic additive by additive effect is important, even if we have not mapped specific interactions so far. Considering the epistasis may be one piece of the missing components, fitting feasible epistasis effects that can account for moderate or large effects seems practical in improving prediction accuracy, especially in our TAGGING framework where the hypothesis spaces can now be more easily searched. Previous studies showed that infinitesimal GP model (GBLUP or ridge regression models) outperformed QTL-based model in predicting complex traits for both multi-family populations (Peiffer et al. 2014) and biparental segregating populations (Lorenzana and Bernardo 2009; Guo et al. 2012). The opposite results were found for traits with major QTL (Zhao et al. 2014). In a plant breeding application, linear combinations of different genomic prediction (including Bayesian) models did not result in noticeable gain in GP accuracy (Heslot et al. 2012). Our results showed that the ensemble learners among well-tuned TAGGING-assisted QTL models and GBLUP models that come from the same genotypic information improved prediction, which suggests their useful complementation in prediction of complex traits. Furthermore, the factorial experiment of combining varied model predictions suggest that aggregating models that use different genotypic information is advantageous in GP in the tested NAM populations. Leveraging the complementary effects among model assumptions and/or genetic information provides one more possible solution to achieve a better model specification in order to approach the ideal heritable variation.
Acknowledgment

We thank Dr. Jason Peiffer for suggesting use of the Nelder–Mead optimization technique and for providing us with HapMap v1 based relationship matrix for NAM panel. We thank Dr. Howard Bondell for helpful discussion about analyses on bias and variance. We used the High Performance Computing Services in NC State University, and thank Dr. Gary Howell for assistance with parallel computing. We appreciate two anonymous reviewers for their thorough evaluations. Research was supported by United States National Science Foundation grants IOS-1238014 and IOS-0820619 and the US Department of Agriculture, Agricultural Research Service. YB is a graduate student at NC State University and supported by NSF grant IOS-1238014 and a Monsanto Co. fellowship.
Figures

**Figure 3-1** Data sampling, map thinning, model building and ensemble schemes. The cross validations were repeated 10 times for **Figure 3-2, 3-3 and 3-4**, and 50 times for **Figure 3-5**. Step 1: Take a stratified random sample 80% of RILs of each NAM families to use as a training set and 20% as test set. Step 2: Thin the original 0.2-cM resolution map is into \( s \) sets of reduced maps with inter-marker distance \( 0.2 \times s \) cM, and calibrate JF and SF models as base learners on the same training sets with various combinations of reduced maps and \( p \) thresholds. Step 3: Independently predict the same test sets with the parameter estimates obtaining from training data. Step 4: Generate ensemble predictions for ESF and EJF models by taking arithmetic means. Step 5: Generate ensemble learning and prediction of QTL-based and GP models. Step 6: Evaluate prediction \( R^2 \) for all models including individual QTL models, TAGGING-assisted QTL models, ensembles of QTL and GBLUP models, as well as subbagging models (see text for details).
Figure 3-2 Prediction $R^2$ for resistance to SLB in biparental and multiple-family prediction, comparing JF, SF and the TAGGING-assisted QTL analyses using multiple map densities and under multiple selection $p$. The number at the top of and dot within each boxplot presents the mean $R^2$ among 25 NAM families for that boxplot. The x-axis represents the densities of linkage maps used for the JF, SF, and TAGGING-assisted QTL methods.
Figure 3-3 Prediction bias\(^2\) and variance in TAGGING-assisted and single QTL models for resistance to SLB. X-axes denote the inter-marker distances for the genetic maps used: “0.2” for single JF or SF models, and others for TAGGING methods. Dot radius was scaled to within-family \(R^2\) calculated based on all RILs in the test sets to permit comparisons of both mean error measures and mean prediction \(R^2\).
Figure 3-4 Marker-trait associations identified by JF and EJF models across ten chromosomes, using different map densities under selection $p$ at 0.001 for maize SLB resistance in NAM panel. Blue, cyan and gray peaks denote associations with RMIP greater than 0.5, 0.3 to 0.5 and less than 0.3, respectively. X-axes denote the genetic positions (cM) across ten chromosomes. RMIP values were summarized at the individual marker (every 0.2 cM) basis.
Ensemble between TAGGING-assisted QTL and GBLUP models showed their complementary effects on prediction ability. The mean within-family prediction $R^2$ values for QTL and GBLUP models were shown in first row and column, and other cells show the $R^2$ values for ensembles with equal weights for the two models in the corresponding rows and columns. The thinned maps of 20-cM inter-marker distances were used for the QTL models, under the best selection stringency ($p = 0.01$ for EJF; $0.05$ for ESF). * denotes $p$-value < 0.0005 in the one-sided binomial tests with the null hypothesis that the ensemble model predicted the same as the corresponding row GBLUP model (≥ 21 out of 25 families). The $R^2$ values were estimated from 50 replicates of cross validation. Standard errors of prediction ability based on variation among families ranged from 0.016 to 0.018.
### Tables

#### Table 3-1 Description of prediction models compared.

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Model Description</th>
<th>Genotypic input</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTL model</td>
<td>JF/EJF</td>
<td>(Ensemble) Joint Linkage analysis</td>
<td>subset(s) of linkage marker genotypes in one consensus map</td>
<td>(Ensemble of) joint multiple-family linkage analysis</td>
</tr>
<tr>
<td></td>
<td>SF/ESF</td>
<td>(Ensemble) Single Family analysis</td>
<td>subset(s) of linkage marker genotypes for each of 25 families</td>
<td>(Ensemble of) single family analysis</td>
</tr>
<tr>
<td></td>
<td>EJF+ESF</td>
<td>Ensemble of EJF and ESF models</td>
<td>As above</td>
<td>Average prediction of the two above</td>
</tr>
<tr>
<td>Genomic prediction model</td>
<td>SGBLUP</td>
<td>linkage map-based GBLUP model (within-family IBD)</td>
<td>25 $G$ matrices from 25 sets of linkage marker genotypes</td>
<td>Single family GP one family at a time</td>
</tr>
<tr>
<td></td>
<td>JGBLUP</td>
<td>Allele calling-based GBLUP model (cross-family IIS)</td>
<td>One $G$ matrix based on the actual genotypes most adjacent to linkage marker positions</td>
<td>Joint family GP</td>
</tr>
<tr>
<td></td>
<td>HGBLUP</td>
<td>Allele calling-based GBLUP model (cross-family IIS)</td>
<td>One $G$ matrix based on 1.6 M HapMap v1 SNPs</td>
<td>Joint family GP</td>
</tr>
</tbody>
</table>
References


Bernardo, R., 2004 What proportion of declared QTL in plants are false? Theoretical and applied genetics 109: 419-424.


Buckler, E., J. Holland, M. McMullen, S. Kresovich, C. Acharya et al., 2009 The genetic architecture of maize flowering time. Science 325: 714 - 718.


Hastie, T., R. Tibshirani, J. Friedman, T. Hastie, J. Friedman et al., 2009 The elements of statistical learning. Springer.


Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan et al., 2014 Usefulness of Multiparental Populations of Maize (Zea mays L.) for Genome-Based Prediction. Genetics 198: 3-16.


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


Schön, C. C., H. F. Utz, S. Groh, B. Truberg, S. Openshaw et al., 2004 Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167: 485-498.


Zou, G., G. Zhai, Q. Feng, S. Yan, A. Wang et al., 2012 Identification of QTLs for eight agronomically important traits using an ultra-high-density map based on SNPs generated from high-throughput sequencing in sorghum under contrasting photoperiods. Journal of Experimental Botany.
CHAPTER 4: Enhancing Genomic Prediction with Genome-wide Association Studies in
Multiparental Maize Populations
Enhancing Genomic Prediction with Genome-wide Association Studies in Multiparental Maize Populations

Yang Bian¹ and James B. Holland¹,²,*

¹ Department of Crop Science, North Carolina State University, Raleigh, NC 27695
² U.S. Department of Agriculture-Agricultural Research Service, Plant Science Research Unit, Raleigh, NC 27695

Keywords: Quantitative trait loci; genome-wide association studies; mixed model; genomic prediction

*corresponding author (james_holland@ncsu.edu)
Abstract

Genome-wide association mapping using dense marker sets has resulted in identification of some nucleotide variants affecting complex traits which have been validated with fine-mapping and functional analysis. Many sequence variants associated with complex traits in maize have small effects and low repeatability, however. In contrast to genome-wide association study, genomic prediction is typically based on models incorporating information from all available markers, rather than modeling effects of individual loci. In this study, we considered methods to integrate results of genome-wide association studies into genomic prediction models in the context of multiple interconnected families. We compared association tests based on a biallelic additive model constraining the effect of a SNP to be equal across all families in which it segregates, with those based on a multiallelic haplotype model in which the effect of a SNP can vary across families. Association SNPs were then included as fixed effects into a genomic prediction model that also included the random effects of the whole genome background. Simulation studies revealed that the effectiveness of this joint approach depends on the extent of polygenicity of the traits. Congruent with this finding, cross-validation studies indicated that genomic prediction including the fixed effects of the most significantly associated SNPs along with the polygenic background was more accurate than the polygenic background model alone for moderately complex but not highly polygenic traits measured in the maize nested association mapping population. Individual SNPs with strong and robust association signals can effectively improve genomic prediction.
Introduction

Genomic prediction (GP) based on genome-wide markers has emerged as a powerful supplement to conventional plant and animal breeding (Hayes and Goddard 2001; Meuwissen et al. 2001; Bernardo and Yu 2007; Heffner et al. 2009; Crossa et al. 2014). Exploiting recent dramatic decreases in genotyping costs, GP can accelerate selection cycles or increase effective population sizes under selection (Wang and Hill 2000) and is likely to increase selection gains per unit of time (Heffner et al. 2010; Butruille et al. 2015). Most genomic selection models assume approximately Fisher’s ‘infinitesimal’ genetic architectures (Hill 2014), in which each marker is assumed to be associated with very small genetic effects. In contrast, quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS) attempt to find individual markers associated with larger amounts of genetic variation than expected for a polygenic background effect.

Genomic prediction and GWAS are attempts to model different aspects of genetic architecture and have complementary advantages. Whereas GWAS may serve as initial evidence for the role of a particular gene in the inheritance of a complex trait, GWAS models are generally poor at representing the combined simultaneous effects of many small effect variants. In particular, with dense marker data sets, researchers often have more markers than observations, resulting in strong dependencies among SNP association tests. In contrast, common GP models rely on assumptions about the genetic architecture that permit simultaneous modeling of high dimensional SNP data and often good prediction ability, but high prediction ability usually comes at the expense of low interpretability. Consequently, it
seems inevitable to trade off model interpretability for model complexity in GP (Gianola and van Kaam 2008; González-Recio et al. 2008).

Genomic Best Linear Unbiased Prediction (GBLUP), or equivalently, Ridge regression Best Linear Unbiased Prediction (RR-BLUP; (Habier et al. 2013)) assumes equal marker variances and no epistasis, but works well for quantitative trait prediction in practice (Lorenzana and Bernardo 2009; Guo et al. 2011; Heffner et al. 2011; Heslot et al. 2012). GBLUP has become a widely used statistical method to predict genotypic values in breeding practice with less computational burden than more complex approaches (VanRaden 2008; Hayes et al. 2009; Piepho 2009). GBLUP uses single nucleotide polymorphisms (SNPs) to calculate the realized genomic relationships between individuals to mimic the genetic relationships at quantitative trait loci (QTL). The prediction ability of GBLUP relies on linkage disequilibrium (LD) between SNP loci and QTL and additive genetic relationships at QTL (Habier et al. 2013). GBLUP is expected to perform well when the quantitative trait is controlled by a large number of loci dispersed across the genome, because in the polygenic genetic architecture, the realized relationship matrix serves as a good approximation of genetic correlation at the numerous QTL.

GP generally outperforms QTL-based marker-assisted selection (MAS) in animal and plant breeding (Lorenzana and Bernardo 2009; Moser et al. 2009; Heffner et al. 2011). The poor predictive accuracy of MAS is due to inadequate power to detect small effect loci and treating markers that do not pass stringent thresholds as having zero effect, while conversely overestimating the effects of markers that are retained (Beavis 1998; Xu 2003; Schön et al. 2004). GWAS has some limitations similar to QTL mapping due to the common use of
stringent thresholds designed to minimize false positive associations, resulting in limited
repeatability of detection results across studies especially for small-effect associations (Bian

A variety of approaches have been proposed to account for variability in SNP-associated effects in genomic prediction. Bayesian models assume the genetic architecture includes a mixture of major and small effect QTL and polygenic background, and apply shrinkage and variable selection techniques, using prior assumptions for marker effect
distributions and hyper parameters for fine tuning the prior (Meuwissen et al. 2001; Park
and Casella 2008; Habier et al. 2011). The Bayesian approaches to GP are more
computationally intensive and do not scale well to large marker data sets. The parameter
estimates from Bayesian or non-parametric methods (kernel methods, random forest etc.) can
be sensitive to priors and unstable across studies (Gianola et al. 2009). More
computationally efficient methods have been proposed to weigh the genomic relationship
matrix to represent the relative size of variance explained by the corresponding loci (Zhang
et al. 2014; Ober et al. 2015; Tiezzi et al. 2015). These methods aim to improve prediction
and are not intended as gene discovery methods.

Methods that specifically integrate GWAS and GP methods might provide a way to
bridge the information coming from opposite ends of the genetic architecture spectrum and
provide more robust associations that could aid gene discovery while also improving
genomic prediction. Spindel et al. (2016) proposed combining a small number of significant
SNPs detected with GWAS as fixed effects in combination with an RR-BLUP model to
capture polygenic effects model in a panel of 332 rice breeding lines, resulting in superior
prediction performance (SPINDEL et al. 2016). The objectives of this study were to compare the prediction ability of models including only polygenic background effects to those integrating polygenic background effects with GWAS-based SNP ‘discoveries’ in the context of a large multiple family interconnected population in maize. A simulation study was conducted to compare the power and false discovery rate of association tests and the prediction ability of different modeling approaches under two different genetic architectures, differing by the effect sizes of a subset of makers. Real data from the maize nested association mapping were analyzed to compare predictive ability of different models for traits with different genetic architectures.
Materials and Methods

Plant phenotypes and genotypes

We used the maize nested association mapping (NAM) population comprising a set of ~5000 recombinant inbred lines (RILs) derived from crosses between a reference parent, inbred line B73, and 25 other diverse founder inbred lines of maize (MCMULLEN et al. 2009). Three quantitative agronomic traits in maize: resistance to southern leaf blight (SLB), Gray Leaf Spot (GLS), and Plant Height (PHT) were previously studied for genetic architectures, and the predicted mean values for NAM RILs were previously reported (BIAN et al. 2014; PEIFFER et al. 2014; BENSON et al. 2015) and adopted in this study. A second generation NAM linkage map consisting of 1476 markers with a uniform 1 cM inter-marker distance was used for linkage analysis (BIAN et al. 2014; SWARTS et al. 2014). An additive genomic relationship matrix (G matrix) for the whole NAM panel computed from the high-density markers in maize HapMap V1 of 1.6 million SNPs (PEIFFER et al. 2014) was also adopted here to represent the polygenic additive genetic relationships (VANRADEN 2008), as previously used in (BIAN AND HOLLAND 2015). A total of 4354, 3225 and 4359 RILs with both genotypic and phenotypic data were available for analyzing SLB, GLS and PHT, respectively. A total of 1,328,174 single-nucleotide markers (most SNPs and small indels) in Maize HapMap V2 genotypes (CHIA et al. 2012) had complete and polymorphic loci for 26 NAM lines. Those markers were referred to as SNPs in this study, and their genotypes were projected into NAM RILs based on the linkage map (BIAN et al. 2014) and used for GWAS of the three agronomic traits. We then sampled one of every ten of the GWAS HapMap V2
SNPs for each cM interval, and ensured at least two markers were kept for each interval, such that a total of 133,509 SNPs were used for the simulation analyses (described below).

*Simulation scheme*

We designed simulation of two distinct quantitative traits using the real genotypes in NAM populations to prelude empirical data analyses. Heritability was assigned at 0.56, with two sources of variances: additive main effects of causal loci and additive polygenic backgrounds. With the extensive coverage across the genome, the *G* matrix accurately measured the pairwise additive genetic relationships and population structure among all RILs. For the additional causal effects, ten independent SNPs were randomly chosen from each of ten chromosomes and assigned identical additive effects as the true variants. For the contribution from the rest of genome, we gave a random variance component to associate with the *G* matrix described above. The attributes of the first trait was described as follows: the overall heritability is 0.56 when having both sources, the proportion of variance explained (*R*^2^) by only polygenic effects (*G* matrix) is 0.32, and the *R*^2^ accounted for by the causal loci is 0.27. Since the causal and background effects have some correlation, their combined variation is less than the sum of the individual component variations. For the second trait, we kept the three attributes the same as in the first trait, but increased the number of causal loci from ten to 100 true variants assigned with equal additive genotypic values. Every chromosome then had 10 nearly independent causal loci that were at least 1 cM apart. We hereafter referred to the first trait as the “oligogenic trait”, and the second trait as the “polygenic trait”. The RIL samples were the same 4354 as in GWAS for SLB. The simulation did not include epistatic interactions.
Genomic BLUP models

Given the massive amount of genotypic data available in this study, any model that became prohibitively computationally intensive was not employed, because of the need to conduct many analyses for simulations and cross-validations. We therefore chose to use GBLUP model to represent a conventional GP model. The GBLUP model was expressed as follows:

\[ \mathbf{y} = \mathbf{1} \mu + \mathbf{Z} \mathbf{u} + \mathbf{e}, \]

where \( \mathbf{y} \) is the \( n \times 1 \) vector for the trait values of \( n \) NAM RILs, \( \mu \) is the \( n \times 1 \) intercept vector, \( \mathbf{u} \) is an \( n \times 1 \) vector of genotype random effects and has covariance structure

\[ \text{var}(\mathbf{u}) = \mathbf{G} \sigma_u^2, \]

where \( \sigma_u^2 \) is the additive variance, \( \mathbf{G} \) is the realized genomic relationship matrix as described above, \( \mathbf{Z} \) is a design matrix (identity matrix here) relating elements of \( \mathbf{y} \) to elements of \( \mathbf{u} \), and \( \mathbf{e} \) is an \( n \times 1 \) vector of error and \( \mathbf{e} \sim N(0, \mathbf{I} \sigma_e^2) \). We used the mixed models to accommodate additive genetic relationships and background polygenic effects by including random effects to account for correlations among lines.

We tested an alternative mixed model GBLUP by adding the NAM family indices as fixed effects into the regular GBLUP model.

\[ \mathbf{y} = \mathbf{1} \mu + \mathbf{r} \alpha + \mathbf{Z} \mathbf{u} + \mathbf{e}, \]

where the new term \( \mathbf{r} \) is the \( n \times 24 \) incidence matrix for the first 24 of the 25 NAM RIL families, and \( \alpha \) is the \( 24 \times 1 \) vector for family mean effects relative to the reference family \( (\mathbf{1} \mu + \mathbf{r} \alpha \) used reference coding). This model was called PGBLUP model, “P” for “population” effects.

111
Main effect and nested effect model in GWAS

We aimed to capture reliable genetic variance for the complex traits with mixed effect models, using GWAS to directly map outstanding variation due to genomic variants beyond polygenic background. Linear mixed model was applied to GWAS framework. The mixed model GWAS differed from its baseline GBLUP and PGBLUP model by searching for additional significant marker effects. The idea of including random effects of line relationships was to accommodate background effects to more reliably identify loci with lower false discovery rate (FDR). The main effect GWAS model was described as,

\[ y = 1 \mu + \tau \alpha + X\beta + Zu + e, \]

where \( X \) is \( n \times 1 \) marker genotype vector for \( n \) RILs at a marker locus, and \( \beta \) is the vector of marker main effect.

For each NAM population, every chromosome pair of a RIL line is a mosaic of the two founder haplotypes (one is B73 and the other is an alternate founder). Marker effects at QTL could vary by populations due to epistatic interactions between the causal variants and other factors in the genetic background, or due to functional allelic or haplotypic variation around the site being tested. To allow for the possible patterns of effects, we developed a nested effect model for GWAS that included separate coefficients representing the additive genetic effects of each founder allele (or the founder’s haplotype that is in tight LD with the test locus). Using this nested effect GWAS model, we can estimate the effects on the phenotype contributed by each of the alternate founder alleles. The marker genotype at a locus, as a continuous variable, is considered nested within population index, using the following mixed model:
\[
\mathbf{y} = \mathbf{1} \mu + \tau \mathbf{a} + \mathbf{W} \gamma + \mathbf{Z} \mathbf{u} + \mathbf{e},
\]
similar to the previous models, but introducing \( \mathbf{W} \), an \( n \times m \) design matrix, which are constructed by multiplying the marker genotype column into the population indices matrix and excluding the columns for populations in which the SNP is not segregating, and \( \gamma \), a vector of marker nested effects, that is, a separate “slope” for the marker loci within each population. In contrast to the main effect model, this model estimates \( m \) allele effects per SNP, where \( m \) ranges from 1 to 25 and represents the number of populations in which the SNP is segregating.

The two GWAS models were implemented by a two-stage procedure. In the first stage, restricted maximum-likelihood (REML) estimates of variance components \( \hat{\sigma}_u^2 \) and \( \hat{\sigma}_e^2 \) were calculated in the mixed model that omitted the \( \mathbf{X} \mathbf{\beta} \) or \( \mathbf{W} \gamma \) term, using EMMA implementation (Kang et al. 2008). The \( \mathbf{1} \mu + \tau \mathbf{a} \) terms were included as fixed effects in this baseline model, where GLS needed an additional flowering time fixed effect as a covariate (Benson et al. 2015). In the second stage, the significance of each marker loci was measured by the additional variation explained, with one degree of freedom for the marker main effect model or with \( m \) degrees of freedom for marker effect nested within population model, based on an F-test conditioning on previously estimated \( \hat{\sigma}_u^2 \) and \( \hat{\sigma}_e^2 \) (Kang et al. 2010; Segura et al. 2012). The variance components were estimated only once, and the same \( \mathbf{G} \) matrix was used throughout the genome scan (Kang et al. 2010; Zhang et al. 2010). The \( p \)-value was recorded for all markers.

Resample Model Inclusion Probability (RMIP) (Valdar et al. 2009) was computed to measure repeatability of detection for GWAS. RMIP was calculated for each SNP as the
proportion of cross-validation training data sets for which the marker was passed a particular \( p \)-value threshold. The power was estimated by the proportion of the 1-cM intervals to which the casual loci had been assigned that had significant GWAS associations, and the FDR was estimated by proportion of the 1-cM intervals that had significant GWAS associations but had not been assigned with a causal locus.

**Three prediction models**

We tested the hypothesis that incorporating association mapping discoveries from our new GWAS models can improve prediction abilities compared to GBLUP models. For the three agronomic traits SLB, GLS, and PHT, we identified the significant loci from training sets (80% of random draws of each population), and predicted the masked phenotypes in test sets (the rest 20% of RILs). We performed prediction using REML-based and MCMC-based linear mixed models for the three traits. First, REML-based linear mixed models fit all main effect GWAS discoveries passing the criteria as the fixed effects. Second, prediction models combining nested SNP association discoveries with the genomic background were fit using REML, however these models often did not converge because of collinearity or not-in-full-rank among fixed marker effects, so we do not report results from this model. Third, Bayesian linear mixed models fit the same main effect GWAS discoveries. Fourth, the Bayesian linear mixed models were applied to the same nested effect GWAS discoveries.

For the Bayesian models, a Gaussian prior (Bayesian Ridge Regression, BRR) or a double-exponential prior (Bayesian LASSO, BL) was assigned to the significant SNPs from GWAS modeling (Park and Casella 2008; De Los Campos et al. 2009; Pérez et al.)
R package ‘BLR’ (DE LOS CAMPOS et al. 2009) used a Gibbs sampler to draw samples from the posterior distributions. Degrees of freedom were set to 5, the hyper-parameter in BL was set to follow a gamma distribution, and we chose the shape \((s)\) parameter of the gamma density as 1.01, which allowed a relatively uninformative prior. We solved for the rate \((r)\) parameter to match the expected proportion of variance accounted for by the causal loci. A flat prior was used for all other fixed effects (covariates). The heritability was assumed to be 0.85 based on previous analyses of NAM populations, and 80\% of the genetic variance was assigned to the \(G\) matrix term (0.68), and the remainder of genetic variance was due to the additional causal loci (0.17) \textit{a priori}. A total of 50000 MCMC samples were drawn from the resulting posterior distribution, the initial 10000 iterations were discarded as burn-in, and the thinning interval was 5. In general, predicted values should be robust with reasonable hyper-parameters, although estimates of marker effects are more sensitive.

To understand the influence of oligogenic versus polygenic architecture on GWAS performance, we examined power and FDR for the two simulated traits at differential \(p\)-value stringencies in 50 runs of cross validation (80\% for training and 20\% for testing). In the 50 cross validation runs for the three agronomic traits and two simulated traits, we used the most significant GWAS associations within each cM interval that passed the \(p\)-value stringencies into the prediction models, avoiding overrepresentation of the same genetic information.
Cross-validation scheme

Prediction abilities of models were evaluated within each of the 25 biparental mapping populations using a cross-validation procedure. Training sets were created by randomly sampling 80% of RILs from each of the 25 NAM families. The remaining lines constituted the validation sample for the models created using the training data set. Within-family predictions were evaluated, and prediction abilities were measured as the proportion of total trait variance explained by the model ($R^2$) in the validation set. The $R^2$ was used to allow direct comparison to heritabilities, and it was estimated by averaging squared Pearson’s correlations ($R^2$) between the observed and predicted trait values. We gave a negative sign to few $R^2$ to correct for the wrong direction of correlations between predicted and observed line values.
Results

Simulation results: power, FDR and prediction $R^2$

The main effect GWAS model was implemented for the two simulated traits. Power and FDR were computed across a range of $p$-value significance thresholds. Power to detect SNP-trait associations in the oligogenic trait was substantially higher than that in the polygenic trait (Figure C-1 and C-2). As the $p$-value threshold decreased from $10^{-4}$ to $10^{-15}$, power decreased from 0.89 to 0.36 and FDR dropped from 0.88 to 0.30 in the oligogenic trait, and for the same amount of change in FDR, the power decreased from 0.53 to 0.01 in the polygenic trait. Power and FDR were highly related, so that there was no clear optimal $p$-value threshold for GWAS. At the Bonferroni adjusted $p$-value of $3.7 \times 10^{-7}$, an average of 27 SNP loci were discovered, of which about seven were true, resulting in an FDR of 0.72 and power of 0.70 for the oligogenic trait. For the polygenic trait FDR was 0.26, and power was 0.01.

Next, we incorporated significantly associated SNPs (selecting only the most significant SNP within 1-cM linkage map interval) into the REML-based linear mixed model to obtain predicted values of the two traits. The prediction abilities of the REML-based linear mixed model were compared to other benchmark models (Figure 4-1). The prediction $R^2$ in the GBLUP model was 0.33 for the oligogenic trait and 0.36 for the polygenic trait. It was expected that GBLUP model performed better for the polygenic trait because most of the trait variation was well approximated by the assumptions of the GBLUP model. Both $R^2$ were greater than the simulated proportion of variance truly due to polygenic background (0.32) because the $G$ matrix is able to absorb some of the variation due to the ‘major’ genes.
We also examined the prediction $R^2$ in a mixed model that fit the exact causal SNPs and $G$ matrix as we had simulated. Because this “causal model” has the correct predictors specified, it can be considered a theoretically optimal model. The ‘causal model’ had prediction ability $R^2$ from 14 – 15 percentage points lower than the heritability (0.56). Although the causal loci were known, the marker effects and background variance components required re-estimation for each training data set. Estimation errors in the causal model can be due to genetic sampling variance (affecting causal allele frequencies and introducing correlations among causal markers) as well as experimental error. Bernardo (2001) observed that for polygenic traits, even knowing which genes are causal does not solve problems in estimating their effects.

The integrated GWAS-GP model performed well with the oligogenic trait: the prediction $R^2$ increased from 0.33 in the basic GBLUP model to a maximum of 0.41 at the optimal $p$-value threshold ($p = 10^{-6}$; Figure 4-1). Within the commonly used $p$-value range, from $10^{-4}$ to $10^{-7}$, the GWAS-GP model had stable prediction ability, although it began to decrease as stringency increased further (Figure 4-1). In contrast, the GWAS-GP model provided very little improvement over GBLUP at its optimal $p$-value threshold, and provided substantially worse prediction ability outside of the $p$-value range between $10^{-4}$ and $10^{-7}$. For both traits, prediction ability was near optimal at the Bonferroni corrected $p$-value, and therefore we used this threshold in the analyses of the three agronomic traits.
GWAS of the three agronomic traits

We implemented two GWAS models for each of two quantitative disease resistance traits (SLB and GLS) and plant height (PHT). The main effect model assumed a constant difference between the SNP allele effects across families, whereas the nested effect model allowed the allele effects to vary across families. The discoveries from the two models were obviously clustered in a relatively few genomic regions (Figure 4-2, C-3, C-4, C-5, C-6, C-7, and C-8). However, main effect SNP GWAS models identified strong and repeatable associations in chromosome 6 for SLB and chromosome 1 for PHT alone, while nested GWAS models appeared more powerful for associations on chromosome 4 for GLS and chromosome 7 for PHT. The complementary discoveries indicate that a mixture of the two genetic assumptions might be closer to reality.

The average number of discovered SNP loci was 21 for SLB (range from 18 to 25 loci), 14 for GLS (range from 7 to 24) and just 0.9 for PHT (range from 0 to 5), when fitting the main effect GWAS model in the 50 training sets. The average number of significant loci was 21 (18 to 26) for SLB, 23 (13 to 32) for GLS, 1 (0 to 4) for PHT, when fitting the nested GWAS model in the 50 training sets. The small number of PHT associations detected and their limited repeatability suggest that the infinitesimal genetic model should be sufficient for PHT (Figure C-7 and C-8).

Prediction $R^2$ in SLB, GLS and PHT

We investigated the impact of fitting the significant repeatable association SNPs into the GBLUP and PGBLUP models and evaluated the performance relative to these two
baseline GBLUP models (Figure 4-3). For SLB, discoveries from both GWAS models significantly improved prediction $R^2$ over the GBLUP/PGBLUP models. For GLS, the nested GWAS associations increased prediction $R^2$ significantly over both baseline GBLUP models, while the main effect model increased prediction ability significantly only over the GBLUP model but not the PGBLUP model.

For PHT, there appeared no improvement over the baseline models, although the new approach did not introduce overfitting or decrease prediction accuracy due to fitting conservative GWAS discoveries. The true effects of underlying QTL should be small for PHT, and so the accuracy with which these effects were estimated was expected to be low although a large sample was used. Being conservative about the GWAS discoveries (< 1 on average) guarded against overfitting problem common seen in prediction practice. The prediction $R^2$ varied by NAM populations for all traits we studied (Figure C-9). This phenomenon may occur because the segregating QTL are population dependent and the founders’ genetic relatedness to the common parental line B73 varies substantially.
Discussion

In this study, we proposed new GWAS models for the maize NAM populations and tested if their results could be used to enhance genomic prediction. The refined genetic architectures comprised a polygenic portion with covariance relationships estimated by genome-wide SNPs (and summarized in the G matrix) and an additional portion due to loci with individually significant effects. Simulation studies showed the new main effect model had good detection power in identifying oligogenicity with even small effects (explaining ~3% phenotypic variations). In contrast, polygenes (100 causal loci, each ~ 0.3% phenotypic variations) were very difficult to detect from the background without a very high false discovery rate. One way to bridge the disconnected interest between mapping gene and genomic prediction is to have an integrative modelling system to detect repeatable association loci to exploit them in prediction performance. We identified reliable associations for the two disease resistance traits and used them to improve prediction accuracy significantly.

Power in main effect vs nested effect GWAS models

Previous GWAS analyses of multiparental RIL populations assumed that the same allele always has the same effect independent of the source of multiparental populations. By nesting SNP genotypes effects in founders, we allow the flexibility of allelic effects across the populations. SNP effects may vary across families because they may tag different clusters (haplotypes) of functional variants in different founders, or they may interact epistatically with the genetic background. Statistically, the new nested effect GWAS model is an
extension of the NAM joint-family linkage model, except that identity in state between parents results in zero SNP effect in the new model, and the genetic background effects are controlled with a random polygenic effect rather than by model selection.

Conversely, the additive SNP effect model may be more realistic in some cases because LD among loci is weak in diverse maize, the current marker coverage is extremely high across the genome, and the importance of epistasis in maize is uncertain. From those perspectives, the main effect model should be used to gain greater power and more precise estimation of the marker effect.

*Genetic architecture and implication to breeding practice*

Although most agronomic traits are quantitatively inherited, relatively large effect genes may still be involved in their inheritance, and their presence has a large impact on which genomic prediction model is optimal. A previous simulation study showed that when a few (1–3) major genes are present for a quantitative trait and each major gene accounts for 10% of genetic variance, fitting these major genes as fixed effects are beneficial to the genome-wide prediction model (Bernardo 2014). In contrast, with polygenicity, many genetic effects must be estimated, and therefore it is unlikely to get good estimates of genetic effects with even large sample sizes. Our GWAS approach distinguished the purely polygenic trait (PHT) from the ones that may have major genes, and by explicitly modeling the effects of repeatable loci, if any, in addition to the infinitesimal model, the new GP model achieved good and robust performance across traits.
The genetic architectures provided by this study imply diverse breeding strategies are needed. Targeted introgression of favorable alleles or elimination of deleterious alleles can be employed in breeding if repeatable marker-trait associations are found. If no major marker-trait association is identified, however, phenotypic selection or GP should be given priority over targeting specific regions or loci. Highly repeatable and robust associations are those that are most likely to improve selection response when incorporated with genomic prediction models. These associations also can be considered higher priority targets for high resolution genetic analysis, to aid in the discovery of genes and genetic pathways underlying complex traits.

Using an extreme density of genome-wide marker set to represent the polygenic contribution may have costed a reduction in detection power. If causal loci are in strong LD with other markers, the relationships modeled by the $G$ matrix will account for some shared proportion of genetic variations at major effect loci as well as the polygenic loci, resulting in a reduction in the power of GWAS. Nevertheless, loci with moderate to strong effects can still stand out of the background and be captured as fixed effects. In the future, we plan to remove the local correction during association scan; each cM interval scan is scanned with a $G$ matrix calculated on all chromosomes but the current one (DELL’ACQUA et al. 2015), and the within-interval SNP-trait associations are to be tested with and without the boundary linkage markers included as covariates (a three-dimensional GWAS scans in the mixed model context).

Although GWAS had low power of detections for PHT due to more polygenic genetic architectures, our strategies incorporating the association mapping discoveries into prediction
did not cause overfitting. It is important to have a robust approach suitable to different traits. In addition, accounting for the different population mean effects using the PGBLUP model did not benefit prediction in NAM populations (Figure 4-3) beyond the conventional GBLUP model, as the $G$ matrix alone appears to mostly account for the effects of variation among family means.

*The Future plans*

This study investigated within-family prediction. Further efforts needs to be put into studying the best practice of modeling in cross-population prediction. How this strategy would perform in training on multiple related or unrelated small biparental families to predict progeny from other crosses is an important question. To address the question, studies need to be done in optimizing the allocation of resources in GP using the current method. Second, it is important to extend the GP framework into multi-trait joint selection (JIA AND JANNINK 2012). Higher hierarchical genetic data such as RNA profiles, transcriptomic data, and other biomarkers may be crucial to enhance the current DNA marker source (RIEDELSHEIMER et al. 2012).

*Acknowledgment*

We thanked the High Performance Computing Services in NC State University. Research was supported by United States National Science Foundation grants IOS-1238014 and IOS-0820619 and the US Department of Agriculture, Agricultural Research Service. YB
is a graduate student at NC State University and supported by NSF grant IOS-1238014 and a Monsanto Co. fellowship.
Figure 4-1 Prediction $R^2$ in the simulated oligogenic and polygenic traits. “Causal” represents the true (ideal) model as the traits were simulated, “GBLUP” represents the “background only model” with no fixed effects of SNPs, the numbers in the horizontal axis indicate the $p$-value thresholds for declaring a significant locus, and “Bon.” represents the Bonferroni correction $p$-value. “Main effect GWAS + REML LMM” denotes linear mixed models combining significant GWAS associations and polygenic background effects. The margins represent 95% confidence intervals of the mean prediction $R^2$ in 50 cross validation runs.
**Figure 4-2 GWAS repeatability plot for the three agronomic traits.** Each data point represents the resample model inclusion probability (RMIP) of a SNP with a significant association in one or more of 50 GWAS analyses of training data sets containing 80% of lines. Two filters were applied to the significant SNPs: first, only one association loci in one cM interval was considered, certifying the chosen SNPs to be the most competitive in the neighboring region, and second, the chosen SNPs’ \(p\)-values needed to pass the Bonferroni correction (\(p\)-value < 3.8 x 10^{-8}). Asteroid data points denote those loci that have average \(p\)-values more significant than the Bonferroni corrected \(p\)-value in 50 GWAS analyses. Venn diagrams show the numbers of highly repeatable loci (RMIP \(\geq 0.05\)) identified in the two models.
Figure 4-3 Within-family Prediction $R^2$ of the three agronomic traits using GBLUP and the new models. The sizes of models were constrained by the same two filters. The different letters indicate the average prediction $R^2$ values are significantly different by Tukey’s HSD (honestly significant difference) test at type I error rate of 0.05. All average prediction $R^2$ were computed based on a total of 25 NAM populations in 50 runs of cross validations. For 21 runs in main effect GWAS model and 29 runs in nested effect GWAS model that did not identify any associations for PHT, the prediction $R^2$ were supplemented by the corresponding GBLUP models.
References


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


APPENDICES
Appendix A: Supplemental Information for Chapter 2

Limits on the reproducibility of marker associations with southern leaf blight resistance in the maize nested association mapping population

Yang Bian1†
Email: ybian2@ncsu.edu

Qin Yang2†
Email: qyang6@ncsu.edu

Peter J Balint-Kurti2,3
Email: Peter.Balint-Kurti@ars.usda.gov

Randall J Wisser4
Email: rjw@udel.edu

James B Holland1,3*
* Corresponding author
Email: Jim.Holland@ars.usda.gov

1 Department of Crop Science, North Carolina State University, Raleigh, NC 27695, USA
2 Department of Plant Pathology, North Carolina State University, Raleigh, NC 27695, USA
3 U.S. Department of Agriculture-Agricultural Research Service, Plant Science Research Unit, Raleigh, NC 27695, USA
4 R.J. Wisser, Department of Plant and Soil Sciences, University of Delaware, Newark, DE 19716, USA
† Equal contributors.

Keywords
Quantitative trait loci, Nested association mapping, Disease resistance, Genome wide association study, Zea mays
Figure A-1 Histograms and box plots of prediction correlation coefficients from 100 random cross-validation analyses sets randomly subsampled from the common set of 4354 RILs. Current QTL Model = Model 7; 2011 QTL Model = Model 1.
File A-1

Supporting data


File A-1: Tables A-1 – A-6 12864_2014_6799_MOESM1_ESM.xlsx. (XLSX 164 KB)
Availability of supporting data

The data sets supporting the results of this article are available at the Panzea.org repository,
http://panzea.org/db/gateway?file_id=Kump_etal_2011_Nat_Genet_SLB_pheno_data (for raw data) and at the LabArchives.com repository,
https://mynotebook.labarchives.com/share/SLB%2520GWAS%2520reanalysis/MjIuMXw0MDg2OC8xNy9UcmVITm9kZS80MTgyNjQ1ODh8NTYuMQ==
(for updated BLUPs and SAS codes to perform NAM joint linkage analysis).
Appendix B: Supplemental Information for Chapter 3

Ensemble Learning of QTL Models Improves Prediction of Complex Traits

Yang Bian\textsuperscript{1} and James B. Holland\textsuperscript{1,2,\ast}

\textsuperscript{1} Department of Crop Science, North Carolina State University, Raleigh, NC 27695
\textsuperscript{2} U.S. Department of Agriculture-Agricultural Research Service, Plant Science Research Unit, Raleigh, NC 27695

\ast corresponding author (james_holland@ncsu.edu)
Figure B-1 Within-family prediction $R^2$ for plant height in biparental and multiple-family prediction, comparing JF, SF and the TAGGING models using multiple map densities at multiple selection $p$. The number at the top of each boxplot was the mean $R^2$ among 25 NAM families. The x axis represents the densities of linkage maps used in either JF, SF and TAGGING models.
Figure B-2 Within-family prediction $R^2$ for days to anthesis in biparental and multiple-family prediction, comparing JF, SF and the TAGGING models using multiple map densities at multiple selection $p$. The number at the top of each boxplot was the mean $R^2$ among 25 NAM families. The x axis represents the densities of linkage maps used in either JF, SF and TAGGING models.
Figure B-3 Prediction bias\(^2\) and variance in TAGGING and discrete QTL models for plant height (top) and days to anthesis (bottom). X-axes denote the inter-marker distances on the linkage maps used; “0.2” refers to discrete JL or SF models that used 0.2 cM density map, and the rest for TAGGING methods. Dot radius was scaled to within-family \(R^2\) calculated based on all RILs in the test sets in order to show the consistency between the expected mean error measures and mean prediction \(R^2\).
Figure B-4 Marker-trait associations identified by JF and EJF models across ten chromosomes, using different map densities under selection $p$ at 0.001 for plant height in NAM panel. Blue, cyan and gray peaks denote associations with RMIP greater than 0.5, 0.3 to 0.5 and less than 0.3, respectively. X-axes denote the genetic positions (cM) across ten chromosomes. RMIP values were summarized at the individual marker (every 0.2 cM) basis.
Figure B-5 Marker-trait associations identified by JF and EJF models across ten chromosomes, using different map densities under selection $p$ at 0.001 for days to anthesis in NAM panel. Blue, cyan and gray peaks denote associations with RMIP greater than 0.5, 0.3 to 0.5 and less than 0.3, respectively. X-axes denote the genetic positions (cM) across ten chromosomes. RMIP values were summarized at the individual marker (every 0.2 cM) basis.
Figure B-6 Resample Model Inclusion Probabilities (RMIP) distributions for genetic resistance to SLB using EJF models with 100 sets of 20 cM reduced maps at selection \( p = 1e^{-5}, 1e^{-4}, 1e^{-3} \) and \( 1e^{-2} \). More association were detected with more relaxed \( p \).
**Figure B-7** Ensemble of QTL and GBLUP models showed complementary effects on prediction ability (left, plant height, right, days to anthesis). The mean within-family prediction $R^2$ values for individual QTL and GBLUP models are shown in first row and column, and the rest of cells show the $R^2$ values for the equally weighted ensembles of the two models in the corresponding rows and columns. The thinned maps of 20-cM inter-marker distances were used for the QTL models, under the best selection stringency ($p = 0.01$ for EJF; 0.05 for ESF). * denote $p$-value < 0.0005 in the one-sided binomial tests with the null hypothesis that the ensemble model predicted the same as the corresponding row GBLUP model ($\geq 21$ out of 25 families). The $R^2$ values were generated from 50 replicates of cross validation described in the materials and methods. Standard errors of prediction ability based on variation among families ranged from 0.013 to 0.016 for plant height and from 0.017 to 0.021 for days to anthesis.

<table>
<thead>
<tr>
<th>Model</th>
<th>Plant height</th>
<th>Days to anthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Individual model</td>
<td></td>
</tr>
<tr>
<td>SGGLUP</td>
<td>0.465</td>
<td>0.465</td>
</tr>
<tr>
<td>JGGLUP</td>
<td>0.481</td>
<td>0.497</td>
</tr>
<tr>
<td>HGGLUP</td>
<td>0.493</td>
<td>0.498</td>
</tr>
<tr>
<td></td>
<td>0.428</td>
<td>0.446</td>
</tr>
<tr>
<td>SGGLUP</td>
<td>0.450</td>
<td>0.458</td>
</tr>
<tr>
<td>JGGLUP</td>
<td>0.461</td>
<td>0.485</td>
</tr>
<tr>
<td>HGGLUP</td>
<td>0.510</td>
<td>0.507</td>
</tr>
</tbody>
</table>

Low to high within family prediction $R^2$ values
Figure B-8 Resample Model Inclusion Probabilities (RMIP) of JF models that used different linkage map densities under selection $p = 1e^{-3}$ for genetic resistance to SLB. The patterns of RMIP landscapes were mostly same among JF using 0.2 – 2 cM density maps, while further reducing map densities compromised precision by losing and merging marker associations of small effects into available marker positions.
Table B-1 – B-6

Available for download as Excel files at www.g3journal.org/lookup/suppl/doi:10.1534/g3.115.021121/-/DC1

**Table B-1** Prediction $R^2$ and other statistics for genetic resistance to SLB, comparing JF, SF and the ensemble QTL analyses using multiple map densities and under multiple selection $p$.

**Table B-2** Prediction $R^2$ and other statistics for plant height (PHT), comparing JF, SF and the ensemble QTL analyses using multiple map densities and under multiple selection $p$.

**Table B-3** Prediction $R^2$ and other statistics for days to anthesis (DA), comparing JF, SF and the ensemble QTL analyses using multiple map densities and under multiple selection $p$.

**Table B-4** Prediction results in EJF, ESF, EJF+ESF, SGBLUP, JGLUP, HGBLUP, and their ensemble models.

**Table B-5** ANOVA for factorial experiment in ensemble learning involving EJF+ESF, SGBLUP, JGBLUP, and HGBLUP models.

**Table B-6** Prediction performance of ensemble of EJF, ESF, SGBLUP models weighted by two natural coefficients and Nelder–Mead optimization toward test set phenotypes.
Supporting data

File B-1. NAMSLB.RData. figshare.com/s/7dde4a3cedc611e4ac5606ec4bbcf141. Mean values for Southern leaf blight disease scores and linkage map marker scores for NAM RILs formatted as an R data object.

File B-2. SLB_genopheno.sas7bdat. figshare.com/s/687458c6edc611e4ae3406ec4bbcf141. Mean values for Southern leaf blight disease scores and linkage map marker scores for NAM RILs formatted as a SAS data set.

File B-3. NAMPHT.RData. figshare.com/s/b5e82c54edc611e4ae3406ec4bbcf141. Mean values for plant height and linkage map marker scores for NAM RILs formatted as an R data object.

File B-4. PHT_genopheno.sas7bdat. figshare.com/s/93761014edc611e4b69206ec4bbcf141. Mean values for plant height and linkage map marker scores for NAM RILs formatted as a SAS data set.

File B-5. NAMDA.RData. figshare.com/s/c367d258edc611e496ee06ec4bbcf141. Mean values for days to anthesis and linkage map marker scores for NAM RILs formatted as an R data object.

File B-6. DA_genopheno.sas7bdat. figshare.com/s/d67633eeedc611e49c4306ec4bbcf141. Mean values for days to anthesis and linkage map marker scores for NAM RILs formatted as a SAS data set.

File B-7. relMat.RData. figshare.com/s/d6069ea4edc511e49c4306ec4bbcf141. Realized additive genomic relationship matrix for NAM lines based on 1.6M HapMap I markers (courtesy of Dr. Jason Peiffer).
**File B-8.** NAMSLB_IIS.RData. figshare.com/s/cc9cdd00edc611e4ae3406ec4bbcf141. Mean values for Southern leaf blight disease scores and Identity In State (IIS) calls for HapMap markers closest to NAM linkage map markers for NAM RILs formatted as an R data object.

**File B-9.** NAMPHT_IIS.RData. figshare.com/s/9bc11e3aedc611e4ac5606ec4bbcf141. Mean values for plant height and Identity In State (IIS) calls for HapMap markers closest to NAM linkage map markers for NAM RILs formatted as an R data object.

**File B-10.** NAMDA_IIS.RData. figshare.com/s/71f708b2edc611e4994406ec4bbcf141. Mean values for days to anthesis and Identity In State (IIS) calls for HapMap markers closest to NAM linkage map markers for NAM RILs formatted as an R data object.

**File B-11.** ESF.sas. figshare.com/s/44c43d7edc611e49c4306ec4bbcf141. SAS code to conduct ensemble single family QTL analysis.

**File B-12.** JSF.sas. figshare.com/s/5dbb14a6edc611e49c4306ec4bbcf141. SAS code to conduct ensemble joint family QTL analysis.

**File B-13.** GBLUP and ensemble model prediction SLB.R. figshare.com/s/2d33f2bcedc611e4ae3406ec4bbcf141. R code to ensemble GBLUP and ensemble-QTL-based predictions.
Appendix C: Supplemental Information for Chapter 4

Enhancing Genomic Prediction with Genome-wide Association Studies in Multiparental Maize Populations

Yang Bian\textsuperscript{1} and James B. Holland\textsuperscript{1,2,*}

\textsuperscript{1} Department of Crop Science, North Carolina State University, Raleigh, NC 27695
\textsuperscript{2} U.S. Department of Agriculture-Agricultural Research Service, Plant Science Research Unit, Raleigh, NC 27695

\textit{Keywords}: Quantitative trait loci; genome-wide association studies; mixed model; genomic prediction

*corresponding author (\texttt{james\_holland@ncsu.edu})
Figure C-1  **Power and false discovery rate in oligogenic trait simulation.** X-axis shows significance claim thresholds in terms of $-\log_{10}(p\text{-value})$. Dots on lines indicate the Bonferroni adjusted $p$-value threshold.
Figure C-2 Power and false discovery rate in polygenic trait simulation. X-axis shows significance claim thresholds in terms of –log10(p-value). Dots on lines indicate the Bonferroni adjusted p-value threshold.
Figure C-3 and C-4 GWAS Manhattan plots for SLB using main effect and nested effect models. Average $-\log_{10}(p)$ shows the average significance over 50 runs, each of which used random 80% of each NAM population.
Figure C-5 and C-6 GWAS Manhattan plots for GLS using main effect and nested effect models. Average $-\log_{10}(p)$ shows the average significance over 50 runs, each of which used random 80% of each NAM population.
Figure C-7 and C-8 GWAS Manhattan plots for PHT using main effect and nested effect models. Average $-\log_{10}(p)$ shows the average significance over 50 runs, each of which used random 80% of each NAM population. The green bar shows the threshold at Bonferroni corrected $p$-value of 0.05.
Figure C-9 Boxplots of average prediction $R^2$ for 25 NAM populations, comparing GBLUP model and main effect GWAS + REML-based linear mixed model. The blue number on the top shows the average number of significant GWAS association loci. The “+” signs and black numbers indicate of the overall mean $R^2$. 