SELECTION AND CHARACTERIZATION OF PREVIOUSLY PLANT-VARIETY-PROTECTED COMMERCIAL MAIZE INBREDS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Travis J. Beckett

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Torbert R. Rocheford, Chair

Department of Agronomy

Dr. William M. Muir

Department of Animal Sciences

Dr. Mitch R. Tuinstra

Department of Agronomy

Dr. Mohsen Mohammadi

Department of Agronomy

Dr. Klaus L. Koehler

Department of Agronomy

Approved by:

Dr. Ronald F. Turco

Head of the Department of Agronomy Graduate Program

To my wife, Charisse, and to my four children, Blake, Ashley, Brayden, and Isaac. I go to work so that I can come home every day (and not the other way around).

ACKNOWLEDGMENTS

I extend my deepest gratitude to my wife, Charisse Beckett, for going through this whole thing with me. She probably deserves this degree more than me. Compared to her, I had the easy job. She had the harder but far more rewarding one–raising our kids in our home. I also thank my kids Blake, Ashley, Brayden, and Isaac for the energy they generate in our home, and for the joy they each bring into my life.

At the conclusion of my official educational experience (though I'll always continue learning) I'd like to mention my two favorite classes. One was an ecology course at Cambridge University in England. The other was a population genetics course at Purdue. Both courses presented the core principles of their respective discipline, gave me the tools to critically evaluate a problem, then required me to take it to the next level by putting the pieces together and making inferences. It is a pattern of learning that I have tried to emulate in every pursuit since.

I thank my graduate committee–namely Drs. Torbert Rocheford, Mitch Tuinstra, Bill Muir, Klaus Koehler, and Mohsen Mohammadi–for guiding me, teaching me, and critically evaluating my work. I am thankful that they allowed me to make a few mistakes for the sake of learning.

I express gratitude to Dow AgroSciences (DAS) for providing the funds for this research. DAS also provided on-site space for growing yield trials, lab supplies and labor for genotyping, expert consultation, and direct mentoring by Jason Morales, Juan Rey, and others.

Finally, I express my gratitude to a few unofficial mentors at Purdue who stepped up to answer my questions, offer valuable advice, and guide me when I needed it most and didn't know where else to turn–Dr. Linda Lee, Dr. Mohsen Mohammadi, and Dr. John Thompson. I will never underestimate the value of a compassionate advocate. I will do the same whenever I am called upon to help someone else.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABBREVIATIONS

ACRE	Agronomy Center for Research and Education
CL	Cob length
CR	Cob rows
DAS	Dow AgroSciences (now Corteva Agrisciences)
DP	Days to pollen shed
EH	Ear height
Ex-PVP	Expired Plant Variety Protection
GD	Genetic diversity
GEBV	Genomic estimated breeding value
GM	Grain moisture at harvest, in percent
GWA	Genome-wide association
GY	Grain yield at harvest, in kg/ha
IBS	Identity by state
MP	Mid-parent value
NAM	Nested association mapping
PH	Plant height
PVP	Plant Variety Protection
RIL	Recombinant inbred line
SL	Spike length
SNP	Single nucleotide polymorphism
SP	Superior progeny
TBN	Tassel branch number
TW	Tassel weight (as used in Ch. 3)
TW	Test weight (as used in Ch. 2)

ABSTRACT

Beckett, Travis J. Ph.D., Purdue University, December 2018. Selection and Characterization of Previously Plant-Variety-Protected Commercial Maize Inbreds. Major Professor: Torbert Rocheford.

The use of genotypic markers in plant breeding has greatly increased in the last few decades. In this dissertation, I report on three topics that illustrate how genotypic marker information can be applied in maize breeding to increase genetic gain. In the first chapter[1], I describe how genotypic and phenotypic data can be used to predict the mean, variance, and superior progeny mean of virtual biparental populations. I use these predictions to identify optimal breeding crosses out of a commercially relevant collection of North American dent inbreds. In the second chapter, within the context of early generation maize inbred development, and using a hybrid testcross data set, I report on the change in genomic prediction accuracy as the size of the training set increases and compare the accuracy of different genomic selection models. In the third chapter[2], I used a multi-variable linear regression approach known as genome-wide association (GWA) analysis to identify particular genetic locations, known as quantitative trait loci (QTL), that are associated with maize inflorescence traits.

[1] Chapter 1 has been submitted for publication in the academic journal *Crop Science*.

[2] Chapter 3 has been submitted for publication in the academic journal *Plant Breeding*.

Prologue

Projections show that by the year 2050, the global population will reach 9 billion (Northoff, 2016). Accordingly, experts have estimated that overall food production must increase by 70%, with a 51% yield increase needed in soybean, 60% in wheat, 46% in rice, and 71% in corn (Bruinsma et al., 2009; Kruse, 2010). Past increases in corn yield have been primarily due to the move from open-pollinated to double-cross then single-cross hybrids, improved agronomic practices, increased fertilizer applications, higher plant populations, and genetic modifications for resistance to pests and herbicides. Future increases in corn yield could come from a number of areas, including precision agriculture, biotechnology, novel genetic variation, and the use of genetic markers.

From a plant breeder's perspective, improvements in corn yield can be described in terms of genetic gain. Genetic gain can be modeled by the following basic formula (Fehr, 1991):

$$G_L = \frac{ir\sigma_A^2}{L} \tag{1}$$

where: G_L is the expected genetic gain per unit time L; i is the selection intensity, or the proportion of individuals chosen to be parents of the next generation; r is the selection accuracy; σ_A^2 is the additive genetic variance; and L is the length of the breeding cycle.

Genomic prediction, defined as the use of genotypic markers and phenotypic data to make predictions for individuals with known genotype but unknown phenotype, can be used to increase genetic gain in several ways. One way is to increase the selection intensity by using predictions to zero in on the individuals with the best predicted performance, and discarding the rest. Another way is to improve selection accuracy by adding predicted performance values to the collection of parameters that breeders traditionally use for selection. Yet another way is to decrease the length of the breeding cycle. One way this is done is by replacing part or all of the first stage of testing with genomic selection–a practice that is common in contemporary commercial maize breeding programs. This effectively reduces the length of that breeding cycle to the length of time that it takes to genotype the seeds and obtain the predictions from the statistical model.

The research projects described in the three chapters in this dissertation all relate to this major theme of applying genomic analysis tools in maize breeding. For all three projects, I used North American dent maize inbreds, the majority of which were formerly elite commercial inbreds with expired Plant Variety Protection certificates (hereafter referred to as ex-PVP inbreds). I supplemented these with a collection of public inbreds that are key progenitors, or founders, of these ex-PVP inbreds. A comprehensive study of the genetic relatedness of these inbreds was presented in the first chapter of my Master's Thesis (T. J. Beckett, 2016) and published shortly thereafter in the journal *PLOS One* (T. Beckett, Morales, Koehler, & Rocheford, 2017).

In Chapter 1[3] of this dissertation, I use the predicted mean, variance, and superior progeny mean of simulated biparental populations to identify optimal breeding crosses for improved hybrid performance with tester PHP02. Several recent studies have reported on similar models that predict progeny variance within simulated or virtual biparental populations (Bernardo, 2015; Mohammadi, Tiede, & Smith, 2015; Lehermeier, Teyssèdre, & Schön, 2017; Osthushenrich, Frisch, & Herzog, 2017). However, none use hybrid traits in the training set. In particular, two of these authors (Bernardo, 2015; Osthushenrich et al., 2017) state that this method could be used to predict topcross performance of simulated biparental populations. This is precisely what is accomplished in Chapter 1 of this dissertation.

In Chapter 2, I detail the possibilities of using genomic prediction (GP) in early generation maize inbred development, by reporting on the change in prediction ac-

[3] Chapter 1 has been submitted for publication in the academic journal *Crop Science*.

curacy when training set size increases and when different selection models are used. GP models have generally been used within crop breeding to: (1) predict additive effects in early generations, thus reducing the time per selection cycle; and (2) predict both the additive and dominance effects of later generations in order to determine the true commercial value of a line (Crossa et al., 2017). This chapter focuses on the former–predicting the additive effects in early generations. Several statistical models have been proposed, each with a slightly different way of solving the overarching $p >> n$ problem (i.e. the number of markers is much greater than the number of individuals) inherent in GP (de los Campos, Hickey, Pong-Wong, Daetwyler, & Calus, 2012; Crossa et al., 2017). In this chapter, I compare the accuracy of four GP models: RR-BLUP (Endelman, 2011), BayesB (Meuwissen, Hayes, & Goddard, 2001), partial-least squares (Wehrens & Mevik, 2007), and Random Forest (Breiman, 1996, 2001; Liaw & Wiener, 2002).

In Chapter 3[4], I report on the results of a genome-wide association study that identified quantitative loci (QTL) associated with inflorescence traits in maize. Maize is a naturally cross-pollinating species, with pollen grains mostly dispersed by wind. Thus, filial-1 (F1) hybrid seed production favors larger tassels on inbreds, with more pollen grains put out per plant. However, over the past several decades as plant breeders have selected for improved hybrid performance, tassel sizes have decreased (Lambert & Johnson, 1978; Fischer, Edmeades, & Johnson, 1987; Meghji, Dudley, Lambert, & Sprague, 1984; Duvick, Smith, & Cooper, 2010). Identification of the particular genetic locations that control both tassel and ear inflorescences could help breeders to better manipulate these traits to increase the efficiency of hybrid seed production and the level of hybrid performance.

Completing these research projects has already proven to be a valuable experience for me as I embark on my career as a commercial plant breeder.

[4] Chapter 3 has been submitted for publication in the academic journal *Plant Breeding*.

1 RE-IMAGINING MAIZE INBRED POTENTIAL: IDENTIFYING OPTIMAL BREEDING CROSSES USING GENETIC VARIANCE OF SIMULATED PROGENY

1.1 Abstract

Proper choice of parents for new breeding populations is essential in developing new maize (*Zea mays* L.) inbreds with improved hybrid performance. Breeders have traditionally chosen breeding populations based on mid-parent (MP) value, or predicted progeny mean. When two breeding populations have the same MP value, an accurate prediction of progeny variance may reveal which population has a greater potential for genetic gain. In this study we used inbred genotypes and hybrid phenotypes from 246 former commercial inbreds with expired Plant Variety Protection certificates along with 39 historically important public North American dent inbreds, all testcrossed to Iodent-type inbred PHP02. We used the R package PopVar to simulate bi-parental populations, perform genome-wide prediction, and predict the progeny mean, genetic variance, and superior progeny mean for grain yield, grain moisture, and test weight within each virtual population. Optimal breeding crosses were identified based on the mean and variance of virtual progeny. Results show that mixing germplasm from different proprietors in new breeding crosses can produce inbreds with improved performance in a hybrid testcross. The simulation and prediction model presented in this study may help breeders to identify parental pairs with the greatest potential for genetic gain in hybrid crop breeding programs.

1.2 Introduction

When selecting maize inbred parents for new breeding populations, breeders have primarily considered the MP valuedefined as the mean value of the trait exhibited by the two parents (Bernardo, 2014; Mohammadi et al., 2015). Other factors that influence choice of inbred parents include heterotic combining patterns, genetic background from pedigree and/or genotypic data, and results from molecular marker-based genome-wide predictions (Bernardo, 2014; Mohammadi et al., 2015; Kadam, Potts, Bohn, Lipka, & Lorenz, 2016; Lehermeier et al., 2017; Osthushenrich et al., 2017). When two breeding populations have similar MP values, however, a prediction of progeny variance can differentiate the two populations in terms of projected value, or potential genetic gain (See Fig. 1.1).

Precise and accurate predictions of progeny variance, however, have proved difficult to obtain (Bernardo, 2015). Most published attempts employed various measures of genetic diversity (GD) based on parentage (Cowen & Frey, 1987; Souza & Sorrells, 1991; Kisha, Sneller, & Diers, 1997; Manjarrez-Sandoval, Carter, Webb, & Burton, 1997; Burkhamer, Lanning, Martens, Martin, & Talbert, 1998; Bohn, Utz, & Melchinger, 1999; Utz, Bohn, & Melchinger, 2001; Gutierrez et al., 2002; Barroso et al., 2003; Hung et al., 2012). The results of these studies, however, showed little to no statistical relationship between progeny genetic variance and either parental genetic distance or coefficient of ancestry.

More recent studies have reported on the prediction of progeny variance within simulated or virtual biparental populations, using genome-wide prediction methods (Bernardo, 2015; Mohammadi et al., 2015; Lehermeier et al., 2017; Osthushenrich et al., 2017). Each of these studies used inbred traits in the training set to predict the performance of simulated populations, yet none use hybrid traits in the training set. Notably, Bernardo (2015) and Osthushenrich et al. (2017) both suggest that the simulation and prediction method could also be applied to prediction of topcross performance of simulated biparental populations. Topcrossing is a widely accepted method for testing newly developed maize inbreds, as well as the standard method for F1 hybrid production of maize in the U.S. Topcross performance is also relevant to breeding program of other species that also use F1 hybrids for commercial production.

In this study, we use a genomic prediction approach and a training set composed of topcrosses to predict hybrid performance of simulated maize breeding populations from a pool of historically elite North American dent inbreds. For each simulated breeding population, we provide predictions of the progeny mean (μ), genetic variance (V_G), and mean of the predicted 10% highest yielding progeny (i.e. superior progeny, or μ_{sp}). Using these statistics, we identify which parental combinations have the highest potential to produce improved inbreds for hybrid combination with a specific tester. We also show that genetic distance between two parents of a breeding population has little predictive value for genetic variance among their progeny. Accurate predictions of genetic variance for simulated progeny populations will help breeders of maize and other hybrid crops to determine which inbreds are best to use as parents to create new breeding populations.

1.3 Materials and Methods

1.3.1 Phenotypic Data

A total of 285 maize inbreds derived from North American dent germplasm were used in this study. This includes 246 former commercial inbreds with expired Plant Variety Protection certificates, known as ex-PVP inbreds (USDA, 2013a), and 39 historically important public inbreds. Detailed information about the inbreds used in this study can be found in (T. Beckett et al., 2017). All 285 inbreds were testcrossed to PHP02, an ex-PVP inbred developed by Pioneer Hi-Bred in the Iodent heterotic group. PHP02 was chosen due to its relative higher level of combining ability and performance in testcrosses with other ex-PVP inbreds.

The experimental hybrids were grown in seven single-replicate environments in 2014 and 2015. Four environments were at Purdue Universitys Agronomy Center for Research and Education (ACRE) in West Lafayette, IN, and three were at Dow AgroSciences hybrid testing locations in Platteville, WI, Rochelle, IL, and Clinton, WI. Each entry was grown in a two-row plot, 5.3 m long with 76 cm spacing between rows. Experimental entries were randomly assigned to one of three blocks within each environment, with each block 6 ranges deep and 24 plots wide. Five commercial hybrid checks were repeated three times within each block. After assignment of experimental entries and checks, any remaining plots within each block were planted with commercial hybrid filler. A modified augmented design (Lin & Poushinsky, 1983) was used.

Three traits were measured over seven environments: plot weight (PW, in kg); grain moisture at harvest (GM, in %); and test weight (TW, in kg/m^3). The trait PW was converted to grain yield (GY, kg/ha) by normalizing to 15% moisture with the following formula:

$$\mathrm{GY} = \mathrm{PW} \cdot \frac{1231 \text{ plots}}{1 \text{ ha}} \cdot \frac{1 \text{ m}^3}{720.8 \text{ kg}} \cdot \frac{100 - \mathrm{GM}}{100 - 15.5\%} \tag{1.1}$$

The software SAS 9.4 (SAS Institute, Cary, NC) was used to calculate best-linear unbiased estimators (BLUPs) for use in predictions, and for estimating variance components used to calculate trait heritabilities. The phenotypic value of genotype i when grown in block k within environment j is represented by:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{k(j)} + \epsilon_{ijk} \tag{1.2}$$

where μ is the population mean, α_i i is the effect of the ith genotype, β_j is the effect of the jth environment, $\delta_{k(j)}$ is the effect of the kth block within the jth environment, and ϵ_{ijk} is the residual error. The genotypic, environmental, and block effects were all considered random effects. Each experimental entry was represented once within each environment. Therefore, the genotype-by-environment interaction was confounded with the residual error ϵ_{ijk}, and the two terms could not be statistically separated by regular ANOVA methods (Bernardo, 2002). Variance components were estimated for the random effect terms on a line-mean basis (Nyquist & Baker, 1991). Broad-sense heritability was calculated by:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_\epsilon^2}{n}} \tag{1.3}$$

where H^2 is the broad-sense heritability, σ_G^2 is the genotypic variance, σ_ϵ^2 is the error variance, and n is the number of environments.

1.3.2 Genotypic Data

The genotypic data set is identical to that used in (T. Beckett et al., 2017). Briefly, the data came from two different genotype-by-sequencing (GBS) runs, with SNP calls aligned to the B73v2 reference genome. The two GBS data sets were merged using TASSEL 5.0, version 20151210 (Bradbury et al., 2007). Heterozygote calls were changed to missing data. At loci that contained more than two alleles, the primary and secondary alleles were kept, and tertiary and greater alleles were replaced with missing data. Markers with minor allele frequencies less than 0.05 were removed, leaving the set with 77,314 SNP markers. The remaining missing data (6.22%) was fully imputed using the R package NAM (Xavier, Xu, Muir, & Rainey, 2015) in Rstudio version 0.98.1103 (R Core Team, 2015). Imputation accuracy, estimated to be 0.83, was calculated by starting with a subset of genotypes with no missing data, masking 6.22% of the data points, imputing, and then comparing the imputed set with the actual set at only the masked loci. This was repeated 100 times, with the mean of the 100 iterations reported as the imputation accuracy. Original genotypic data for all inbreds is available in the online GBS data repository at www.panzea.org (Zhao et al., 2006).

1.3.3 Genetic Map

Our original genotypic data set was based on physical map coordinates (in bp). To simulate RIL progeny genotypes, however, a genetic map based on recombination frequencies (in centimorgans, or cM) was required. As recombination patterns in maize are ancient and predictable (Rodgers-Melnick et al., 2015), any genetic map that includes inbreds genetically related to the 349 inbreds in our set will be sufficient to model recombination. Thus, we chose the Nested Association Mapping (NAM) collection which involves 26 diverse maize inbreds, including several historically important North American dent maize progenitors (Yu, Holland, McMullen, & Buckler, 2008). This collection was created by crossing 25 inbreds with the inbred B73 and self-pollinating the progeny over six generations, resulting in 200 recombinant inbred lines (RILs) per F1 family for approximately 5,000 RILs overall (McMullen et al., 2009).

The NAM genetic map was obtained from MaizeGDB (http://www.maizegdb.org). A best-fit polynomial equation was chosen to model the data for each chromosome (physical location vs. genetic location), by minimizing the residuals between the predicted value and the actual value given in the NAM genetic map. Using the polynomial equation for each respective chromosome, genetic locations (in cM) were imputed to the physical locations of each of the 77,314 GBS loci. The genome was divided into windows of 1 cM each, and one marker chosen to represent each 1 cM window. This resulted in 1,466 markers in the final genotypic set to be used for simulation and prediction.

1.3.4 Simulation of Breeding Populations

We used the R package PopVar (Tiede, Kumar, Mohammadi, & Smith, 2015; Mohammadi et al., 2015) which uses the R/qtl package (Broman et al., 2003) to simulate the RIL populations, based on the genetic map provided. The method used to simulate the meiotic events is based on Stahls model (Stahl, 1979). The general procedure begins with creation of a generic RIL population, followed by creation of specific RIL populations based on parental genotypes. The simulated individual RIL genotypes are recombinant inbreds self-pollinated to an infinite number of generations. Simulated populations were created from all possible bi-parental crosses using the 285 inbreds.

1.3.5 Statistical Estimation and Validation of Genomic Estimated Breeding Values

The ridge-regression best linear unbiased prediction method (rrBLUP) (Endelman, 2011) was employed to produce an estimate of the genomic estimated breeding values (GEBVs) (Hayes & Goddard, 2001). The linear regression model to estimate marker effects is as follows:

$$Y = \mu + Xg + \epsilon \tag{1.4}$$

where Y is the $N \times 1$ vector of phenotypic means, μ is the overall mean of the training set, X is the $N \times Nm$ marker matrix, g is the $Nm \times 1$ marker effects matrix, and ϵ is the $N \times 1$ vector of residual effects. Cross validation was accomplished by reporting on the accuracy of 60:40 split for prediction of traits.

1.3.6 Prediction of Parameters in Simulated RIL Populations

The breeders equation presented by Falconer, Mackay, and Frankham (1996) can be used to narrow the target for selection of optimal breeding population:

$$R = ir_{PA}\sigma_A \tag{1.5}$$

where R is the response to selection, i is then intensity of selection, r_{PA} is the correlation between phenotypic and breeding values, and σ_A is the additive genotypic standard deviation, or the square root of the additive genetic variance. In context of the usefulness criterion described by (Schnell & Utz, 1976), we can then derive the following:

$$\mu_{sp} = MP + ir_{PA}\sigma_A \qquad (1.6)$$

where μ_{sp} is the mean of the superior progeny, or usefulness criterion, MP is the mid-parent value, and the remaining values represent R, or the response to selection. Therefore, our target when selecting optimal breeding populations is the mean of the superior progeny, or μ_{sp}.

Maximization of the usefulness criterion μ_{sp} depends on the variables on the right side of Eq. 1.6. If the traits for both parents are known, MP is easily calculated. Intensity of selection i can be changed to suit the goal of the breeding scheme. Improvement in r_{PA}, or prediction accuracy, can be achieved by replicating over environments (i.e. each unique combination of location and year) to reduce the impact of genotype-by-environment interaction or by using a training set composed of relatives. If the prediction accuracy is high, then selecting parents that produce progeny with high predicted genotypic variance will result in the maximum usefulness criterion.

Superior progeny (sp) was defined as the top 10% among the GEBVs within each RIL progeny population (see Eq. 1.6). The mean of the superior progeny (μ_{sp})–effectively equivalent to the the usefulness criterion–was then used for ranking potential breeding crosses. All population parameters (μ,V_A,and μ_{sp}), are calculated by taking the mean from 20 iterations of the simulation.

1.3.7 Relationship Between Genetic Distance and Genetic Variance

First, we calculated GD using Tassel 5.0, version 20151210 (Bradbury et al., 2007). The software program Tassel determines GD at a single locus by $1 - p(IBS)$, with $p(IBS)$ defined as the probability that randomly selected alleles at the same locus in two different individuals are the same, or identical by state. To get an overall measure of GD, individual terms are summed over all loci. To examine the potential relationship between GD and the usefulness criterion or V_A, we produced a scatterplot of GD vs. V_A.

1.4 Results

1.4.1 Summary Statistics

Summary statistics for the six traits per environment are provided in Table 1.1. Mean GY was 11,725 kg/ha. The mean GY for the commercial hybrid checks replicated within the trials was 13,307 kg/ha. Mean GM was 19.2% and mean TW was 676.7 kg/m3. Grain yields were higher in environments 3, 4, and 5, likely due to higher fertilizer inputs than at Purdues ACRE, where environments 1, 2, 6, and 7 were grown. Variance components and heritabilities are included in Table 1.2. Heritability was highest for GY, at 0.86, GM was next with 0.64 followed by TW at 0.63.

All trait distributions appeared approximately normal. The strongest correlation between traits, at 0.61, was observed for GY and GM. The other two trait correlations were negative: -0.25 between GY and TW, and -0.32 between GM and TW. All correlations were significant at the $p < 0.001$ level.

1.4.2 Predicted Performance of Simulated Populations

Among the simulated progeny populations, the mean predicted progeny GY ranged from 13,276 kg/ha for LH213/PHR58 to 8,442 kg/ha IBB15/Q381. The additive standard deviation (SD_A, or the square root of the genetic variance V_A) ranged from 523.0 kg/ha for IBB15/PHR58, to 15 kg/ha for B73/F42. The highest mean predicted GY among superior progeny subsets was 13,674 kg/ha for LH213/PHR58. Table 1.3 shows a subset of the highest yielding simulated bi-parental populations (as well as a few other notable biparental populations) ordered by mean progeny GEBV, equivalent to MP. Correlated trait responses represent the mean trait value among the superior progeny selected by GY. While all trait values are shrunken BLUP values, they are sufficiently informative to for purposes of comparison. To highlight the difference between expected performance based on MP and performance indicated by the predicted mean of the superior progeny, a scatterplot of MP vs. μ_{sp} for grain yield is included as Fig. 1.2.

For GM, the maximum mean predicted progeny value was 20.4%, for PHHH9/WIL901, and the minimum was 17.5%, for 779/W117Ht. The largest SD_A was 0.349%, for ICI441/PHN37, and the smallest SD_A was 0.022%, for PHP76/PHV07. The lowest GM among the superior progeny subsets was 17.2%, for 779/W117Ht. A summary of the predicted statistics for a few selected populations is given in Table 1.4.

For TW, the maximum mean predicted progeny value was 692.5 kg/m^3, for ND203/RS710. This particular biparental population also had the greatest mean predicted TW among superior progeny subsets, at 694.1 kg/m^3. Among all populations, the minimum mean predicted TW was 672.5 kg/m^3, for MM402A/ LH215. The largest predicted SD_A was 2.4 kg/m^3 within the progeny population, resulting from MM402A/RS710. Within the MM402A/RS710 population, selecting the superior progeny for TW would result in a selection gain ($\mu - \mu_{sp}$, or $MP - \mu_{sp}$) of only 4.2 kg/m^3. A list of selected simulated breeding populations is provided in Table 1.5..

1.4.3 Genetic Distance and Genetic Variance

The results depicted in Fig. 1.3 show a weak relationship between GD among parents and progeny V_A. The density of the plotted points is indicated by color, with the least dense areas on the plot orange, and the most dense areas red. The Adjusted R^2 between these two variables is 0.16, statistically significant at $P0.001$. The range of possible V_A values greatly increases as the genetic distance between parents becomes larger. It is clear that there is no consistent correlation. Thus, GD between parents is not a good predictor of progeny V_A, especially as GD between parents increases.

1.5 Discussion

A training set of known traits and genotypes, along with a set of simulated progeny genotypes, can be used to predict the GEBV of the progeny, and thus the V_A. We can use the predicted V_A, or more directly the usefulness criterion–defined as the mean of the superior progeny μ_{sp}–to identify which bi-parental populations are likely to produce a set of inbreds that will maximize performance in a topcross with Iodent tester PHP02.

1.5.1 Using Genetic Diversity to Predict Genetic Variance

Our results confirm that GD is not a good predictor of genetic variance. Fig. 1.3 reveals that there is only an association-and not a correlation-between GD and V_A. The relationship between V_A and GD only has an Adjusted R^2 of 0.16. It follows that the accuracy of predicting V_A using GD would only be 0.16. It is clear in Fig. 1.3 that high genome-wide GD for a particular parental combination is not a guarantee of high V_A. Therefore, for the purposes of a breeding selection decision, genome-wide GD is not a reliable predictor of V_A.

1.5.2 Application

Among the top 50 parental combinations ranked by predicted superior progeny grain yield, 41 populations (82%) had parents from different proprietors and only nine populations (9/50 = 18%) had both parents from the same company. The two most common inbreds appearing in the top 50 parental combinations were PHR58, appearing in 23 crosses, and LH213, appearing in 20 crosses. The next most frequent parents were LH214 and 8M129, each appearing in six crosses. The remaining 22 inbreds appeared between one and four times in potential crosses. Thus a small set of inbreds constitute the best performers. Thus an important contribution of this simulation and prediction analysis is to confirm known and/or reveal untested crosses.

Another useful application of this simulation and prediction method is to identify and eliminate "false positives" that are not readily apparent among the possible biparental breeding populations. These false positives are the breeding populations that, after consultation of the MP value and pedigree, would at first appear to be potentially valuable populations. However, when we examine the predictions, we find a smaller V_A and thus a lower μ_{sp}. If grown out and tested, these populations would not be ranked as high in a topcross with PHP02 as MP and pedigree would suggest.

Consider the potential breeding population derived from a cross between two non-stiff stalk (NSS) inbreds, LH216, developed by Holden's Seed, and PHR58, from Pioneer Hi-Bred. LH216, a Lancaster, is descended from LH51 and to a smaller extent LH123Ht, and PHR58, located in the Pioneer Mixed NSS group, descends from a cross between PH383 and PHG16 (USDA, 2015). With MP at 13,035 kg/ha (good enough for 19th-best), but with limited public information about Pioneer Hi-Bred proprietary inbreds PH383 and PHG16, it appears possible that LH216 and PHR58 could create a productive breeding population. The results of the simulation and prediction indicate otherwise. The predicted μ_{sp} is 13,390 kg/ha, coming in at 39th place for μ_{sp}. Thus the ensuing breeding population is predicted to have a narrow variance, leading to a lower μ_{sp} when compared to other potential breeding populations with similar MP.

One population predicted by V_A to outperform its MP-based expectations is LH194/LH213. Both inbreds were developed by Holden's Foundation Seed. LH194 (LH117/LHE137) is located in the B73 sub-group of the SS heterotic group, while LH213 (LH123Ht/LH51) is located in the Mixed sub-group of the NSS heterotic group (USDA, 2015; Beckett et al. 2017). MP places this breeding population at 53rd, with 12,953 kg/ha. A predicted V_A of 93,221 kg2/ha2 and a μ_{sp} of 13,483 kg/ha ranks this cross at 11th place among all potential breeding populations. If the primary goal was to produce an inbred that performs well in a hybrid cross with only PHP02, the predicted progeny V_A and μ_{sp} identify this parental combination as a promising breeding cross. However, commercial breeders also need to keep heterotic group divisions clear in order to preserve complementary alleles and haplotypes for future population development. As LH194 and LH213 are from very different heterotic groups, this cross is would not normally be made by a commercial breeder.

Selection for GY will also affect other traits. For example, the superior progeny for GY of the cross LH213 by PHR58 (see Table 1.3) have a predicted correlated response for GM of 19.5%, which is 0.2% higher than the MP for this cross. This is not an unexpected response, as it is well known that later-maturing corn varieties, generally have both higher grain moisture and higher grain yield values (Daynard et al., 1971). Similarly, for the same subgroup of superior progeny for GY, TW shows a MP value of 679 kg/m3, which is 2 kg/m3 lower than the mean TW for all progeny of LH213 by PHR58, 677 kg/m3. Both of these responses to selection are reasonable in context of the sign and magnitude of the calculated trait correlations.

Grain moisture shows potential for substantial change of up to 0.6%, by applying this model and using predicted V_A to identify suitable populations for selection. It is likely possible to select among progeny to drive a population to an earlier relative maturity. However, caution should be exercised to ensure that the correlation with grain yield is not too high within that population, or the gains in shifting to an earlier maturity will be offset by losses in GY. One approach is to start with the superior progeny for GM from each population (i.e. the subsets of the lowest 10% by GM), then sort the results by highest predicted GY response. By doing so, we identify the population 2FACC/LH213, where the superior progeny by GM is predicted to have a relatively high GY at 13,047 kg/ha. The correlation between GM and GY is 0.33 for this population. Another example is the superior progeny for GM of the cross ICI441/LH213, at 12,982 kg/ha, with a correlation between GM and GY of 0.28. Both of these populations have a selection differential ($\mu - \mu_{sp}$) of -4% GM. Selection within either of these populations has the potential to identify new inbreds that will show high GY performance with an earlier relative maturity. Again, while these two crosses are across heterotic groups and would not normally be made by commercial breeders, they do represent an opportunity to take two populations predicted to be high performers and drive them to an earlier relative maturity.

Given the relatively small predicted variances for TW among simulated biparental populations, direct selection for TW is not likely to produce much gain. Instead, the goal for TW among commercial breeders is to keep it from falling below a certain threshold (generally around 721 to 734 kg/m3) while improving other traits. Therefore, two approaches could be taken. The first is to identify populations with favorable performance by GY that also have a favorable to minimal correlation between TW and GY. Populations with a negative correlation between GY and TW, even if they had a high predicted GY, would be discarded. This practice will ensure that TW is not compromised when other traits are improved. The second approach is to find an inbred that can be a donor for high TW. For example, the inbred RS710 is one of the parents for 77 of the top 100 breeding populations ranked by mean predicted progeny TW. Therefore, RS710 has potential to serve as a donor of high TW.

Overall, given the small predicted variances for TW and general high correlation between GY and GM within this germplasm set, the best general approach within this data set is to use GM as an indicator for maturity groupings, GY as the primary trait for ranking and selection and within each maturity group, and monitor correlated response in TW. Alternatively, as has already been illustrated, different approaches can derive value from these results by providing predictions that inform selections toward a specific breeding goal.

A set of F1 hybrid phenotypic traits, pedigrees, genotype-based clustering, *GD*- or even all four-is not enough to accurately predict V_A and μ_{sp}. Pedigrees and GD provide a solid foundation of germplasm knowledge that should not be disregarded; however, they only provide an incomplete picture of precisely which parents alleles were incorporated into the progeny through the cycles of inbred development. Pedigrees provide a general idea of the historical origin of a portion of the genome of an inbred, but such estimations tell us little about the impact of these particular sequences on the performance of the plant *per se* and, more importantly, in testcross hybrids. Similarly, although GD incorporates a genome-wide marker set, many are non-influential to the trait of interest (Charcosset, Lefort-Buson, & Gallais, 1991; Burkhamer et al., 1998; Hung et al., 2012). The simulation and prediction method we present here provides the ability to leverage the genetic markers that are influential to our trait of interest by using marker effect estimates to predict the value of a simulated bi-parental breeding population. The PopVar approach sorts variations of influential genomic regions in determination of traits of interest. Using an appropriate training set and genetic markers evenly spaced across the genome, we provide predictions of two factors central to breeding population decisions-V_A and the usefulness criterion, or μ_{sp}.

1.5.3 Revisiting Potential Issues

Training Set Composition and Trial Locations

This germplasm set includes inbreds previously protected by a Plant Variety Protection certificate, as well as some historically relevant public inbreds. We did not filter the inbreds by maturity or zone of adaptation. Furthermore, our trial locations are confined to the central Corn Belt. Inbreds and their test crosses in this set that are more adapted to areas unlike our trial locations were being grown as part of testcrosses in non-optimal environments. The experimental design, then, favors those inbreds that are adapted to areas similar to our trial locations. The results on optimal parental combinations for breeding populations are likely most relevant to those inbreds that are well adapted to environments similar to our trial locations. When using predictive models such as this, it is useful to remember that the training set locations should be representative of the target environments for the simulated populations upon which predictions are made.

Evaluating and Improving Prediction Accuracy

Revisiting Eq. 1.5, we can examine in more detail r_{PA}, which represents the correlation between phenotypic and breeding values. This part of the breeders equation can be redefined to reflect the parameters that are subject to control by plant breeders as opposed to animal breeders. In animal breeding, each individual is represented only once. However, in plant breeding, each individual is replicated by multiple genetic clones within the same plot, and plot means are used instead of individual measures. Therefore, we can indicate the use of plot means by changing Eq. 5 to:

$$R = i \cdot \frac{\sigma_A}{(\sigma_P / \sqrt{n})} \cdot \sigma_A \tag{1.7}$$

As plant breeders increase the replication by the use of plot means and numerous environments, n will increase. As n becomes large, the $\sigma_A / \sqrt{n}$ term will converge to zero. Therefore, the correlation between genotype and phenotype, or the prediction accuracy, can in theory be increased to a more desired level. That leaves σ_A as the only variable that cannot be manipulated by the breeder; therefore, a prediction of genetic variance would indeed be very valuable information for selection decisions that affect genetic gain. The precision of the estimate of genetic variance would increase as replicates increase across the target environments. Additionally, the results of this study were based on a relatively small number of trial locations. A greater number of locations, reps and years would likely produce a more robust training set for the predictions. Essentially, an experimental design with a wider scale of testing could have produced different results.

1.5.4 Suggestions for Future Study

Further research to integrate this approach into a plant breeding pipeline could include developing a selection index with weights for each trait. This would enable a breeder to make more precise selections of optimal parental pairs under the parameters of the particular selection index. Such an approach would hold a greater practical value than predictions of the best biparental populations based on only one trait and correlated responses of other traits.

This study could also be expanded by including predictions of performance on additional testers. Additional testers will add a measure of genetic replication, as additive variance is the key target of using testers, and additional hybrid testcross data will allow a better estimate of GCA. Such relevant testers should be chosen from within the opposite heterotic pool that is opposite the inbred populations.

There are additional inbreds whose PVP certificates expire each year. Many of these newly expired-PVP inbreds are highly genetically related to those used in this study. Therefore, it would be a simple extension of this predictive model to use the genotype of a new ex-PVP inbred to predict the same parameters for simulated biparental populations with the new ex-PVP inbred as one of the parents. Additional data from hybrid testcrosses with these newly expired PVP inbreds would ensure that the training model captured the performance of any new additions to the germplasm pool, such as introgression of exotic germplasm from other geographic zones.

The simulation and prediction model presented here is constrained by using only data from one tester. Thus, the predictions can reasonably be applied only to topcross performance under this single tester. The genetic gains made by using these predictions would be realized within only one-half of the resulting hybrid per breeding cycle, as crosses are predicted just for one parental side of a single cross F1 hybrid. If breeders simultaneously improve both parents of a new hybrid, the overall rate of genetic gain in the breeding program would increase.

Kadam et al. (2016) describe a model where a novel population developed from single heterotic crosses in many combinations was used to predict the performance of untested single heterotic crosses, thus providing an accurate predictive model of single-cross performance based on genotype. Current breeding practices generally improve only one side of a heterotic cross at a time by evaluating and selecting among hundreds of experimental entries crossed to several elite testers. Phenotypic data on several hundred randomly chosen heterotic crosses could serve as a training set to predict not only which parental crosses would be best to create new breeding populations, but also which simulated breeding populations would combine well with other simulated breeding populations. Essentially such a breeding scheme would strive for simultaneous improvement of both sides of a heterotic cross. If prediction accuracy was high enough, this approach could increase the rate of genetic gain and speed up the time to commercialization by decreasing reliance upon testcrossing for selection of improved varieties.

1.5.5 Summary

In breeding for hybrid crops, genomic selection can be used to predict the breeding value of an untested but genotyped inbred. This can be extremely valuable information when breeders are looking to make selection and advancement decisions. In our data set, where all potential parents are tested, it is likely that the pairs of parents chosen for breeding populations by traditional methods would be very similar to the pairs of parents identified based on predictions of progeny V_A and μ_{sp}, especially if the breeder making the decisions had extensive knowledge of pedigrees and combining patterns. There are, however, a few notable exceptions, as have been discussed in detail. The two areas then, where this simulation and prediction model can provide the most value is (1) in the case of an untested but genotyped inbred, one that is not found within the training set; and (2) to help eliminate false positive populations, those that appear promising by MP value and pedigree but in fact would have a small V_A and therefore, a relatively lower μ_{sp} than expected. When both MP and progeny V_A are both considered when choosing initial biparental crosses, it is more likely that genetic gain will be maximized.

Table 1.1.
Testcross yield trial summary statistics by environment.

		Environment[a]							
Trait	**Stat.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Overall**
Grain Yield (kg·ha^{-1})	Mean	9992	10789	13393	12155	13860	9437	11405	11725
	SD[b]	1911	1911	2863	2424	2427	1947	2103	2720
	Min	5122	5192	5398	4496	6057	4689	5505	4496
	Max	15387	16024	19717	18282	19364	14101	17289	19717
Grain Moisture (%)	Mean	16.7	17.3	22.0	19.1	25.3	14.5	17.5	19.2
	SD	1.1	2.7	2.8	2.3	2.7	1.6	3.0	4.1
	Min.	15.1	12.8	14.3	13.9	14.2	12.0	13.1	12.0
	Max.	22.7	25.3	30.5	26.4	31.5	20.5	32.3	32.3
Test Weight (kg·m^{-3})	Mean	666.8	665.2	675.7	673.7	654.8	706.2	711	676.7
	SD	21.9	26.4	24.3	16.9	20.1	26.4	27	29.9
	Min	608.2	578.5	610.8	631.2	594.8	576.3	624.9	576.3
	Max	743.7	807.5	764.3	731.4	705.1	766.4	784.7	807.5

[a] Environments 1 and 2 were grown in 2014 West Lafayette, IN, at Purdue Agronomy Center for Research and Education (ACRE) fields 67 and 98, respectively. Environments 3-5 were grown in 2014 at Dow AgroSciences trialing locations, in: Platteville, WI; Rochelle, IL; and Clinton, WI, respectively. Environments 6-7 were grown in 2015 at ACRE fields 68 and 59, respectively.

[b] SD=Standard deviation.

Table 1.2.
Variance components and heritabilities for six traits measured in testcross yield trials used as the phenotypic training set.

	Statistic			
Trait	**N**	$\boldsymbol{\sigma^2_G}$	$\boldsymbol{\sigma^2_\epsilon}$	$\boldsymbol{H^2}$
Grain Yield (kg/ha)	7	37258	40794	0.86
Grain Moisture (%)	7	1.2	4.5	0.64
Test Weight (kg/m^3)	7	7.8	32.0	0.63

N=Number of environments in which the trait was collected; σ^2_G=Genotypic variance; σ^2_ϵ=Residual error variance; and H^2=Broad-sense heritability.

Table 1.3.
Predicted grain yield and correlated trait responses for selected simulated breeding populations in a topcross with Iodent tester PHP02.

Parent1		**Parent2**		**Predicted Grain Yield (kg/ha)**			**Corr. Resp.**	
Name	**Pedigree**[a]	**Name**	**Pedigree**[a]	$\boldsymbol{\mu}^{(\text{Rank})}$	$\boldsymbol{V_A}$	$\boldsymbol{\mu_{sp}}^{(\text{Rank})}$	**GM (%)**	**TW (kg/m^3)**
LH213	LH123Ht/LH51	**PHR58**	PH383/PHG16	13,276$^{(1)}$	825	13,674$^{(1)}$	19.5	678.7
LH213	LH123Ht/LH51	**LH214**	LH123Ht/LH51	13,212$^{(2)}$	74	13,329$^{(63)}$	19.5	675.1
LH214	LH123Ht/LH51	**PHR58**	PH383/PHG16	13,190$^{(3)}$	742	13,569$^{(2)}$	19.6	678.4
2FACC	4676A/PB80	**LH213**	LH123Ht/LH51	13,179$^{(4)}$	791	13,563$^{(3)}$	19.5	677.8
2FACC	4676A/PB80	**PHR58**	PH383/PHG16	13,158$^{(5)}$	837	13,557$^{(4)}$	19.6	681.3
8M129	78060A/88144	**LH213**	LH123Ht/LH51	13,116$^{(6)}$	785	13,504$^{(10)}$	19.8	675.5
ICI441	PHI3377/LH132	**LH213**	LH123Ht/LH51	13,115$^{(7)}$	988	13,545$^{(5)}$	19.9	675.7
8M129	78060A/88144	**PHR58**	PH383/PHG16	13,095$^{(8)}$	942	13,517$^{(7)}$	19.9	678.9
ICI441	PHI3377/LH132	**PHR58**	PH383/PHG16	13,094$^{(T9)}$	921	13,509$^{(8)}$	20.0	679.2
2FACC	4676A/PB80	**LH214**	LH123Ht/LH51	13,094$^{(T9)}$	704	13,457$^{(16)}$	19.7	677.3
LH195	LH117/LH132	**LH213**	LH123Ht/LH51	13,086$^{(11)}$	1079	13,541$^{(6)}$	19.8	676.7
LH216	LH123Ht/LH51^3[b]	**PHR58**	PH383/PHG16	13,036$^{(19)}$	650	13,390$^{(39)}$	19.7	680.3
2FACC	4676A/PB80	**ICI441**	PHI3377/LH132	13,002$^{(29)}$	353	13,257$^{(122)}$	20.1	677.9
LH194	LH117/LHE137	**LH213**	LH123Ht/LH51	12,953$^{(53)}$	93,221	13,483$^{(11)}$	19.8	676.7
B73	BSSS Cycle 5	**F42**	B73 Mutation	12,359$^{(2,677)}$	226	12,385$^{(8,940)}$	19.4	678.7
IBB15	J6/W70884	**PHR58**	PH383/PHG16	10,854$^{(28,044)}$	273,498	11,764$^{(21,487)}$	19.0	681.5
IBB15	J6/W70884	**Q381**	PHI3369	8,442$^{(\text{last})}$	7,048	8,546$^{(\text{last})}$	18.2	678.9

Abbreviated column headings from left to right: $\mu^{(\text{Rank})}$=Predicted progeny mean, with rank in parentheses; V_A=Predicted additive genotypic variance; μ_{sp}=Mean of the superior progeny, or top 10% by grain yield, with rank in parentheses; GM=Correlated response for grain moisture at harvest; TW=Correlated response for test weight.

[a] Pedigrees are from PVP certificates (USDA, 2013b); the pedigree for B73 is from (Gerdes, Tracy, Coors, Geadlemann, & Viney, 1993).

[b] The full pedigree of LH216 is ((LH123Ht/LH51^2)-S2)/LH51, but has been abbreviated within this table for formatting purposes.

Table 1.4.
Predicted grain moisture and correlated trait responses for selected simulated breeding populations in a topcross with Iodent tester PHP02.

Parent1		**Parent2**		**Predicted Grain Moisture (%)**			**Correlated Response**	
Name	**Pedigree**[a]	**Name**	**Pedigree**[a]	$\boldsymbol{\mu}$	$\boldsymbol{V_A}$	$\boldsymbol{\mu_{sp}}$ **(low)**	**GY (kg/ha)**	**TW (kg/m^3)**
PHHH9	PHJ29/PHBT4	**WIL901**	Mo17/Tuxpeño	20.4	0.041	20.1	12,278	679
779	CM-24/W117	**W117Ht**	643/MN13	17.5	0.017	17.2	9,876	683
ICI441	PHI3737/LH132	**PHN37**	CM11/041Ht	19.1	0.122	18.5	11,992	680
PHP76	G50/PHEJ8	**PHV07**	PHG41/G21	18.7	0.001	18.7	10,322	683
2FACC	4676A/PB80	**LH213**	LH123Ht/LH51	19.4	0.032	19.0	13,028	678
ICI441	PHI3737/LH132	**LH213**	LH123Ht/LH51	19.8	0.062	19.4	12,982	676

Abbreviated column headings from left to right: μ=Predicted progeny mean; V_A=Predicted additive genotypic variance; μ_{sp} (low)=Predicted mean of the lowest 10% by grain moisture; GY=Correlated response for grain yield at harvest; TW=Correlated response for test weight.

[a] All pedigrees were obtained from PVP certificates, available at ars.grin.gov (USDA, 2013b).

Table 1.5.
Predicted test weight and correlated trait responses for selected simulated breeding populations in a topcross with Iodent tester PHP02.

Parent1		Parent2		Predicted Test Weight (kg/m^3)			Corr. Resp.	
Name	**Pedigree**[a]	**Name**	**Pedigree**[a]	$\boldsymbol{\mu}^{(\mathrm{Rank})}$	$\boldsymbol{V_A}$	$\boldsymbol{\mu_{sp}}^{(\mathrm{Rank})}$	**GY (kg/ha)**	**GM (%)**
ND203	Haney's MN13	**RS710**	PAG1202/A641	$692.5^{(1)}$	0.83	$694.1^{(1)}$	9,717	17.9
LH162	ND246/Mo17	**RS710**	PAG1202/A641	$691.7^{(3)}$	1.21	$693.7^{(2)}$	10,351	18.1
PHGG7	PHT64/PHG49	**RS710**	PAG1202/A641	$691.8^{(2)}$	1.16	$693.6^{(3)}$	10,173	17.9
A554	Wf9/WD^3	**RS710**	PAG1202/A641	$690.5^{(5)}$	1.38	$692.5^{(\mathrm{T4})}$	10,170	18.1
LH160	ND246/Mo17	**RS710**	PAG1202/A641	$690.3^{(8)}$	1.63	$692.5^{(\mathrm{T4})}$	10,536	18.3
PHBA6	PHZ51/PHG47	**RS710**	PAG1202/A641	$690.0^{(16)}$	1.94	$692.5^{(\mathrm{T4})}$	11,196	18.2
MM402A	LH38/MANS	**RS710**	PAG1202/A641	$683.1^{(5,002)}$	5.77	$687.3^{(1,207)}$	10,955	18.2
2FACC	4676A/PB80	**LH213**	LH123Ht/LH51	$677.7^{(35,730)}$	1.63	$679.9^{(35,337)}$	13,195	19.4
MM402A	LH38/MANS	**LH215**	R177/Mo17C2	$672.5^{(\mathrm{last})}$	1.08	$674.3^{(\mathrm{last})}$	12,002	19.2

Abbreviated column headings from left to right: $\mu^{(\mathrm{Rank})}$=Predicted progeny mean, with rank in parentheses; V_A=Predicted additive genotypic variance; μ_{sp}=Mean of the superior progeny, or top 10% by test weight, with rank in parentheses; GY=Correlated response for grain yield; GM=Correlated response for grain moisture at harvest.

[a] All pedigrees were obtained from PVP certificates (USDA, 2013a), with the exception of ND203 and A554, which were found in Gerdes et al. (1993).

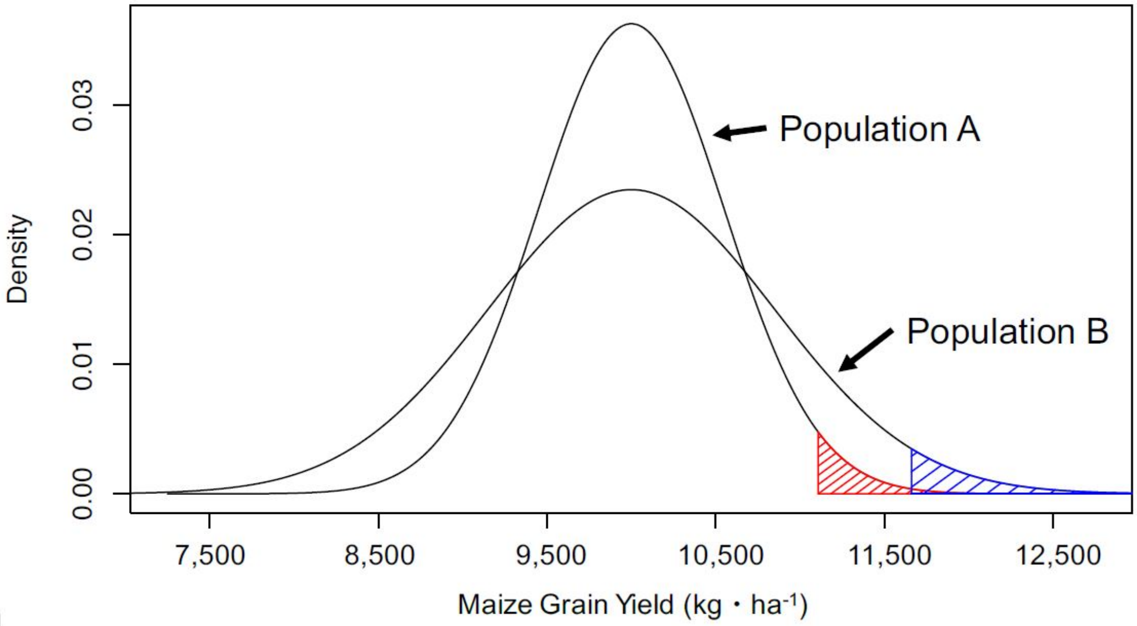

Figure 1.1. Population distribution in context of selection of superior individuals. Here we compare two breeding populations with the same progeny mean of approximately 10,000 kg/ha, but different progeny variances. The red shaded area represents approximately the 10% highest yielding individuals (superior progeny) of Population A, and the blue area represents the same for Population B. Thus, while both breeding populations have the same overall progeny mean, Population B has a greater potential for genetic gain due to greater progeny variance.

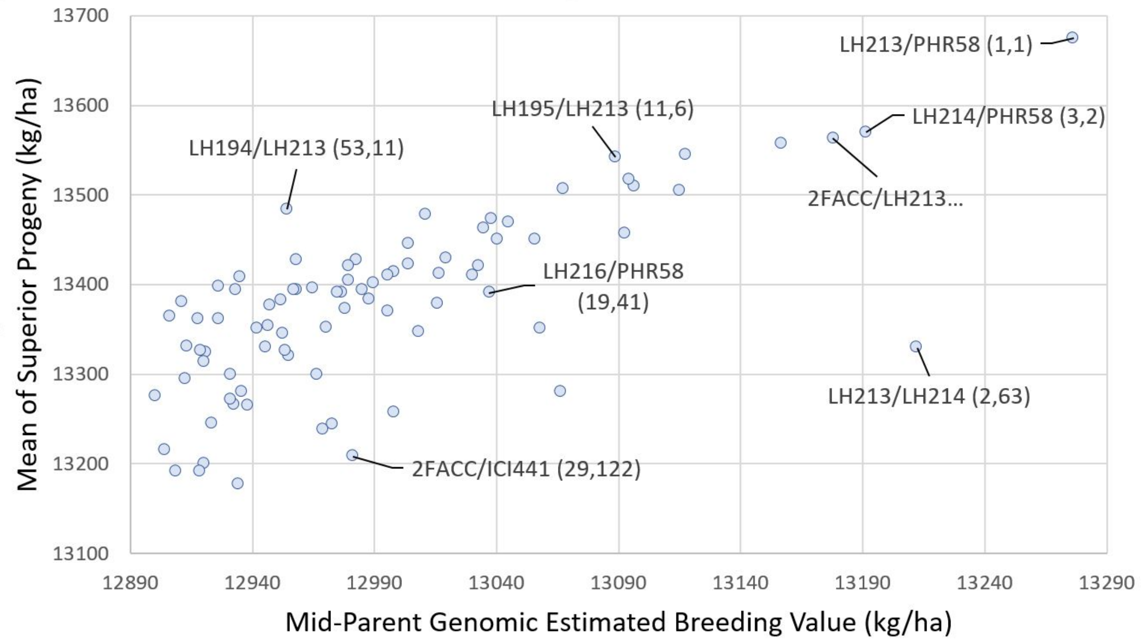

Figure 1.2. Scatterplot of mid-parent value (MP) vs. mean of the superior progeny (μ_{sp}). As selection of breeding populations in a commercial program is limited to the top performers, only potential breeding populations with $MP > 12,900$ kg/ha are shown. A few notable bi-parental combinations are identified, followed by two numbers in parentheses. The first is rank by MP, the second is rank by μ_{sp}. For example, the breeding population with parents LH195/LH213 has the 11th-highest MP, and the 6th-highest μ_{sp}.

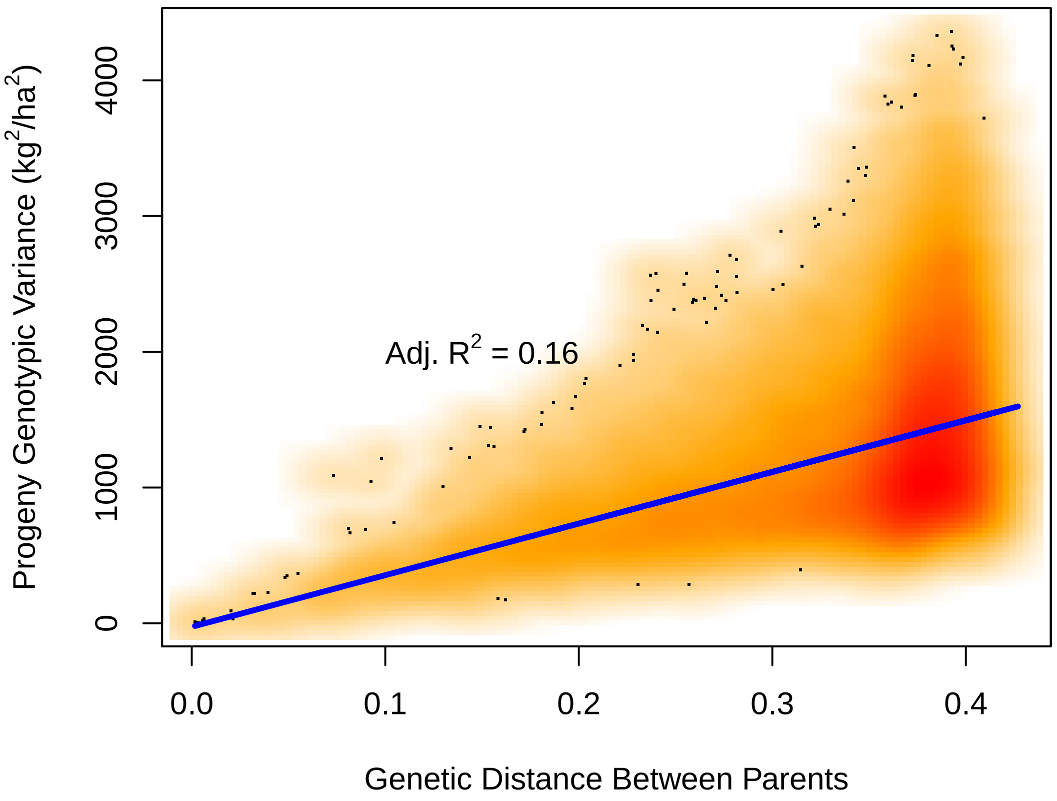

Figure 1.3. Genetic distance between parents vs. progeny genotypic variance. The X-axis shows the genetic distance (calculated by 1-IBS, where IBS=Identity by state) between potential parents of simulated breeding populations. The Y-axis gives the predicted genotypic variance among the progeny in each simulated breeding population. Given the large number of data points, the plot view was adjusted to show density, with the red areas indicating the highest concentration of data points, and the orange areas the lowest. A best-fit linear trendline is shown in blue, with an Adjusted R^2 value of 0.16.

2 SPLIT ’N PREDICT: COMPARING GENOMIC SELECTION MODELS IN EARLY GENERATION MAIZE BREEDING

2.1 Abstract

Genomic prediction (GP) models have recently become an important part of the selection process in both private and public plant breeding programs. GP can increase selection accuracy and decrease the length of the breeding cycle. Here, we use genomic selection in the context of early generation maize inbred development to: (1) model how prediction accuracy changes as the size of the training set changes; and (2) compare the accuracy of different selection models. We previously developed a set of F4 maize lines from a biparental cross between LH51 and PHG35, two formerly Plant Variety Protected (ex-PVP) non-stiff stalk inbreds. These progeny lines were then testcrossed to two ex-PVP stiff-stalk inbreds, PHHB9 and 2FACC. We compared prediction accuracies at graduated training set sizes using four models: ridge regression best linear unbiased prediction (RR-BLUP), BayesB, partial least squares (PLS), and Random Forest (RF). Results suggest that using the RR-BLUP method with 150 individuals in the training set will extract the maximum value out of GP in an early generation maize breeding pipeline by optimizing the balance between trialing cost and prediction accuracy.

2.2 Introduction

Genomic prediction (GP), introduced by Meuwissen et al. (2001), is the process of using known phenotypes and genetic markers spread across the genome (i.e. the training set) to build a regression model that will generate genomic estimated breeding values (GEBV) of lines with known genotypes but unknown phenotypes (i.e. the testing, or validation set) (Leng, Lübberstedt, & Xu, 2017). This method has proven quite valuable to plant breeders, as GEBVs are closely correlated with true breeding values (Heffner, Sorrells, & Jannink, 2009; Heffner, Lorenz, Jannink, & Sorrells, 2010), even in selection models that include multiple correlated traits (Jia & Jannink, 2012). GP models have rapidly become a reliable way for maize breeders to identify superior progeny–and eliminate inferior progeny–within early generation maize breeding based on genotype alone.

In the late 20th century and into the 21st, marker-assisted selection (MAS) was proposed as a cost-effective way to use genetic information to predict and select for phenotypic traits. The MAS approach called for identification of individual marker-trait associations (known as quantitative trait loci, or QTL), then incorporation of these QTL in a statistical model to predict the performance of lines with a known genotype but unknown phenotype. Ultimately, while MAS proved successful at predicting qualitative traits (Flint-Garcia, Darrah, McMullen, & Hibbard, 2003), its usefulness remained limited for prediction of complex quantitative traits like grain yield in early generations of breeding (Stromberg, Dudley, & Rufener, 1994; Jannink, Lorenz, & Iwata, 2010; Bernardo, 2016; Crossa et al., 2017). While MAS uses only a subset of markers related to a particular trait, GP models use all available markers across the genome in an attempt to capture all of the genetic variance for the target trait (Heffner et al., 2010). Primarily due to the increased marker coverage of the genome in recent years, GP has proven superior to MAS in the improvement of traits controlled by many loci (Bernardo & Yu, 2007; Lorenzana & Bernardo, 2009; Lorenz, 2013; Poland & Rutkoski, 2016; Leng et al., 2017). Accordingly, many recent publications have demonstrated the effectiveness in applying GP to prediction of complex hybrid performance traits in maize (Albrecht et al., 2011, 2014; Jacobson, Lian, Zhong, & Bernardo, 2014; Riedelsheimer et al., 2012; Kadam et al., 2016).

Genomic prediction models have generally been applied to plant breeding pipelines in two different ways. First, additive effects in early generations can be predicted in order to speed up the selection cycle within a specified interval of time. The established approach of early generation hybrid testing is to select high performing inbreds based on testcross performance. However, when genomic selection accuracies are at least moderate and the correlation between the genomic-estimated breeding value (GEBV) and the true breeding value is high enough, genomic prediction could reasonably replace or drastically decrease the size one of an early generation testcross trial (Heffner et al., 2009, 2010). In GP applications to early generation yield trials, the *breeding* value or GEBV is the predicted metric, rather than total genetic value of a line. Accordingly, parametric prediction models that deal directly with additive effects–and disregard interaction effects such as epistasis and dominance–are successful. The second way that GP models have been applied to plant breeding pipelines is in the determination of the commercial value of a line. When a prediction method includes both the non-additive effects and the additive effects, the output is a measure of the *total* genetic value of that line. Non-parametric models can ascertain subtle and hidden genotypic correlations and thus predict total genetic value.

There are statistical challenges with using GP associated with the large number of markers. In standard multiple linear regressions of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{\varepsilon} \tag{2.1}$$

the least squares solution is represented by

$$\boldsymbol{B} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \tag{2.2}$$

In order to calculate the least squares estimates of marker effects, the marker matrix $\boldsymbol{X}$ must have a greater number of individuals n than number of variables p, and no columns can be a linear combination of another column. In GP models, the number of markers (p) is much greater than the number of individuals (n). In addition, given the large number of markers in such sets, it is likely that some of the markers are linearly predicted by others with a high degree of accuracy, causing multicollinearity within the matrix. Thus, in data sets considered for GP, the marker matrix $\boldsymbol{X}$ is rank deficient, and the parameters of the model cannot be estimated by normal means.

To solve this problem in GP, regression models require either (i) penalization; (ii) selection of variables; or (iii) reduction of dimensionality (de los Campos et al., 2012; Crossa et al., 2017). Penalization is applied in the ridge-regression best linear unbiased predictor model (RR-BLUP), one of the most commonly used mixed linear models in GP (Endelman, 2011). The RR-BLUP model assumes that all markers have equal variance and that covariance between markers is equal to zero. Thus, the model penalizes, or shrinks the marker effects equally. This reduces the risk of overfitting the model.

One disadvantage of the RR-BLUP model is that by assuming equal marker variances, RR-BLUP can underestimate GEBV in the case of a trait with large-effect loci. To address this, the BayesB model (Meuwissen et al., 2001) includes both penalization and variable selection, making it useful in the case of a trait that falls somewhere within the qualitative to quantitative continuum of genetic inheritance. Essentially, BayesB solves the problem of $p >> n$ in two ways: by estimating parameters from a prior distribution, and by including variable selection by both allowing a unique variance for each marker and allowing some markers to have an effect equal to zero. If there exist large QTL effects for any of the three traits in our study, and if the markers we used are in high LD with the QTL, the BayesB model should perform at a higher level of accuracy in GP.

To allow normal calculation of least-squares estimates, reduction of dimensionality can be applied such as in the model Partial Least Squares (PLS). Falling in the category of supervised learning algorithm, PLS works by regressing the phenotypic response matrix $\boldsymbol{Y}$ not on the marker matrix $\boldsymbol{X}$, but on decomposed scores of the marker matrix (Wehrens & Mevik, 2007). By so doing, the problem of $p >> n$ and rank deficiency of $\boldsymbol{X}$ has been averted, and regression coefficients can be calculated, sufficient for application in GP. It is also important to note that the PLS model determines the latent variable values such that the relationship strength between the latent variables and the response is as strong as possible (Jannink et al., 2010).

Linear models similar to the three discussed predict the GEBV, which is essentially the sum of the additive effects across loci. None of these parametric models take into account genetic mechanisms beyond additivity, such as dominance or epistasis. Non-parametric models, on the other hand, attempt to incorporate non-additive genetic interaction factors that are difficult to explicitly model, thus producing a prediction of the total genetic value. One such model is the machine learning algorithm known as Random Forest (RF) (Liaw & Wiener, 2002). Proposed by Breiman (2001), RF is founded on bootstrap aggregated sampling, or boosting (Breiman, 1996), and works by building a series of regression trees based on the original training set data. Because the trees are built in a progressive manner, the effect of each marker is ascertained in concert with the state of other markers. This allows RF to capture non-additive effects among pairs or groups of loci such as dominance or epistasis. Therefore, if epistatic effects account for a large amount of genetic variation of any of the three traits we measured, RF is expected to perform better at GP.

The objective of this study is two-fold: (1) to model how GP accuracy changes as the size of the training set changes; and (2) to compare the accuracy of different GP models. Within this study, we explore the former application of GWS to the plant breeding pipeline: prediction of early-generation maize hybrid performance. We derived a biparental population from a breeding cross of two non stiff-stalk inbreds with expired Plant Variety Protection (ex-PVP) certificates, LH51 and PHG35. We testcrossed the population progeny to two stiff-stalk testers, PHHB9 and 2FACC, and measured three traits: grain yield, grain moisture, and test weight. We ran genomic selection on various sizes of training sets to predict the performance of a training set of 150 individuals, and measured the prediction accuracy. We report on the change in prediction accuracy as the training set size increases, and suggest an optimal size training population size for efficient incorporation of genomic selection with partial-population testcrossing, in place of whole-population testcrossing in early generation maize breeding.

2.3 Materials and Methods

2.3.1 Inbred Plant Material

The plant material is composed of progeny of a biparental cross of two ex-PVP inbreds, LH51 and PHG35. The parental inbred LH51 was developed by Holden's Foundation Seeds, Inc., and originates from a Mo17 backcross 5 recovery (USDA, 2013a). The public inbred Mo17 descends from a cross between C.I.187-2 and C103 (Gerdes et al., 1993). The other parental inbred PHG35 was created by Pioneer Hi-Bred International, Inc., by a cross of proprietary lines G3BD2 and H7FS6. Both LH51 and PHG35 are members of the non-stiff stalk heterotic group, with LH51 located in the Lancaster-type sub-group and PHG35 in the Pioneer Mixed sub-group (T. Beckett et al., 2017).

2.3.2 Breeding Population Development

The breeding population of 10 F2:3 families used in this study was originally developed for a related project (Morales, 2013) that concluded that genomic selection is superior to phenotypic selection in early generation maize breeding. A short summary of the population development follows. (For more detail, see Morales, (2013).) The progeny from an initial cross between LH51 and PHG35 was used to derive a population of 358 F2:3 families. These 358 families were testcrossed to LH119, PHG39, and an inbred proprietary to Dow AgroSciences. All testcrosses were evaluated in at least six environments. Based primarily on hybrid grain yield, 10 F2:3 families were chosen for advancement. Subsequently, in the summer of 2014, 100 individual F3 seeds representing each of the top 10 F2:3 families was planted at Purdue University's Agronomy Farm for Research and Education (ACRE) in West Lafayette, IN. All plants were self-pollinated; 707 lines successfully produced F4 seed.

2.3.3 Hybrid Testcross Yield Trials

The 707 F4 lines were planted in two isolations at ACRE in 2015 and testcrossed to two stiff-stalk testers, PHHB9 and 2FACC. Both testers are in the B37 sub-group of the stiff-stalk heterotic group (T. Beckett et al., 2017). PHHB9 descends from a cross between PHW52 and PHG86, and 2FACC is derived from a cross between inbreds 4676A and PB80. Out of the 707 LH51/PHG35 F4 lines in the two isolations, 566 produced enough seed for the 2FACC testcross yield trials, and 434 produced enough seed for the PHHB9 testcross yield trials.

Testcross hybrid yield trials were grown in four environments at ACRE–two in 2016 and two in 2017. Three phenotypic traits were collected: grain yield (GY), in kg/ha; grain moisture at harvest (GM), in %; and test weight at harvest (TW), in kg/m^3. Descriptions of these traits is given in Table 2.1. Due to harvest combine equipment failure, no TW data was recorded for environment number 4.

To correct for spatial field variation, the R package 'lme4' v. 1.1-14 (Bates, Mächler, Bolker, & Walker, 2015) in RStudio version 0.98.1103 (R Core Team, 2015) was used fit a linear model with random effects and obtain a best-linear unbiased prediction (BLUP) (Henderson, 1975) for each entry. The phenotypic value of genotype i when grown in environment j, block k, block row l, and block range m is given by:

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \delta_{k(j)} + \gamma_{l(k(j))} + \rho_{m(k(j))} + \epsilon_{ijklm} \tag{2.3}$$

where μ is the population mean, α_i is the ith genotypic effect, β_j is the jth environmental effect, $\delta_k(j)$ is the kth subgroup effect within the jth environment, $\gamma_{l(k(j))}$ is the lth range effect within the kth subgroup and the jth environment, $\rho_{m(k(j))}$ is the mth row effect within the kth subgroup and the jth environment, and $\epsilon_i jk$ is the residual error effect. Genotype, environment, subgroup, range, and row were all considered random variables. The resulting variance components were used to estimate trait heritabilities. Broad-sense heritability on a line-mean basis was calculated by:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_\epsilon^2}{n}} \tag{2.4}$$

where H^2 is the broad-sense heritability, σ_G^2 is the genotypic variance, σ_ϵ^2 is the error, and n is the number of environments. The BLUPs were used instead of raw phenotypes for the genomic prediction analysis.

2.3.4 Genotypic Data

Plant tissue was sampled from F3 plants at ACRE in 2014. DNA extraction and whole-genome-sequencing (WGS) genotyping was accomplished by Dow AgroSciences in Indianapolis, IN, using proprietary protocols. Genotypic markers that had greater than 50% missing data or that were monomorphic for the parents LH51 and PHG35 were removed from the data set. Following filtering and imputation, 3,391 markers remained. Any remaining missing data was then fully imputed by replacing the missing data point with the mean value of the population at that marker.

2.3.5 Genomic Prediction Methods

The genomic estimated breeding value (GEBV) for each individual line was calculated as the sum of the effects of markers across the genome. The general statistical model is given by:

$$y_i = \mu + \sum_k \beta_k x_{ik} + e_i \tag{2.5}$$

where y_i is the breeding value of the ith individual, β_k is the effect on the breeding value of the kth SNP, x_{ik} is the genotype of the ith individual at the kth SNP, and e_i is the residual error of the ith individual.

For tester 2FACC, 24 different levels (sizes) of training sets were used, and 25 were used for tester PHHB9. The number of levels were chosen to best represent a curve that models the change in prediction accuracy as the number of individuals included in the training set increases. Cross validation at each training set level was accomplished by the Monte Carlo or repeated random sub-sampling method (Leberg, 2002; Belkhir, Dawson, & Bonhomme, 2006). Each validation set included 150 randomly selected individuals. The remaining individuals were then randomly assigned to the training set until the set was full, with the number of individuals specified by level. Number of individuals per training set level are listed in Tables 2.4 and 2.5. The training data was then used to fit the model, and predictions were made using the respective model.

The package 'rrBLUP' (Endelman, 2011) was used in RStudio version 0.98.1103 (R Core Team, 2015) for fitting the RR-BLUP model and calculating the associated predictions. The R package 'PLS' (Wehrens & Mevik, 2007) was used to implement the PLS model. The number of dimensions was reduced using the first five principal components (PCA). The R package 'BGLR' (Pérez & de Los Campos, 2014) WAS used to implement the BayesB model. The R package 'ranger' (Wright & Ziegler, 2015) WAS used to implement the RF model. Unless otherwise stated, the default settings were used within each respective function.

2.3.6 Estimating Prediction Accuracy

Prediction accuracy for each GP model was measured as the Pearson correlation between the predicted phenotypic values and the adjusted phenotypic values (BLUPs) within the validation set. For all models, the prediction algorithm was repeated for 500 cycles, with the overall prediction accuracy at each training set level calculated as the mean of prediction accuracies over the 500 cycles.

2.4 Results

2.4.1 Trait Distribution and Correlations

Distributions of phenotypic traits were all approximately normal (Fig. 2.1 and Fig. 2.2). Bivariate phenotypic correlations were weak and limited. For the topcrosses with tester 2FACC, the only statistically significant correlation was between grain yield and test weight, with a value of 0.16 (statistically significant at $p < 0.001$). For the topcross with tester PHHB9, GY and GM had a 0.27 correlation, statistically significant at $p < 0.05$, and GY and TW had a 0.11 correlation, significant at $p < 0.1$.

Among the summary statistics by environment (Table 2.2), environment 4 had higher GY and GM than other environments. While the harvest date for environment 4 was relatively earlier than the other environments, it may not fully explain the observed difference. As was noted earlier, test weight was not collected in environment 4 due to equipment failure. There may be error introduced into the measured GY and GM values due to sub-optimal calibration of the combine measuring equipment.

2.4.2 Variance Components and Heritability

Table 2.3 gives the variance components and heritability values for the three traits for both testers. For the 2FACC testcross group, heritability was 0.69 for GY, 0.58 for GM, and 0.90 for TW. For the PHHB9 testcross group, heritability was 0.61 for GY, 0.58 for GM, and 0.78 for TW.

2.4.3 Prediction Accuracy

Prediction accuracies are provided in Table 2.4 for the 2FACC testcross data set, and Table 2.5 for the PHHB9 testcross data set. Plots of prediction accuracy per number of individuals in training set are included as Figures 2.3, 2.4, and 2.5 for the 2FACC testcross data set, and Figures 2.6, 2.7, and 2.8 for the PHHB9 testcross data set.

For the topcross trial with tester 2FACC, the following methods had the highest prediction accuracy the traits cited: RR-BLUP for GY and GM; and RR-BLUP and BayesB for TW. For the topcross trial with tester PHHB9, the following methods had the highest prediction accuracy for the traits cited: RR-BLUP and RF for GY; RF for GM; and RR-BLUP and BayesB for TW.

For Ridge Regression-BLUP, maximum prediction accuracies were 0.40 for GY, 0.56 for GM, and 0.81 for TW among the 2FACC testcrosses, 0.32 for GY, 0.64

for GM, and 0.60 for TW among the PHHB9 testcrosses. For BayesB, maximum prediction accuracies were 0.40 for GY, 0.55 for GM, and 0.80 for TW among the 2FACC testcrosses, 0.31 for GY, 0.64 for GM, and 0.59 for TW among the PHHB9 testcrosses. For Partial Least Squares, maximum prediction accuracies were 0.39 for GY, 0.54 for GM, and 0.75 for TW among the 2FACC testcrosses, 0.30 for GY, 0.64 for GM, and 0.58 for TW among the PHHB9 testcrosses. For Random Forest, maximum prediction accuracies were 0.39 for GY, 0.53 for GM, and 0.77 for TW among the 2FACC testcrosses, 0.33 for GY, 0.66 for GM, and 0.57 for TW among the PHHB9 testcrosses. For all methods and across all traits, gains in prediction accuracy within the 2FACC and PHHB9 testcross groups appear to be marginal when greater than approximately 150 individuals are included in the training set (see Figures 2.3-2.8).

2.5 Discussion

Trends in Figures 2.3-2.8 show a pattern of marginal improvements in prediction accuracy when greater than 150 individuals are added in the training set. This is more evident in the GP accuracy plots for GM and TW, and less so for the GP accuracy plot for GY. Prediction accuracies within both testcross groups are lower for GY than for GM and TW. Grain yield is a much more complex trait; the relative prediction accuracies are consistent with this.

When comparing GP methods, the parametric model RR-BLUP is among the top performers for all traits across both tester groups. The BayesB method showed a level of performance similar to the RR-BLUP model, with the exception of GY on tester PHHB9. However, BayesB did not exceed the performance of all linear models. This was expected, as the genetic inheritance patterns of GY, GM, and TW are quantitative in nature. Overall, the GP prediction accuracy of the PLS model tended to return lower accuracies when less individuals were included in the training set. As only five PCAs were used to build the predictive model, use of smaller training sets would degrade the relationship between the training set and the testing set, causing them to appear less related than they actually are. RF returned the top prediction accuracy for GY on tester PHHB9 and performed reasonably well on the rest of the trait/tester combinations, with the exception of GY on tester 2FACC. While RF has the added feature of incorporating non-additive genetic interactions, that is not essential in this application, as the primary goal of early-generation GP is to ascertain the inheritable GEBV of each line. RF may, however, offer an advantage over parametric models in late-generation GP of commercial value of elite hybrids by including non-heritable yet important genetic effects such as dominance.

In summary, RR-BLUP appeared to be the most consistent top performer in terms of prediction accuracy. Among the others, however, there was no consistent pattern of relative performance. Therefore, it appears most appropriate to test a variety of models in GP applications, and proceed with the model that returns the highest accuracy in cross-validation tests within the training set for the crop and traits of interest.

Some have suggested that more complex non-linear prediction models will show higher prediction accuracies than simpler linear models when dealing with complex traits (Crossa et al., 2017). That is not the case in these results. One possibility is that the phenotypic measurement error was substantial enough to confound the subtle and hidden genotypic interactions that the non-linear prediction models are optimized to pick up. Reduction of phenotypic measurement error may be achieved by more regular calibration of equipment. Another way to address this is to increase the number of individuals in the training set or the number of replications per entry–both are increases in the overall sample size.

There is a substantial gap between the calculated heritability and the GP accuracy for each trait. Some have referred to this gap as the ”missing heritability” (Manolio et al., 2009; Makowsky et al., 2011). The gap is largest for GY: for the 2FACC testcrosses, the top-performing GP method explains 0.40 of the variation for GY while the heritability is 0.69; and for the PHHB9 testcrosses, the top-performing method explains 0.33 of the variation for GY while heritability is 0.61. For GM, the gap is much smaller, with only 0.02 difference between GP accuracy within the 2FACC testcrosses, and only 0.08 difference for the PHHB9 testcrosses. For TW, the gap is 0.09 for the 2FACC testcrosses and 0.18 for the PHHB9 testcrosses.

The reasons why GP accuracy is so much lower than the heritability for some of these traits merits further examination. Factors that affect prediction accuracy of GP include choice of statistical model, marker density, training population size (n), effective population size (i.e. a measure of genetic diversity), and relationship between training and testing populations (Roorkiwal et al., 2016). Another factor that can affect prediction accuracy is the experimental design of the training set–are the environments diverse and representative of the target environment of the testing population? In this experimental design, the four locations were relatively homogeneous. All four environments–two each in two successive years–were within a 2-mile radius on Purdue's ACRE farm. Ideally, to build a robust training set, the experimental hybrid entries should be grown in several years across several geographically separated locations, with each successive cycle of data added to the overall training set. As long as the target environment remains similar to the environments used in the training set, the predictions should increase in accuracy with each successive batch of data added into the training set.

In summary, these results show that genomic prediction would be most efficient in replacing part of an early generation testcross hybrid yield trial when at least 150 of related individuals are grown in a testcross and used to predict performance of the remainder of the set. This finding agrees with the suggestion by Bernardo and Yu (2007) to have a minimum of 100 to 150 lines in a training set in order to obtain the optimal prediction accuracy. Therefore, we recommend that the RR-BLUP method with a minimum of 150 individuals in the training set will extract the maximum value out of GP in an early generation maize breeding pipeline by optimizing the balance between trialing cost and prediction accuracy. Predictive approaches such as this may also provide value to breeding pipelines for other commercial hybrid crops.

Table 2.1.
Description of phenotypic traits.

Abbrv.	**Trait**	**Description**
GY	Grain Yield (kg/ha)	Grain yield, adjusted to 15.5% moisture.
GM	Grain Moisture (%)	Percent grain moisture at harvest.
TW	Test Weight (kg/m^3)	Weight in kg of 1 m^3 of grain.

Table 2.2.
Testcross yield trial summary statistics by environment.

			Env. no.#				
Tester	**Trait**	**Stat.**	**1**	**2**	**3**	**4**	**Overall**
2FACC	Grain Yield (kg/ha)	Mean	11,347	11,336	11,065	13,497	11,792
		SD	1,460	1,325	2,045	2,114	2,005
		Min	6,876	7,197	6,570	7,592	6,570
		Max	16,741	15,293	17,539	19,657	19,657
	Grain Moisture (%)	Mean	16.8	16.1	18.8	20.3	18.0
		SD	0.3	0.3	0.6	1.6	1.7
		Min	16.1	15.5	16.6	15.9	15.5
		Max	18.2	18.8	21.8	27.0	27.0
	Test Weight (kg/m^3)	Mean	766.7	759.7	753.0	-	760.0
		SD	14.4	15.8	15.2	-	16.1
		Min	712.6	712.6	704.9	-	704.9
		Max	799.0	798.9	793.8	-	798.9
PHHB9	Grain Yield (kg/ha)	Mean	11,699	11,258	11,291	13,417	11,482
		SD	1,790	1,492	2,082	1,777	1,830
		Min	7,182	7,691	7,234	9,851	7,182
		Max	17,084	16,571	18,183	19,048	19,048
	Grain Moisture (%)	Mean	15.3	16.1	17.7	19.4	16.4
		SD	0.5	0.3	0.7	1.5	1.3
		Min	14.2	14.7	16.3	16.3	14.2
		Max	18.6	17.2	21.1	22.2	22.2
	Test Weight (kg/m^3)	Mean	732.0	749.5	757.0	-	745.5
		SD	21.4	16.8	20.7	-	22.2
		Min	671.7	699.7	690.6	-	671.7
		Max	799.6	800.2	798.9	-	800.2

Environments 1 and 2 were grown in 2016, and environments 3 and 4 were grown in 2017.

Table 2.3.
Variance components and heritabilities.

Tester	**Trait**	**N**	$\boldsymbol{\sigma_G^2}$	$\boldsymbol{\sigma_\epsilon^2}$	$\boldsymbol{H^2}$
2FACC	Grain Yield (kg/ha)	4	5.85×10^5	1.04×10^6	0.69
	Grain Moisture (%)	4	0.11	0.32	0.58
	Test Weight (kg/m^3)	4	118	54.0	0.90
PHHB9	Grain Yield (kg/ha)	4	4.87×10^5	1.24×10^6	0.61
	Grain Moisture (%)	4	0.06	0.16	0.58
	Test Weight (kg/m^3)	4	112	130	0.78

N=Number of environments in which the trait was collected; σ_G^2=Genotypic variance; σ_ϵ^2=Residual error variance; and H^2=Broad-sense heritability.

Table 2.4.
Genomic prediction model accuracy for performance traits for F4 breeding lines in a hybrid testcross with 2FACC.

	Grain Yield				**Grain Moisture**				**Test Weight**			
Train.	**RR**	**BB**	**PLS**	**RF**	**RR**	**BB**	**PLS**	**RF**	**RR**	**BB**	**PLS**	**RF**
5	0.09	0.09	0.10	0.09	0.19	0.17	0.20	0.18	0.28	0.26	0.29	0.30
10	0.16	0.15	0.15	0.15	0.28	0.26	0.25	0.25	0.42	0.40	0.42	0.41
15	0.19	0.18	0.17	0.19	0.33	0.30	0.29	0.33	0.48	0.48	0.49	0.49
20	0.21	0.20	0.19	0.20	0.36	0.34	0.32	0.34	0.54	0.54	0.53	0.54
30	0.24	0.23	0.21	0.24	0.40	0.38	0.36	0.39	0.60	0.61	0.60	0.60
40	0.27	0.26	0.22	0.27	0.42	0.41	0.38	0.42	0.65	0.64	0.64	0.64
50	0.28	0.28	0.24	0.29	0.44	0.44	0.40	0.44	0.68	0.67	0.66	0.66
60	0.31	0.28	0.24	0.30	0.46	0.45	0.42	0.42	0.69	0.68	0.68	0.67
70	0.32	0.29	0.26	0.32	0.47	0.46	0.43	0.46	0.71	0.70	0.69	0.69
80	0.32	0.30	0.27	0.33	0.48	0.46	0.44	0.47	0.71	0.71	0.70	0.69
90	0.33	0.31	0.27	0.32	0.49	0.47	0.45	0.47	0.73	0.72	0.70	0.70
100	0.34	0.32	0.28	0.34	0.49	0.48	0.46	0.49	0.73	0.73	0.71	0.71
110	0.34	0.33	0.30	0.34	0.50	0.50	0.47	0.49	0.74	0.73	0.71	0.72
130	0.36	0.34	0.31	0.35	0.50	0.50	0.47	0.49	0.75	0.75	0.72	0.72
150	0.36	0.34	0.32	0.35	0.52	0.51	0.48	0.50	0.76	0.76	0.73	0.73
170	0.37	0.36	0.34	0.36	0.52	0.51	0.49	0.50	0.77	0.77	0.73	0.74
190	0.37	0.36	0.33	0.37	0.52	0.52	0.50	0.51	0.77	0.77	0.73	0.75
210	0.38	0.36	0.35	0.37	0.53	0.53	0.50	0.51	0.78	0.77	0.73	0.75
230	0.38	0.37	0.35	0.38	0.54	0.53	0.51	0.52	0.78	0.78	0.73	0.75
270	0.39	0.38	0.36	0.38	0.54	0.54	0.52	0.52	0.79	0.79	0.74	0.76
310	0.40	0.38	0.37	0.39	0.55	0.54	0.53	0.52	0.80	0.79	0.74	0.76
350	0.40	0.39	0.38	0.39	0.55	0.55	0.53	0.52	0.80	0.80	0.75	0.76
390	0.40	0.39	0.38	0.39	0.55	0.55	0.53	0.53	0.80	0.80	0.75	0.77
416	0.40	0.40	0.39	0.39	0.56	0.55	0.54	0.53	0.81	0.80	0.75	0.77

Train.=Number of individuals in training set.

RR=Ridge Regression Best Linear Unbiased Predictor; BB=BayesB; PLS=Partial Least Squares; RF=Random Forest.

Table 2.5.
Genomic prediction model accuracy for performance traits for F4 breeding lines in a hybrid testcross with PHHB9.

	Grain Yield				**Grain Moisture**				**Test Weight**			
Train.	**RR**	**BB**	**PLS**	**RF**	**RR**	**BB**	**PLS**	**RF**	**RR**	**BB**	**PLS**	**RF**
5	0.10	0.09	0.09	0.08	0.22	0.18	0.20	0.22	0.16	0.15	0.16	0.18
10	0.14	0.12	0.12	0.13	0.34	0.31	0.33	0.32	0.23	0.24	0.26	0.24
15	0.16	0.13	0.13	0.16	0.39	0.39	0.39	0.37	0.29	0.29	0.31	0.29
20	0.18	0.16	0.14	0.19	0.44	0.43	0.42	0.44	0.34	0.34	0.35	0.33
25	0.20	0.18	0.16	0.20	0.48	0.47	0.46	0.47	0.37	0.37	0.38	0.36
30	0.21	0.18	0.16	0.21	0.50	0.50	0.48	0.50	0.39	0.40	0.41	0.39
35	0.23	0.20	0.16	0.22	0.52	0.51	0.49	0.53	0.42	0.42	0.42	0.42
40	0.23	0.20	0.18	0.23	0.53	0.53	0.52	0.55	0.44	0.45	0.44	0.43
50	0.24	0.22	0.18	0.24	0.55	0.54	0.53	0.56	0.47	0.47	0.47	0.46
60	0.25	0.23	0.19	0.25	0.57	0.56	0.54	0.59	0.49	0.48	0.48	0.48
70	0.27	0.24	0.19	0.26	0.58	0.58	0.56	0.60	0.51	0.50	0.50	0.50
80	0.27	0.25	0.22	0.27	0.59	0.58	0.57	0.61	0.52	0.51	0.50	0.51
90	0.27	0.25	0.22	0.28	0.60	0.59	0.58	0.63	0.53	0.53	0.52	0.51
100	0.28	0.26	0.22	0.28	0.61	0.59	0.58	0.63	0.54	0.53	0.52	0.52
110	0.28	0.26	0.22	0.29	0.61	0.60	0.59	0.63	0.54	0.54	0.53	0.52
120	0.29	0.27	0.24	0.29	0.61	0.61	0.60	0.64	0.55	0.54	0.54	0.53
130	0.29	0.27	0.24	0.30	0.62	0.61	0.60	0.64	0.56	0.56	0.54	0.54
145	0.30	0.27	0.25	0.30	0.62	0.61	0.61	0.64	0.56	0.55	0.55	0.54
160	0.30	0.28	0.26	0.30	0.62	0.62	0.61	0.65	0.57	0.56	0.55	0.55
175	0.31	0.28	0.26	0.31	0.63	0.62	0.62	0.65	0.57	0.57	0.56	0.55
190	0.31	0.29	0.26	0.31	0.64	0.63	0.62	0.64	0.58	0.57	0.57	0.56
205	0.31	0.29	0.27	0.31	0.63	0.63	0.62	0.64	0.58	0.58	0.56	0.56
225	0.31	0.30	0.27	0.32	0.63	0.63	0.63	0.66	0.59	0.58	0.57	0.57
245	0.31	0.30	0.28	0.32	0.64	0.63	0.63	0.66	0.59	0.59	0.57	0.57
284	0.32	0.31	0.30	0.33	0.64	0.64	0.64	0.66	0.60	0.59	0.58	0.57

Train.=Number of individuals in training set.

RR=Ridge Regression Best Linear Unbiased Predictor; BB=BayesB; PLS=Partial Least Squares; RF=Random Forest.

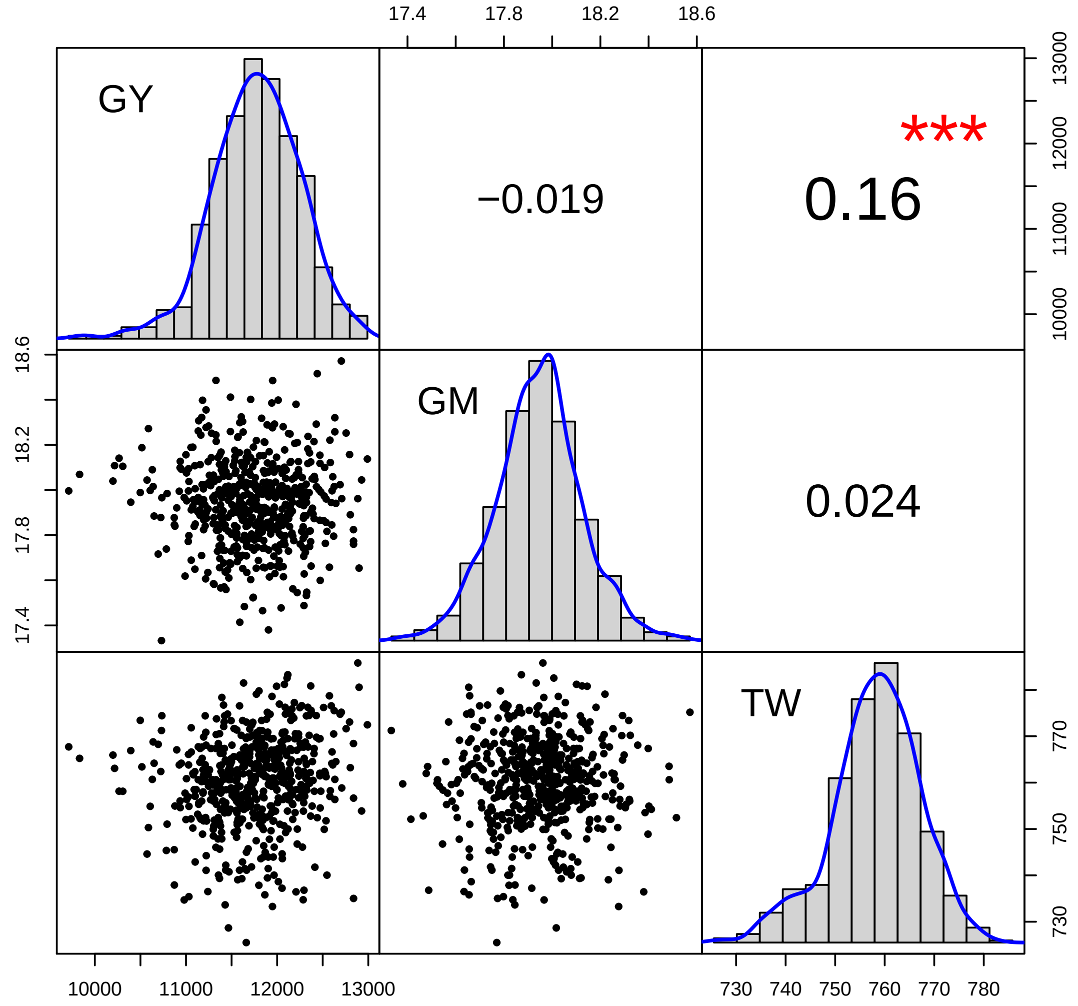

Figure 2.1. Phenotypic trait correlations for LH51/PHG35 F4 progeny in a hybrid yield trial with tester 2FACC.

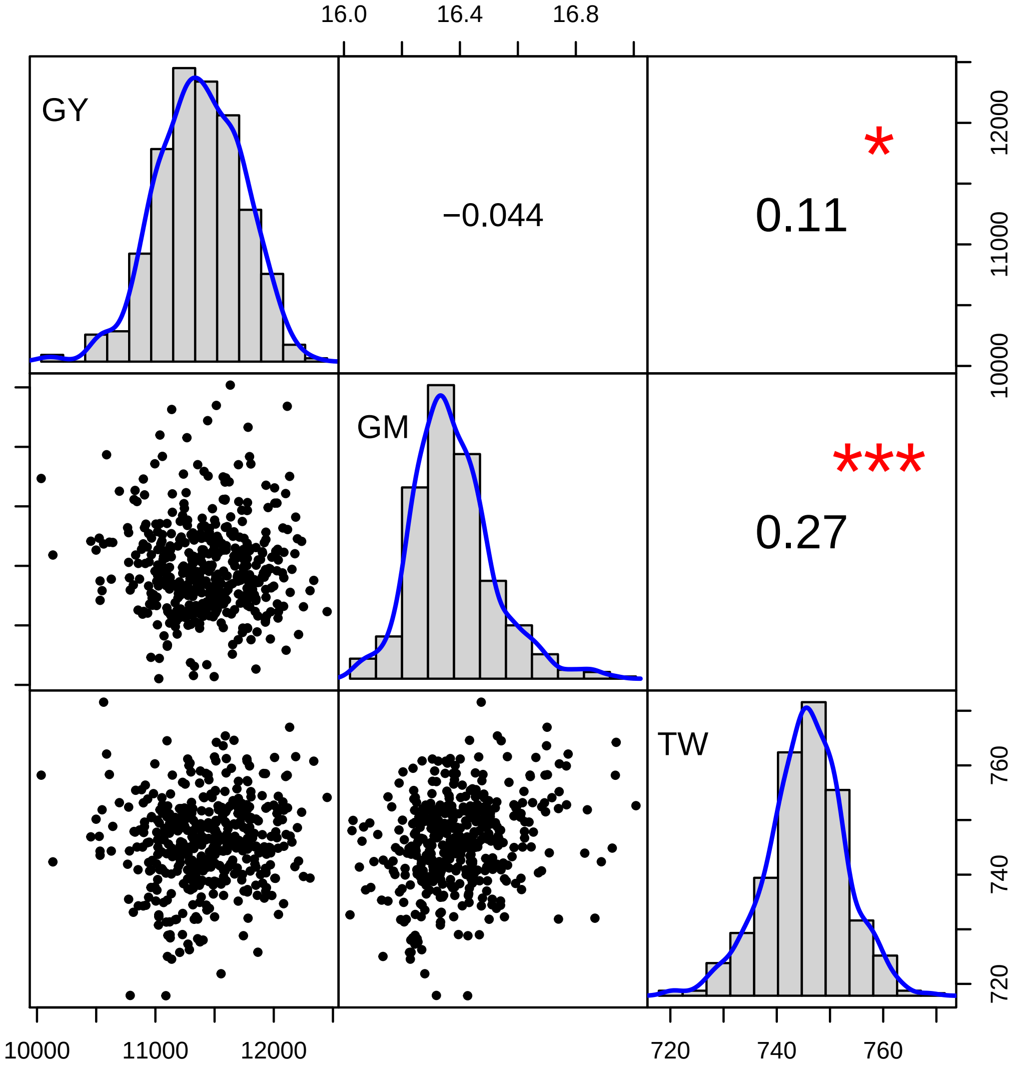

Figure 2.2. Phenotypic trait correlations for LH51/PHG35 F4 progeny in a hybrid yield trial with tester PHHB9.

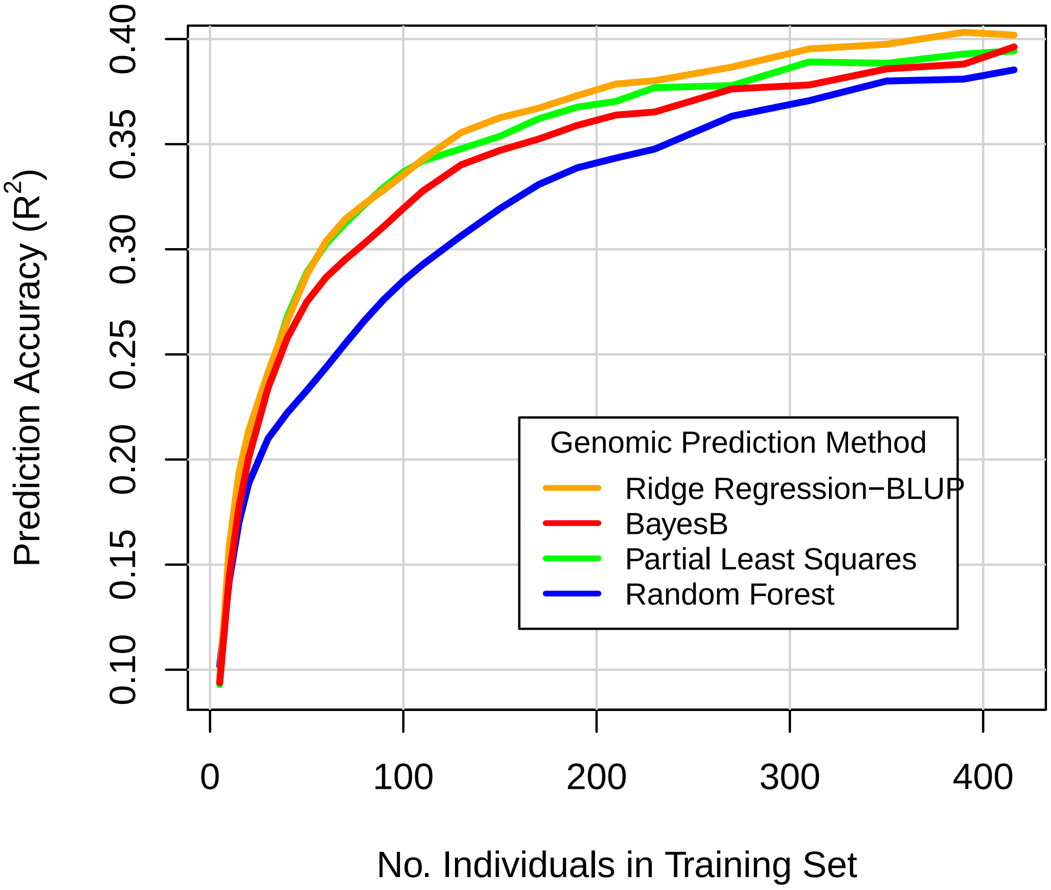

Figure 2.3. Genomic prediction accuracy for hybrid grain yield of LH51/PHG35 F4 progeny in a topcross trial with tester 2FACC.

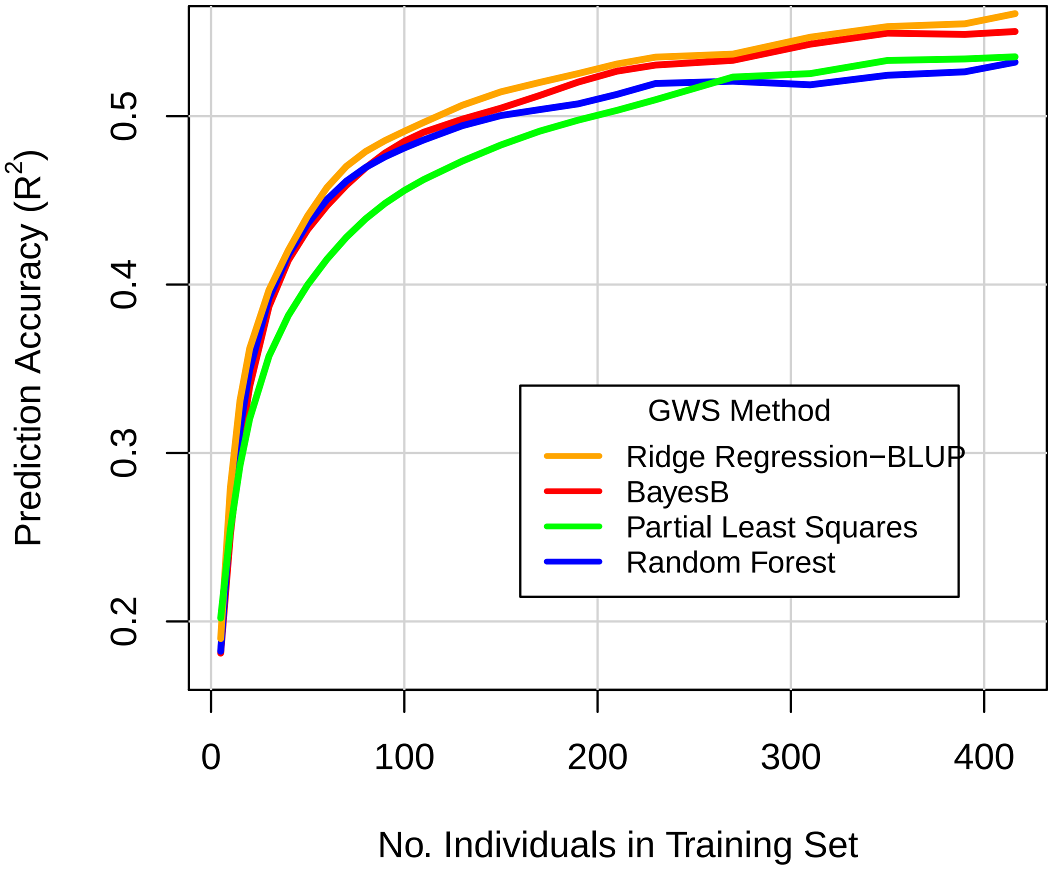

Figure 2.4. Genomic prediction accuracy for hybrid grain moisture of LH51/PHG35 F4 progeny in a topcross trial with tester 2FACC.

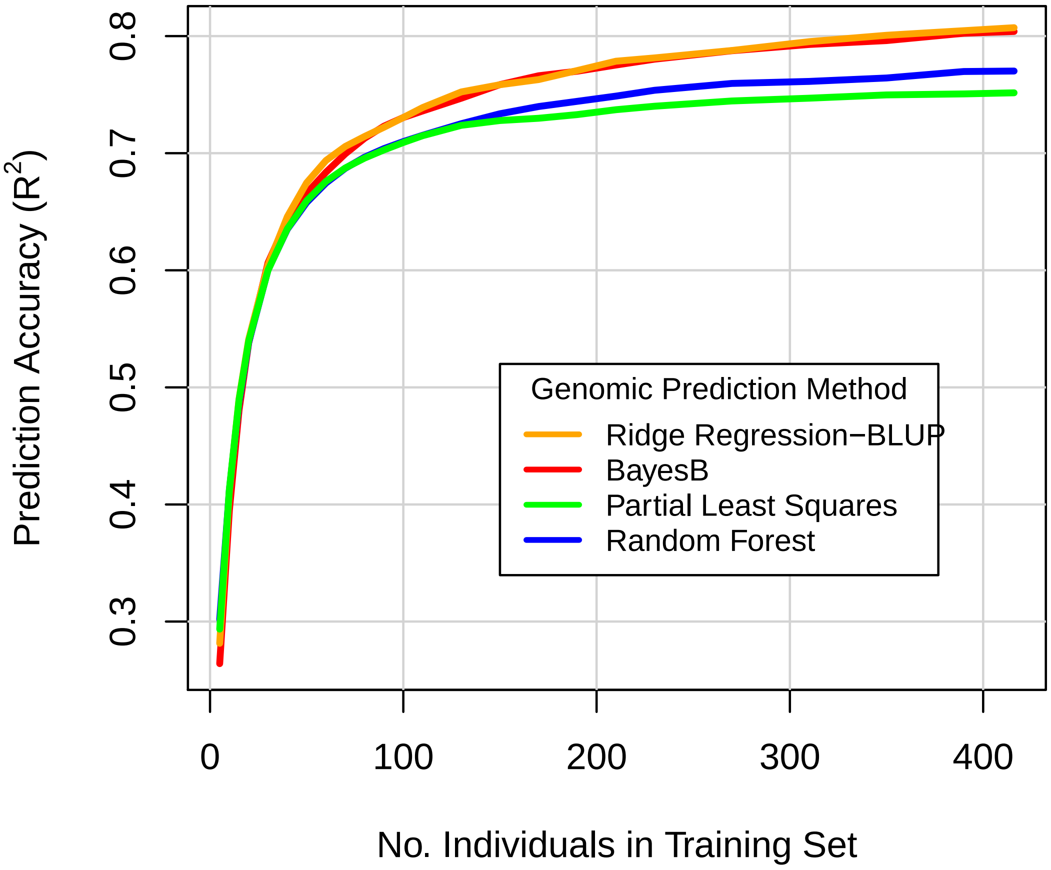

Figure 2.5. Genomic prediction accuracy for hybrid test weight of LH51/PHG35 F4 progeny in a topcross trial with tester 2FACC.

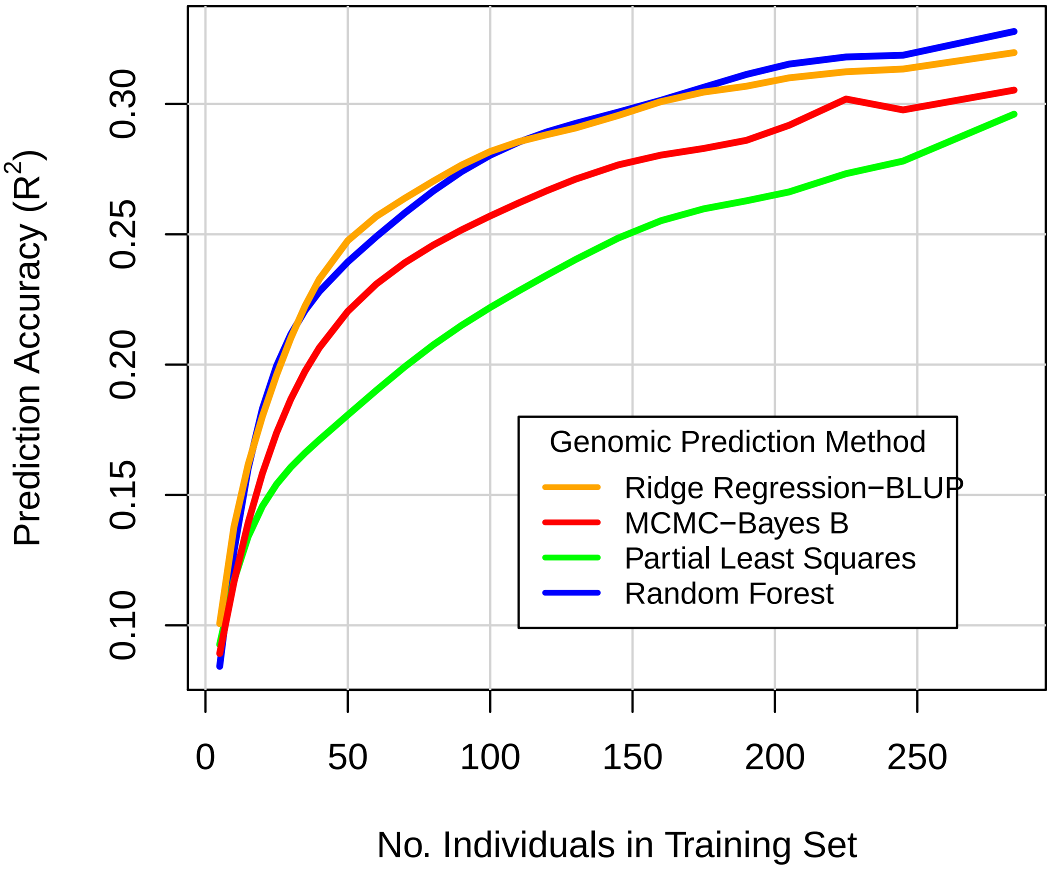

Figure 2.6. Genomic prediction accuracy for hybrid grain yield of LH51/PHG35 F4 progeny in a topcross trial with tester PHHB9.

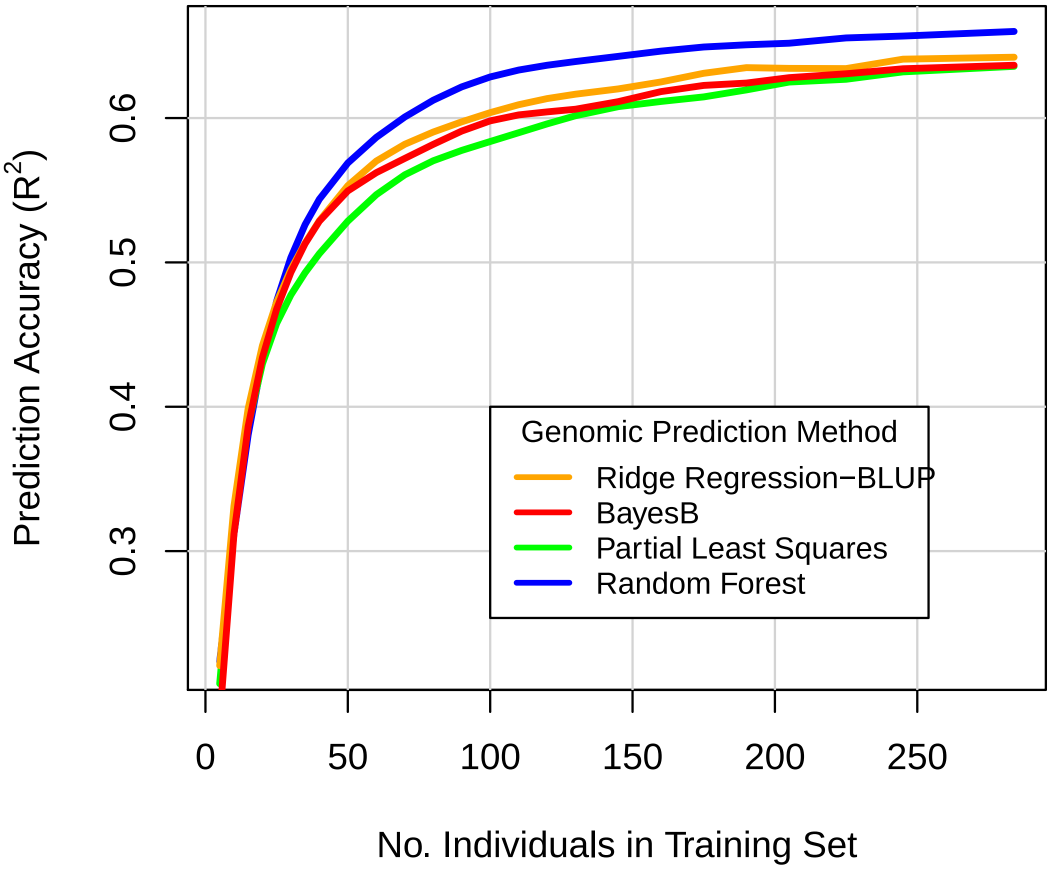

Figure 2.7. Genomic prediction accuracy for hybrid grain moisture of LH51/PHG35 F4 progeny in a topcross trial with tester PHHB9.

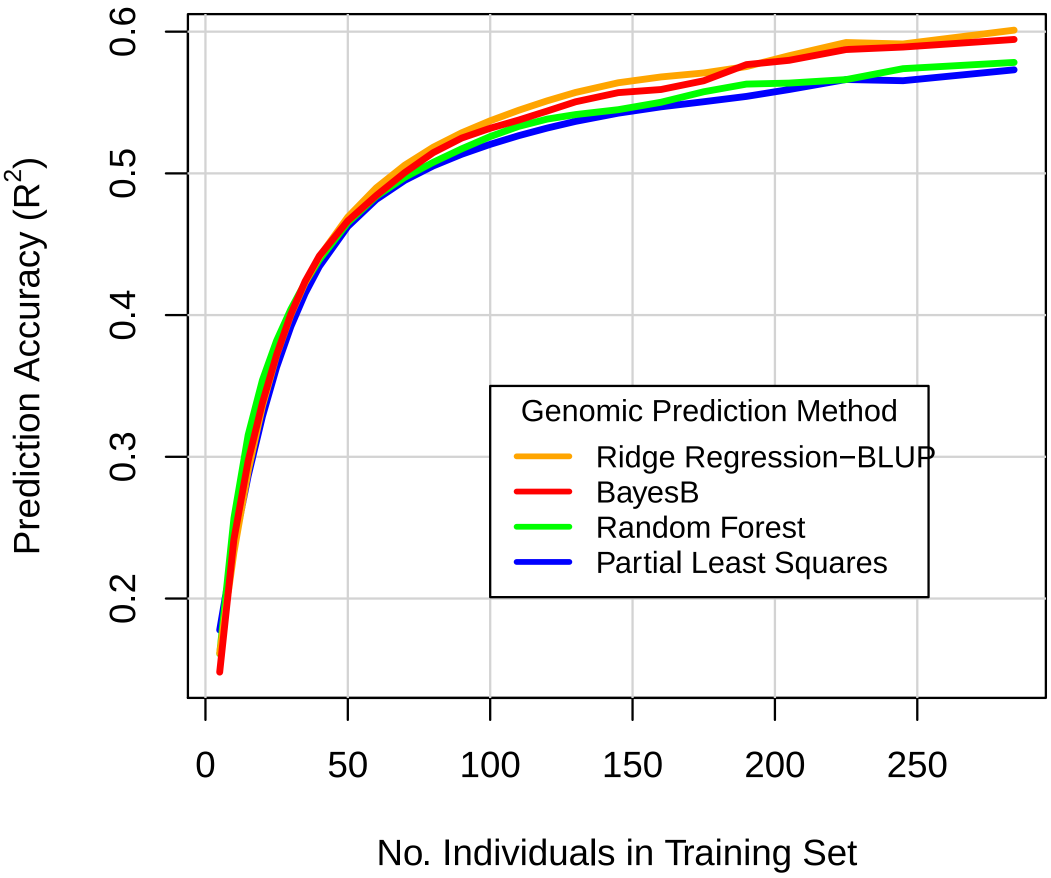

Figure 2.8. Genomic prediction accuracy for hybrid test weight of LH51/PHG35 F4 progeny in a topcross trial with tester PHHB9.

3 GENETIC ANALYSIS OF MAIZE INFLORESCENCE TRAITS IN FORMERLY ELITE COMMERCIAL INBREDS

3.1 Abstract

Inflorescence architecture in maize (*Zea mays* subsp. *mays*) influences seed production and grain yield. Understanding the genetic basis of inflorescence architecture can help breeders better manipulate maize plants to improve seed production and increase grain yield. In this study, we performed a genome-wide association analysis of 349 North American maize inbreds, using 77,329 polymorphic markers produced by genotyping-by-sequencing (GBS). We present three main outcomes: (1) sixty-three quantitative trait loci (QTL) associated with eight inflorescence-related traits; (2) a list of candidate genes for each marker-trait association; and (3) allelic frequency differences at QTL associated with inflorescence traits in North American maize. The results of this study provide a solid foundation for future research to explore applications in marker-assisted selection of inflorescence traits.

3.2 Introduction

The inflorescence structure in maize (*Zea mays* supsp. *mays*) is composed of physically separate and distinct male and female organs. The male inflorescence, found at the apex of the stem, is the pollen-producing tassel, while the female inflorescence, usually located about halfway up the stalk, consists of an ear upon which the seeds are produced (Vollbrecht & Schmidt, 2009). Due to the physical separation of the two inflorescences, a high proportion of pollination events occur between different plants. Maize is thus characterized as a cross-pollinating species.

Filial-1 (F1) hybrid seed production favors larger tassels on plants used as males to maximize number of wind-dispersed pollen grains. This in turn leads to a higher number of successful pollination events, thus maximizing production of seed kernels ((Uribelarrea, Carcova, Otegui, & Westgate, 2002)). Increasing the pollen grain production rate per male plant means more seed-bearing female inbreds can take their place in the seed-production field.

Over time plant breeders have indirectly selected for smaller tassels in hybrid maize, effectively increasing grain yield by redirecting a portion of the plant's photosynthetic assimilates to fill grain on the ear ((Fischer et al., 1987)). Meghji et al. (1984) show a consistent decrease in tassel size among Corn Belt inbreds and hybrids through the 1970s. Duvick et al. (2010) report a similar consistent decrease extending into the early 2000s. Such observations agree with negative associations between tassel size and grain yield (Lambert & Johnson, 1978).

Inflorescence architecture exhibits a wide range of natural phenotypic and allelic diversity (Vollbrecht & Schmidt, 2009), making these traits a promising target for genome-wide association study (GWAS) to discover the underlying quantitative trait loci (QTL). For most traits in maize, including inflorescence, genetic control resides in a large number of genes, each explaining a small part of the observed variation (Brown et al., 2011; Wallace, Larsson, & Buckler, 2014).

Using molecular genotypes and traits collected from a set of formerly elite commercial maize inbreds, we present an association analysis that links genotype with phenotype. Three primary outcomes are discussed: (1) QTL associated with inflorescence architecture; (2) candidate genes near these QTL; and (3) observations of allele frequency differences at tassel trait QTL between North American maize and worldwide maize germplasm. These results can be a starting point for research efforts in developmental and evolutionary biology, genetics and breeding, and maize hybrid seed production.

3.3 Materials and Methods

3.3.1 Plant Material

Three-hundred forty-nine inbreds were used in this study, 283 of which were elite commercial inbreds with expired Plant Variety Protection certificates (also known as ex-PVP inbreds). Sixty-six public inbredskey progenitors of these 283 commercial inbredswere also included. Seed was obtained from the USDA-ARS National Genetic Resources Program (USDA, 2013a). Pedigrees and accession numbers for each inbred are found in Supporting Information 1 and 2.

3.3.2 Experimental Design

All plants were grown in five single-replicated environments at Purdue Agronomy Center for Research and Education (West Lafayette, Indiana), and laid out in a randomized incomplete block design. Two environments were grown in 2014, one in 2015, and two in 2016. Each inbred was represented by no more than 13 plants in a 3.048 m long plot. Rows were planted 0.762 m apart, with a 0.762 m alley. Public maize inbreds B73 and Mo17 were used as replicated checks throughout each environment.

3.3.3 Phenotypic Data

Three plant architecture traits were measured: plant height (PH); ear height (EH); and days to pollen shed (DP). For PH and EH, three random plants per plot were chosen for the measurements. Four weeks after anthesis, three random tassels were selected from each plot and placed in a forced-air dryer for four days. Three tassel phenotypes were measured: tassel weight (TW); tassel branch number (TBN); and tassel spike length (SL). At harvest, three ears from each plot were randomly selected and placed in a forced-air dryer for four days. Two ear phenotypes were measured: cob length (CL); and cob rows (CRW). Trait descriptions are included in Table 1.

3.3.4 Genotypic Data

Genotypic data was obtained from two sources. The first contained 224 inbreds for which PVP certificates had expired up to 2010, as well as 67 public founder inbreds. Genotyping-by sequencing (GBS) data was obtained from www.panzea.org (Zhao et al., 2006). Tissue sampling, DNA extractions, and genotyping-by-sequencing (GBS) were done according to the protocol described by Elshire et al. (2011).

The second genotypic data source was for 58 additional inbreds for which PVP certificates had expired between 2010 and the commencement of this study. Tissue sampling and DNA extraction procedures were performed with adherence to same protocols just cited. Genotypes were obtained from the Cornell University Institute for Genomic Diversity (Ithaca, New York).

The two GBS data sets were merged using Tassel 5.0 version 20151210 (Bradbury et al., 2007). SNPs with a minor allele frequency less than 0.05 were. Genotypes at heterozygous loci and loci with tertiary alleles were changed to missing. The resulting GBS data set contained 77,314 markers with a mean proportion of missing genotype calls per inbred of 0.06 (T. Beckett et al., 2017).

3.3.5 Data Analysis

Statistical Model and Heritability Estimates

The R package 'lme4' v. 1.1-14 (Bates et al., 2015) was used in RStudio version 0.98.1103 (RStudio Team, 2015) to fit a linear model with random effects to obtain variance components for heritability estimates. The phenotypic trait value of genotype i when grown in environment j and subgroup k is given by:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{(k(j))} + \epsilon_{ijk} \tag{3.1}$$

where μ is the population mean, α_i is the effect of the ith genotype, β_j is the effect of the jth environment, $\delta_(k(j))$ is the effect of the kth subgroup within the jth environment, and $\epsilon_i jk$ is the residual error. The genotypic, environmental, and subgroup effects were all treated as random variables.

Broad-sense heritability was calculated on a line-mean basis for each phenotypic trait according to the following formula:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_\epsilon^2}{n}} \tag{3.2}$$

where H^2 is the broad-sense heritability, δ_G^2 is the genotypic variance, σ_ϵ^2 is the residual error variance, and n is the number of environments.

Genome-Wide Association Analysis

Preliminary analyses were performed to determine how to effectively minimize bias due to population structure in the GWAS. A genetic-based cluster was created based on Ward's minimum-variance method (Ward Jr, 1963). Principal component analysis showed that three sub-groups explained 79.4 % of the genotypic variation. These two analyses confirmed that dividing the population into three sub-groups would be optimal.

GWAS was performed using the R package 'GAPIT' (Lipka et al., 2012). Missing genotypic data points were replaced with an intermediate value. Using a mixed linear model, the first three principal components were categorized as covariates to control for population structure. One GWAS was performed for each phenotypic trait in each environment. A simple mean of trait values across all environments was then calculated, and an additional GWAS was run on each trait for this simple mean.

SNPs that fulfilled the following two criteria were identified as QTL: (1) a mean $-log(P.value)$ greater than 3.0; and (2) a greatest single $-log(P.value)$ greater than 4.0. False-discovery simulated GWAS were run to verify the validity of these thresholds for identification of QTL, with 100 iterations. No SNPs from these simulated GWAS runs met either of these criteria. Therefore, these two criteria were considered sufficient to eliminate false-positives from consideration among real results.

Candidate Gene Analysis

Linkage disequilibrium (LD) was estimated using the R package 'NAM' (Xavier et al., 2015), with LD decay reaching a value of $r^2 = 0.2$ at 1.2 kilo base pairs (kbp). Other studies with similar germplasm sets found 1 to 3 kbp (Truntzler et al., 2012) and 10 kbp (Romay et al., 2013). We chose a distance of 10 kbp on either size of the trait-associated SNPs. Candidate genes, as well as orthologues for both rice (*Oryza sativa*) or Arabidopsis (*Arabidopsis thaliana*), were identified using the MaizeGDB genome browser (Sen et al., 2010).

Allele Frequencies at Tassel Trait QTL

We compared allele frequencies at tassel trait QTL within these North American inbreds with allele frequencies in a set of globally sourced inbreds maintained at the USDA-ARS North Central Regional Plant Introduction Station (NCRPIS) in Ames, Iowa as of 2010 (Romay et al., 2013). This population includes North American dent and flint germplasm, tropical inbreds from CIMMYT (The International Maize and Wheat Improvement Center), semi-exotic inbreds from the Germplasm Enhancement of Maize (GEM) program, as well as inbreds from Spain, France, China, Argentina, Canada and other countries.

Tassel 5.0, version 20151210 was used for the direct allele frequency comparisons. The worldwide population included 274 ex-PVP and public inbreds that were also used in this study. Prior to generating allele frequencies for comparison, these 274 inbreds were removed from the worldwide population data set.

3.4 Results

3.4.1 Summary Statistics

Phenotypic distributions, bivariate scatter plots, and pairwise correlation statistics of averages across environments are shown in Figure 1. Among inflorescence traits, TSL and CL had a correlation of $r = 0 : 41$, TBN and TSL were at $r = 0.39$, and TW and TBN, were at $r = 0.62$. PH and EH had a correlation of $r = 0.77$, DP and PH were at $r = 0.56$, and DP and EH were at $r = 0.50$. All of the above mentioned correlations were significant at $p < 0.0001$. Summary statistics per environment are provided in Supporting Information 3.

Phenotypic variance components and broad-sense heritability are listed in Table 2. TBN had the highest heritability at 0.97, CL had the lowest heritability, at 0.86, and the mean heritability for all traits was 0.91.

3.4.2 Genome-Wide Association and Candidate Gene Analyses

Manhattan plots with the -log(P.value) for each SNP-trait association are given in Fig 2. Twenty QTL were found for DP; 12 for PH; 6 for EH; 10 for TW; 8 for TBN; 1 for SL; 4 for CL; 2 for CRW. TW and TBN had two QTL in common; DP and PH had one; and PH and EH had two.

Selected QTL and candidate genes are presented in Table 3. Four tassel trait loci (qTW2, qTW9, qTBN7, and qTBN8) are located near QTL previously identified by (Wu et al., 2016). The TW locus with the highest statistical value, qTW7, was also associated with TBN (qTBN4). Two DP loci (qDP8 and qDP17) were previously identified by (Li et al., 2016). Locus qDP17 was also identified by(Bouchet et al., 2016). All QTL identified in this study are provided in Supporting Information 4. All candidate genes are listed in Supporting Information 5.

3.4.3 Allele Frequencies at Tassel Trait QTL

Frequencies of alleles that increased tassel size and weight were consistently lower (mean difference of -0.12) in the North American commercial germplasm (NA) when compared to the worldwide population (see Table 4). Locus qTW8, had completely different alleles in the worldwide population (C major and T minor) than in the NA commercial population (G major and A minor). For qTW2; qTW6; and qTBN4, alleles that increased TW or TBN were not found in the NA commercial germplasm SS heterotic group. The largest difference in allele frequency (0.28) was found at qTBN7.

3.5 Discussion

3.5.1 Phenotypic Correlations

There was a moderate correlation between TW and TBN. When a tassel has more branches, the additional plant matter means a greater dry weight. There was a negative correlation of $r = 0.39$ between TBN and TSL. When more branches fill up the tassel and spike zone, the region without tassel branches becomes smaller. The moderate correlations observed between DP and EH and DP and PH were expected. Troyer and Larkins (1985) note a strong association between plant height and flowering time. As internodes cease to form following floral initiation, earlier-flowering maize inbreds will be shorter than later-flowering maize inbreds.

3.5.2 Heritability

The environments were all at the same general locationPurdue's ACRE site. Thus, non-genetic variance was minimized due to environmental homogeneity. This could result in artificially high heritabilities. However, a similar study of maize inflorescence traits by Brown et al. (2011), across 8 diverse locations across the US mainland and the territory of Puerto Rico, found similar levels of heritabilities (ranging from the high 80s to the low 90s). The existence of similarly high heritabilities in both Brown et al. (2011) as well as our study indicates that these traits are in reality highly heritable.

3.5.3 Notable QTL and Candidate Genes

Two primary results support the validity of the QTL found in this study: (1) several QTL are found in common across correlated traits (Supporting Information 5); and (2) a large number of QTL were found within 10 kbp of QTL identified in previous studies, including several at the exact same SNP (Supporting Information 6, 7, and 8).

Tassel Traits

The QTL with the highest -log(P.value) (qTW7) is also a QTL for TBN (qTBN4). This locus is approximately halfway between two tassel architecture candidate genes, 19 Mb downstream from *BIF4* and 17 Mb upstream from *tsh1*. *BIF4* is an auxin-signaling module that regulates maize inflorescence (Galli et al., 2015). *tsh1* is involved in development of the inflorescence leaf, or bract (Whipple et al., 2010). Depending on the extent of LD as well as the density of marker coverage in this region, it is possible that the emergence of this QTL is caused by the presence of both genes in this region. Another locus that is a QTL for multiple tassel traits was identified as both qTW10 and qTBN8. However, there are no reasonable candidate genes within 10 kbp on either side of this locus. For a list of tassel-trait QTL that were found in the same region as previously reported QTL (Wu et al., 2016), see Supporting Information 6.

Days to Pollen Shed

First, one of the candidate genes for qDP17, *pebp8* (phosphatidylethanolamine-binding protein8) is involved in origen activity, and promotes flowering at short days (Navarro et al., 2017). Bouchet et al. (2016) also found a flowering-time QTL only 54 bp away. Second, located 593 bp upstream of qPH9, the gene *MADS50* is contributes to the transition to flowering in Arabidopsis and rice (Sun et al., 2012). Third, the gene associated with qDP12, known as *ras11B2*, contributes to tapetal programmed cell death and pollen development in rice (Ko et al., 2014). Fourth, other DP QTL identified in our study are within 4,000 kb of flowering time QTL previously reported by (Li et al., 2016). A full list is included hereafter as Supporting Information 7.

Connecting Plant Height, Ear Height, and Days to Pollen Shed

Comparison of PH and EH results in this study with an earlier study yields some valuable insights. Among the QTL reported by Peiffer et al. (2014), five plant height-related QTL were within 6,000 kbp of QTL found in this study (see Supporting Information 8). The SNP associated with locus qDP8 was also identified as a QTL for PH in Peiffer et al. (2014). Previous research cites a strong association between plant height and flowering time (see Section 4.1); our data show the same association.

3.5.4 Novel QTL

While other GWAS on inflorescence traits used worldwide germplasm, this study used former commercial inbreds and public inbred progenitors. Novel QTL (i.e. those not found in previous studies) may have resulted from intense selection pressure applied to create commercial inbreds. Due to the homogeneous composition of the environments in this study, these novel QTL may also be specific to these five environments only. If these experiments were repeated in different environments, it is possible that different QTL would be identified. The set of commercial inbreds we used also has a wide range of maturities. By growing all the inbreds in the same zone, it creates bias, as the genotypes will be expressed differently than if appropriate subsets were grown in their respective native regions.

3.5.5 Allele Frequencies at Tassel Trait QTL

Given the consistent decrease in tassel weight in North American commercial maize since the 1930s (see Introduction), it follows that alleles that increase tassel size should be less prevalent among NA commercial maize than in a worldwide maize population. Across all tassel-trait QTL identified in this study, the alleles associated with larger tassel size (either TBN or TW) were 12% less frequent in the NA commercial population than in the worldwide population (Table 3). For two QTL associated with both TBN and TW, alleles that increased tassel size were 18% and 19% less common in the NA commercial population than in the worldwide population (Table 3). The data show a consistent trend that alleles conferring larger tassels are less common within North American maize commercial germplasm than in global maize germplasm.

3.5.6 Application to Breeding

One way to use these results to improve seed production traits is to employ a marker-assisted breeding scheme to select for inflorescence traits in the desired direction. Another approach is genomic selection. For example, a breeder could simulate progeny genotypes from potential bi-parental breeding populations, then use the data in this study as a training population to predict which parents would create the best-performing breeding population (Bernardo, 2014).

Determining gene action at these loci would be very valuable for breeding. For example, suppose that a breeder desired large-tasseled males and small-tasseled females for seed production purposes, but a small tassel in the ensuing hybrid. If a small-tassel allele is dominant over a large tassel allele, then successful divergent selection on that locus would produce a large-tasseled male, a small-tasseled female, and a small-tasseled grain-producing hybrid. Such physiological remodeling of plant architecture for improved partitioning efficiency would be ideal for both maximum seed production and improved grain yield.

3.5.7 Summary

This study identifies a large number of QTL and candidate genes associated with inflorescence traits in maize. We hope these results will serve as a foundation for further work to validate these QTL, leading to characterization of novel genes that contribute to control of maize inflorescence development. Obtaining a better understanding of the genetic architecture of maize inflorescences should enable geneticists, breeders, physiologists, and others to work together to design future inbreds and hybrids with more efficient seed production and optimal dry-matter partitioning.

Table 3.1.
Traits collected.

Trait	**Abbv.**	Description	Units
Tassel Weight	TW	Weight of dry tassel.	g
Tassel Branch Number	TBN	Number of total tassels.	count
Tassel Spike Length	SL	Dist. b/n main rachis and top 1° branch.	mm
Cob Length	CL	Length of the cob.	mm
Days to Pollen Shed	DP	Accumulated growing degree days between planting and 50% pollen shed.	AGDD
Plant Height	PH	Distance from ground to ligule of flag leaf	cm
Ear Height	EH	Distance from ground to ear node	cm

Table 3.2.
Phenotypic variance components and heritability.

Trait	**n**	$\boldsymbol{\sigma^2_G}$	$\boldsymbol{\sigma^2_\epsilon}$	$\boldsymbol{H^2}$
Tassel Weight (g)	5	1.07	0.33	0.94
Tassel Branch Number (count)	5	12.8	2.4	0.97
Tassel Spike Length (cm)	5	1012	536	0.90
Cob Length (mm)	4	258	194	0.84
Cob Rows (count)	4	2.2	1.4	0.86
Days to Pollen Shed (GDD)	3	8616	2815	0.90
Plant Height (cm)	4	403	105	0.94
Ear Height (cm)	4	159	65	0.91

n=Number of unreplicated environments in which the trait was collected; σ^2_G=Genotypic variance; σ^2_ϵ=Error variance; and H^2=Broad-sense heritability, calculated by $H^2 = \sigma^2_G/[\sigma^2_G + (\sigma^2_\epsilon/n)]$.

Table 3.3.
Selected QTL and Candidate Genes.

Trait	QTL	SNP[a]	P-val.[b]	Eff.[c]	Candidate Gene	Gene Product Description
Tassel Weight (g)						
	qTW2	S1_278131948	4.11	0.47	GRMZM2G328500	UDP-glucose 6-dehydrogenase
	qTW6	S4_239413702	4.43	0.46	GRMZM2G073571	Phosphatidylinositol transfer protein
					GRMZM2G073731	-
					GRMZM2G374074	DUF1645 domain containing protein
	qTW7	S6_149473281	7.30	0.57	GRMZM2G106140	Sec23/Sec24, trunk domain protein
					GRMZM2G106190	Ferredoxin-6, chloroplastic
					GRMZM2G106218	T-snare
	qTW8	S7_156740065	4.27	0.32	GRMZM2G153438	Equilibrative nucleoside transporter
	qTW9	S8_153860376	4.39	0.27	GRMZM2G059590	DUF292 domain containing protein
	qTW10	S9_105192237	5.76	0.45	GRMZM2G029912	G11 protein
Tassel Br. No. (count)						
	qTBN7	S8_145795246	4.12	0.97	GRMZM2G162347	CTD-phosphatase
	qTBN8	S9_105192237	5.75	1.51	GRMZM2G029912	Gl1 protein
Tassel Spike Length (cm)						
	qTSL1	S3_2583127	5.34	11.15	GRMZM2G013045	Disulfide oxidoreductase/monooxygenase
Days to Pollen Shed (GDD)						
	qDP8	S3_159555813	4.84	27.55	AC188753.3_FG004	Cons. gene of unknown function
	qDP17	S8_123506087	5.26	28.36	GRMZM2G179264	ZCN8
					GRMZM2G179274	6b-interacting protein 1
					GRMZM2G479987	Cons. gene of unknown function
Plant Height (cm)						
	qPH2	S1_211673059	4.60	7.53	GRMZM2G047019	CCR4-NOT transcr. complex subunit 8
					GRMZM2G047238	Stromal cell-derived factor 2
	qPH5	S1_273786380	4.65	9.15	GRMZM2G131525	Knolle protein
					GRMZM2G131575	ATP synthase
Ear Height (cm)						
	qEH2	S1_211673059	4.61	4.52	GRMZM2G047019	CCR4-NOT transcr. complex subunit 8
					GRMZM2G047238	Stromal cell-derived factor 2
	qEH3	S1_273786380	4.82	5.48	GRMZM2G131525	Knolle protein
					GRMZM2G131575	ATP synthase

[a] SNP=Chromosome number and physical position, in bp.

[b] P.val.=$-log_{10}$(P-value)

[c] Eff.=Effect of the major allele.

Table 3.4.
Allele frequency comparisons for tassel trait QTL.

Trait	QTL	SNP	Allele	Global	PVP	Diff.[a]
Tassel Branch Number						
	qTBN1	S1_219371053	A	0.18	0.18	0
	qTBN2	S3_194048392	G	0.68	0.49	0.19
	qTBN3	S5_50335549	A	0.28	0.35	-0.07
	qTBN4	S6_149473281[b]	A	0.25	0.07	0.18
	qTBN5	S6_162140459	G	0.37	0.16	0.21
	qTBN7	S8_145795246	G	0.63	0.35	0.28
	qTBN8	S9_105192237[b]	T	0.41	0.22	0.19
Tassel Weight (g)						
	qTW1	S1_14863005	C	0.39	0.35	0.04
	qTW2	S1_278131948	T	0.24	0.10	0.14
	qTW3	S3_208616512	G	0.41	0.26	0.15
	qTW4	S3_217293300	C	0.58	0.39	0.19
	qTW5	S4_236395269	C	0.20	0.22	-0.02
	qTW6	S4_239413702	C	0.10	0.08	0.02
	qTW8	S7_156740065[c]	A	-	0.27	-
	qTW9	S8_153860376	C	0.52	0.26	0.26
		Overall Mean	**0.37**	**0.25**	**0.12**	

Global=Global inbred population; PVP=North American formerly elite commercial inbred population.

[a] Difference in allele frequency between the Global and PVP.

[b] QTL found for both TBN and TW

[c] Allele not found in global population

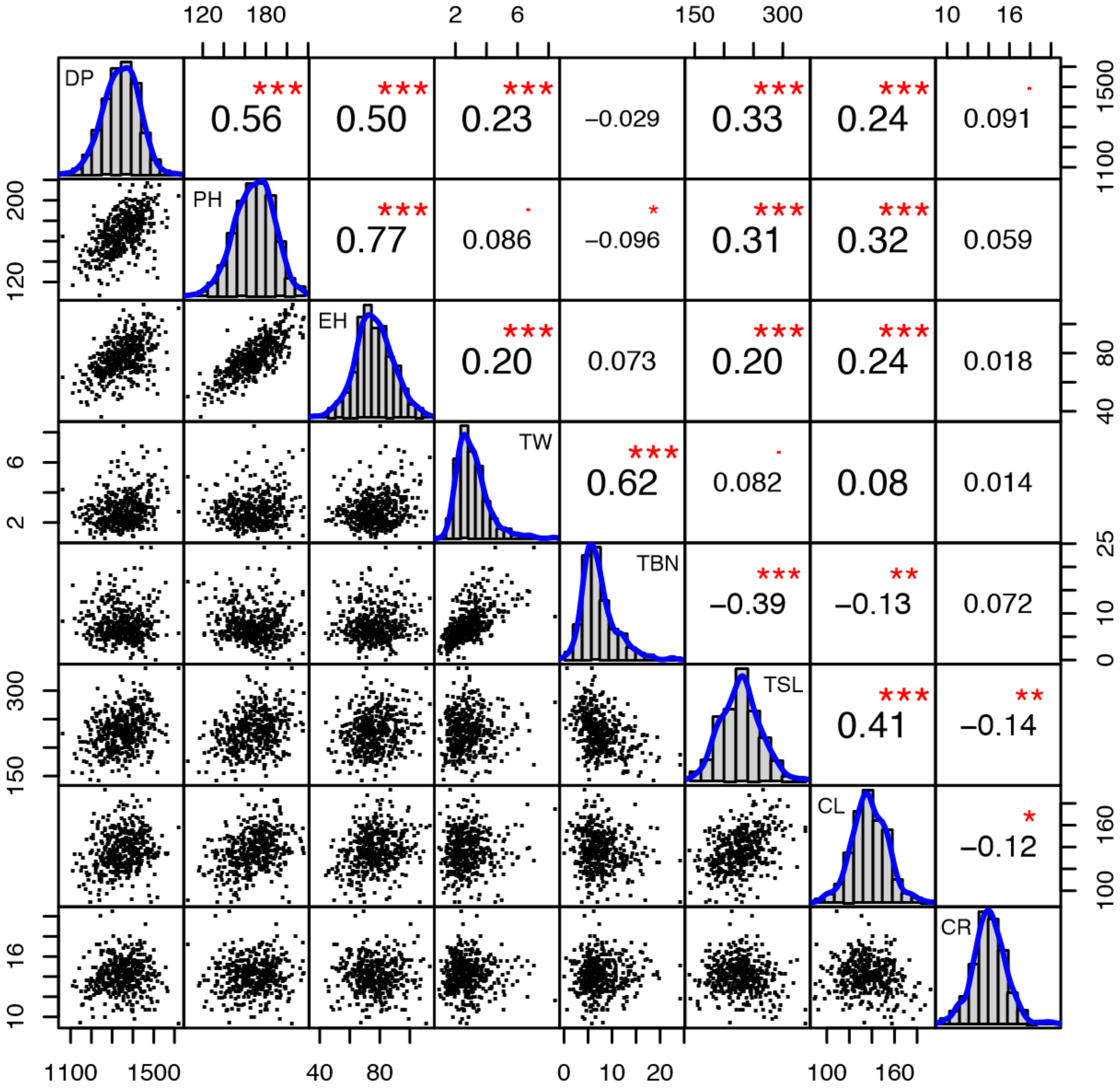

Figure 3.1. Phenotypic distributions and correlations. Histograms are on the diagonal; bivariate scatter plots are below the diagonal; and Pearson's correlation statistic (r) for each pairwise trait comparison is above the diagonal. P-value for each Pearson correlation statistic is indicated by the following: $* * * = p < 0 : 001,_* * = p < 0 : 01; * = p < 0 : 05$.

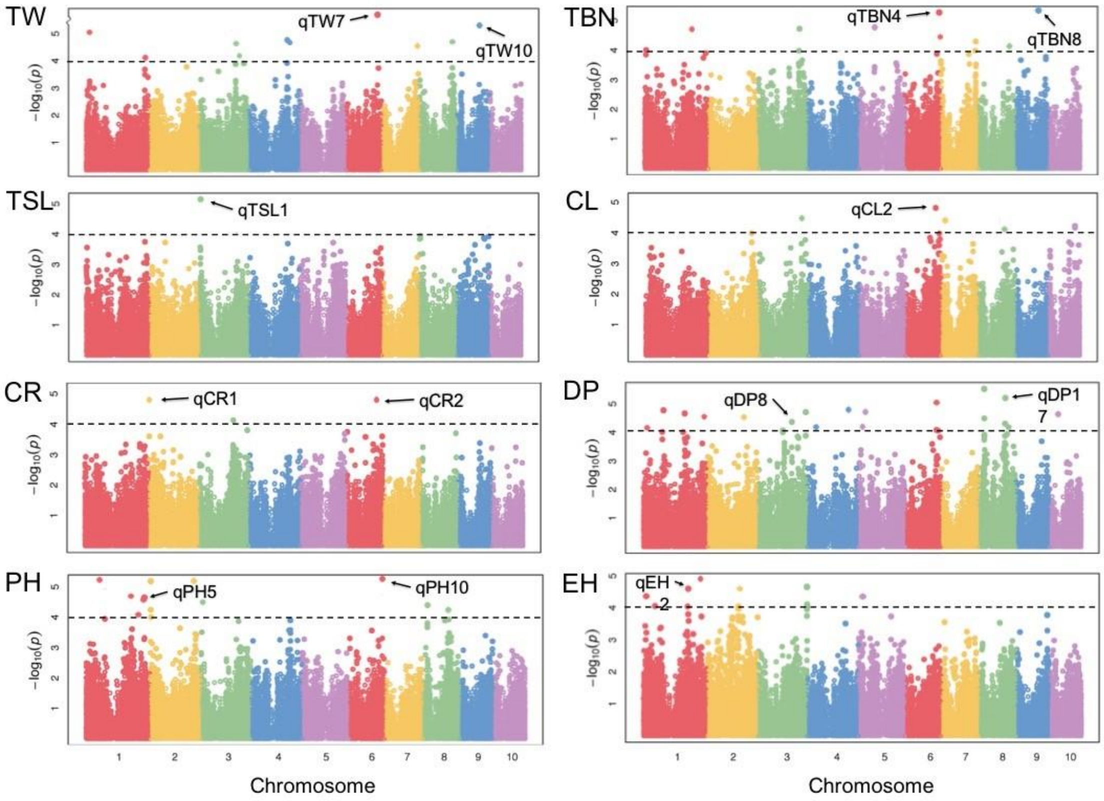

Figure 3.2. Manhattan plots from GWAS. Chromosome and relative SNP position is on the X-axis; -log(P.value) is on the Y-axis. Notable QTL discussed in the text are highlighted.

Table 3.5.: Inbreds with expired Plant Variety Protection certificates used in this study.

Inbred	Proprietor[a]	PVP#	GRIN ID[b]	Pedigree[c]
207	Pioneer	8300144	PI 601005	G3BD2/G3RZ1
740	Novartis	8800028	PI 601489	Mo17/Mexican Deep Kernel
764	Novartis	8700036	PI 601374	235/B73
778	Novartis	8700045	PI 601375	W117/B37Ht
779	Novartis	8700041	PI 601376	CM7-24/W117
787	Novartis	8800029	PI 601490	VA17/VA29
790	Novartis	8800030	PI 601491	235/B73
793	Novartis	8800031	PI 601492	235/B73
794	Novartis	8700046	PI 601377	235/B73
807	Novartis	8700151	PI 601430	W117/B37Ht
904	Northrup	9200123	PI 560317	PHI3737
911	Northrup	9200012	PI 557556	PHI3737
912	Northrup	9200013	PI 557557	PHI3737
1538	United Agriseeds	8900075	PI 601658	PHI3901/AS3(Syn.)
2369	Cargill	8800178	PI 601559	2702H/B73(1)
5707	Asgrow	8600036	PI 601269	C123Ht/Va59
6103	Asgrow	8500005	PI 601159	(AS10631/A632)/RB14AHt
11430	Cargill	8800177	PI 601558	Oh43/H99/Mo17 Composite
78004	DeKalb-Pfizer	8500125	PI 601210	B73/A634
78010	DeKalb-Pfizer	8500126	PI 601211	B73/A634
29MIBZ2	DeKalb	9100124	PI 548830	B87/PHI3901
2FACC	DeKalb	9000016	PI 601808	4676A/PB80
2FADB	DeKalb	9300084	PI 564751	4676A /PB80
2MA22	DeKalb-Pfizer	8800193	PI 601560	4780/5P9-1
2MCDB	DeKalb	9300091	PI 565088	2MA22/4780 Composite
3IBZ2	DeKalb	9100223	PI 554616	IBC2/ZZZ38
3IIH6	DeKalb	9300087	PI 564754	PHI3737
3IJI1	DeKalb	9300086	PI 564753	IBI8/PHI3603
4676A	DeKalb-Pfizer	8600092	PI 601300	1067-1/B-line Composite
4N506	Funk Seeds	8900263	PI 601745	B73/BSSS2
6F629	DeKalb	9100036	PI 546483	88051B/4608H
6M502	DeKalb-Pfizer	8800191	PI 601561	MAWU/4913
6M502A	DeKalb	9100037	PI 546484	MAWU/4913
78002A	DeKalb-Pfizer	8600091	PI 601301	B73/A634
78371A	DeKalb-Pfizer	8700172	PI 601438	4726/Iowa Long Ear
78551S	DeKalb	8800195	PI 601562	78060A/LH38
83IBI3	DeKalb	9100256	PI 555651	IBC2/IBI2
87916W	DeKalb-Pfizer	8800189	PI 601563	W37-2/B73(2)
8M129	DeKalb	9300090	PI 565087	78060A /88144
91IFC2	DeKalb	9300083	PI 564750	FR23/IBC2
AQA3	DeKalb	9300082	PI 564749	ABA10/FBAB
B09	Pioneer	8300142	PI 601007	555/031
B47	Pioneer	8300141	PI 601009	B37/SD105
BCC03	Novartis	9100002	PI 544065	3224/LH51
CQ702rc	United Agriseeds	9300186	PI 566938	KS1030/3535
CR14	J.C. Robinson	8900095	PI 601683	(B73/CM105)/(B73/CQ187)
CR1Ht	J.C. Robinson	8400042	PI 601080	W117Ht/Mo17Ht

continued on next page

Online Resource 1 *continued*

Inbred	**Proprietor**[a]	**PVP#**	**GRIN ID**[b]	**Pedigree**[b]
CS405	United Agriseeds	9200059	PI 559916	B73/K81
CS608	United Agriseeds	9200122	PI 560316	(Oh514/B68Ht)/CD1
DJ7	Funk Seeds	8500086	PI 601191	BS16(Syn.)/B73(3)
E8501	Novartis	8900233	PI 601724	387/FRMo17
F118	DeKalb	9100248	PI 555462	B73/T220
F42	FFR	8300157	PI 601026	B73 Mutation
FAPW	DeKalb-Pfizer	8200152	PI 600958	B14AH/B37H
FBHJ	DeKalb-Pfizer	8700173	PI 601439	B84/FBAB(1)
FBLA	DeKalb	9100035	PI 546482	1094-H x A656
FR 19	IL Found. Seeds	8000011	PI 600772	W438/A635
G35	Pioneer	8300140	PI 601008	PHG3BD2/H7FS6(aka PH595)
G39	Pioneer	8300115	PI 600981	B37/B14/B96/I205/IDT
G50	Pioneer	8300143	PI 601006	848/207
G80	Pioneer	8400128	PI 601037	495/331
H8431	Novartis	8800152	PI 601610	(377/B386)/347
HB8229	DeKalb	8800190	PI 601564	8200/A634H
HBA1	DeKalb-Pfizer	8500069	PI 601172	PHI3195/PHI3199
IB014	DeKalb-Pfizer	8500123	PI 601208	H99/3901(1)
IB02	DeKalb-Pfizer	8700197	PI 601457	IBI/7309B
IBB14	DeKalb	8800192	PI 601565	PHI3710/PHI3732
IBB15	DeKalb-Pfizer	8700196	PI 601458	J6/W70884
IBC2	DeKalb-Pfizer	8700198	PI 601459	Mo17Ht/J6(1)
ICI 193	Advanta	9200037	PI 559380	PHI3732/CB59G
ICI 441	Advanta	9200038	PI 559381	PHI3377/LH132
ICI 581	Zeneca	9300049	PI 564697	LH39/LH58
ICI 740	Advanta	9200039	PI 559382	PHI3377/LH132
ICI 893	Advanta	9200040	PI 559383	Pa91/B73(1)
ICI 986	Zeneca	9200041	PI 559384	PHI3540
J8606	Novartis	8900226	PI 601725	P101/C103G
L 127	Lifaco Seed	8900201	PI 601726	PHI3901/W117
L 135	Lifaco Seed	8900202	PI 601727	PHI3901/W117
L 139	Lifaco Seed	8900203	PI 601728	PHI3901/PHI3780
L 155	Limagrain	9100163	PI 550695	P-3901/A632
LH1	Holden's	7600047	PI 644101	B37/644
LH38	ISU RF	8000066	PI 600791	A619HT/L120
LH39	ISU RF	8000067	PI 600944	Oh43/L120
LH51	Holden's	8200062	PI 600955	Mo17 Backcross 5 recovery
LH52	Holden's	8700020	PI 601360	610/Mo17(2)
LH54	Holden's	8600128	PI 601316	610/Mo17(2)
LH57	Holden's	8600129	PI 601317	(Mo17/H99)/LH53
LH59	Holden's	8700213	PI 601466	(Mo17/H99)/LH53
LH60	Holden's	8700087	PI 601404	LH55/LH47
LH61	Holden's	8700137	PI 601416	ASA/Mo17(3)
LH65	Holden's	8800050	PI 601494	(Mo17/LH18)/LH53
LH74	Holden's	8200063	PI 600957	A632/B73
LH82	Holden's	8500037	PI 601170	610/LH7
LH85	Holden's	8700088	PI 601405	PHI3978
LH93	Holden's	8500038	PI 601171	BS11 FRC3 OPV
LH119	Holden's	8200064	PI 600954	H93/B73 (2)
LH123HT	Holden's	8400030	PI 601079	PHI3535

continued on next page

Online Resource 1 *continued*

Inbred	**Proprietor**[a]	**PVP#**	**GRIN ID**[b]	**Pedigree**[b]
LH127	Holden's	9000064	PI 538007	LH58/L122 (1)
LH128	Holden's	9100067	PI 547086	LH51/(BS11LHC3-S4)
LH132	Holden's	8300148	PI 601004	H93/B73(2)
LH143	Holden's	8300138	PI 601003	A635Ht/A632Ht(2)
LH145	Holden's	8300102	PI 600959	A632Ht/CM105
LH146Ht	Holden's	8700089	PI 601402	B73/CM105 (1)
LH149	Holden's	8800053	PI 601493	((A662/B73)-S1)/B73(2)
LH150	Holden's	8500153	PI 601230	PHI3147
LH156	Holden's	8700090	PI 601403	Va85/Pa91
LH159	Holden's	9200247	PI 562377	PHI3160
LH160	Holden's	9000122	PI 539920	ND246/Mo17
LH162	Holden's	9000123	PI 539921	ND246/Mo17
LH163	Holden's	9000065	PI 538008	PHI3720
LH164	Holden's	9100265	PI 555659	PHI3901
LH165	Holden's	9200248	PI 562378	LH82/LH51
LH166	Holden's	9300035	PI 564539	LH82/LH124
LH172	Holden's	9200249	PI 562379	LH122/LH82(1)
LH181	Holden's	9100068	PI 547087	LH58/LH122
LH183	Holden's	9300088	PI 564755	LH122/LH51(1)
LH184	Holden's	9300038	PI 564542	LH123Ht/LH51 (1)
LH190	Holden's	9000124	PI 539922	((B68Ht/B73Ht)-S2)/B73
LH191	Holden's	9000139	PI 539925	LH132/PHI3184
LH192	Holden's	9000140	PI 539926	LHE137/LHE136
LH193	Holden's	9000141	PI 539927	LHE137/LHE136
LH194	Holden's	9000125	PI 539923	LH117/LHE137
LH195	Holden's	9000047	PI 537097	LH117/LH132
LH196	Holden's	9000066	PI 538009	LH74/LH119
LH197	Holden's	9200020	PI 557562	LH132/B84
LH198	Holden's	9200021	PI 557563	B84/LH132(2)
LH199	Holden's	9200024	PI 557566	(LH117/LHE137)/LH132
LH202	Holden's	9000126	PI 539924	((A662/B73)-S1)/B73(2)
LH204	Holden's	9000048	PI 537098	(CB59G/LH1)/B73
LH205	Holden's	9000049	PI 537099	LH74/LH119
LH206	Holden's	9000067	PI 538010	(CB59G/LH1) /B73
LH208	Holden's	9100069	PI 547088	LH74/CB59G
LH209	Holden's	9100218	PI 554612	LH74/LH119
LH210	Holden's	9000050	PI 537100	LH51/(BS11LHC3-S3)
LH211	Holden's	9000051	PI 537101	Mo17/PHI3535
LH212Ht	Holden's	9100070	PI 547089	LH24/LH123Ht(1)
LH213	Holden's	9100071	PI 547090	LH123Ht/LH51
LH214	Holden's	9100266	PI 555660	LH123Ht/LH51
LH215	Holden's	9100201	PI 552815	R177/Mo17C2
LH216	Holden's	9200028	PI 557569	((LH123Ht/LH51(2))-S2)/LH51
LH220Ht	Holden's	9000068	PI 538011	LH74 x LH145Ht
LH222	Holden's	9200032	PI 559375	((CM174/LH74(1))-S1)/LH74
LH223	Holden's	9200250	PI 562380	CB59G/CM105
LH224	Holden's	9200251	PI 562381	LH74/CB59G(2)
LIBC4	DeKalb	9100255	PI 555650	MBNS/PHI3901
LP1CMSHT	Pfister	7800019	PI 600755	A635 Cms Ht/A632Ht(5)
LP1NRHT	Pfister	7800020	PI 600729	A632Ht/PN042

continued on next page

Online Resource 1 *continued*

Inbred	Proprietor[a]	PVP#	GRIN ID[b]	Pedigree[b]
Lp215D	Wilson Hybrids	9100084	PI 547107	Mo17/Lp216D
Lp5	Claeys Semences	8700031	PI 601378	GLAMOS/B73Ht(1)
MB-	DeKalb-Pfizer	8500127	PI 601209	Mo17Ht/MDA-28
MBPM	DeKalb-Pfizer	8700175	PI 601440	400M Composite
MBSJ	DeKalb	9100134	PI 548838	LH38/5P9-1
MBST	DeKalb-Pfizer	8800194	PI 601566	LH38/4726-1
MBUB	DeKalb-Pfizer	9100135	PI 548839	LH38/MANS
MBWZ	DeKalb	9300081	PI 564748	HBA1/IB014
MDF-13D	DeKalb-Pfizer	8200151	PI 600956	H4101/800M
ML606	United Agriseeds	9400242	PI 583774	LK2/LH38
MM402A	DeKalb	9100222	PI 554615	LH38/MANS
MM501D	DeKalb	9300085	PI 564752	LH38/88121A
MQ305	United Agriseeds	9200060	PI 559917	PHI3901/CB59G
NL001	DeKalb	9100038	PI 546485	(1089HT/A634)/B73
NQ508	United Agriseeds	9200061	PI 559918	PHI3713
NS501	DowElanco	8800149	PI 601583	A634/K1-172B
NS701	DowElanco	8700134	PI 601417	A632/B73Ht
OQ101	United Agriseeds	9200062	PI 559919	PHI3906/ND246
OQ403	United Agriseeds	9200063	PI 559920	PHI3901/K81-336
OQ603	DowElanco	8800150	PI 601584	PHI3713
OS602	United Agriseeds	9200064	PI 559921	PH3901/CM105
PB80	DeKalb-Pfizer	8700174	PI 601441	(1067-1/B73)/(B73Ht.1BC6)
PHAW6	Pioneer	9300104	PI 565100	PHN82/PHM49
PHBA6	Pioneer	9200078	PI 559935	PHZ51/PHG47
PHBB3	Pioneer	9400089	PI 578029	PHK29/PHW52
PHBW8	Pioneer	9200079	PI 559936	PHJ40/PHW52
PHEG9	Pioneer	9400090	PI 578030	PHG86/PHW52
PHEM7	Pioneer	9400092	PI578032	PHT64/PHW23
PHG29	Pioneer	8600047	PI 601270	806/207(1)
PHG47	Pioneer	8600131	PI 601318	041/MKSDTE C10
PHG71	Pioneer	8400157	PI 601150	A632Ht/207
PHG72	Pioneer	8600134	PI 601319	891/207
PHG83	Pioneer	8500152	PI 601229	814/207
PHG84	Pioneer	8600130	PI 601320	848/595
PHG86	Pioneer	8700170	PI 601442	B64/B73
PHGG7	Pioneer	9200081	PI 559938	PHT64/PHG49
PHGV6	Pioneer	9200082	PI 559939	PH814/PHG65
PHGW7	Pioneer	9200083	PI 559940	PHR25/PHR64
PHH93	Pioneer	8800216	PI 601567	PH806/207
PHHH9	Pioneer	9300109	PI 565105	PHJ29/PHBT4
PHHV4	Pioneer	9200084	PI 559941	PHG69/PHM44
PHJ31	Pioneer	8900307	PI 601773	B97/595
PHJ33	Pioneer	8900308	PI 601774	PHG83/CE18
PHJ40	Pioneer	8600133	PI 601321	B09/B36
PHJ65	Pioneer	9000245	PI 543840	PHG63/PHG65
PHJ70	Pioneer	8900309	PI 601775	AC26/B73Ht
PHJ75	Pioneer	8900310	PI 601776	207/G96
PHJ89	Pioneer	9100092	PI 548798	PHT77/PHG47
PHJ90	Pioneer	9100093	PI 548799	G50/PHK42
PHJR5	Pioneer	9300110	PI 565106	PHG73/PHT10

continued on next page

Online Resource 1 *continued*

Inbred	Proprietor[a]	PVP#	GRIN ID[b]	Pedigree[b]
PHK05	Pioneer	8800001	PI 601467	CM7/051
PHK29	Pioneer	8700214	PI 601468	B47/AC54
PHK35	Pioneer	8900311	PI 601777	AC34/G93H
PHK42	Pioneer	8800035	PI 601495	806/207 (1)
PHK46	Pioneer	9000246	PI 543841	PHG65/207
PHK56	Pioneer	9000247	PI 543842	PHG47/PHG35
PHK74	Pioneer	9200085	PI 559942	PHFA0/PHG72
PHK76	Pioneer	8800036	PI 601496	AD18/B02
PHK93	Pioneer	9100094	PI 548800	PHB72/PHT60
PHKE6	Pioneer	9300111	PI 565107	PHG29/PHG47
PHM10	Pioneer	8900312	PI 601778	PHG39/207
PHM49	Pioneer	8800211	PI 601568	PHB81/PHR33
PHM57	Pioneer	8900313	PI 601779	B97/595
PHM81	Pioneer	9100095	PI 548801	PHG72/PHG68
PHN11	Pioneer	8800037	PI 601497	806/207 (1)
PHN29	Pioneer	8900314	PI 601780	PHG69/PHG40
PHN34	Pioneer	9000248	PI 543843	SC359/PH157
PHN37	Pioneer	8900315	PI 601781	CM11/041Ht
PHN41	Pioneer	9300113	PI 565109	PHDK6/PHNN2
PHN47	Pioneer	8800217	PI 601569	207/PHB60
PHN66	Pioneer	9100096	PI 548802	PHG53/PHG21
PHN73	Pioneer	8900316	PI 601782	041/PHG35
PHN82	Pioneer	8900317	PI 601783	PHG29/HD38
PHP02	Pioneer	8800212	PI 601570	PHG44/PHG29
PHP38	Pioneer	9000250	PI 543844	PHG39/PHK29
PHP55	Pioneer	8900318	PI 601784	PHG44/PHG29
PHP60	Pioneer	8900319	PI 601785	AT2/805
PHP76	Pioneer	9000251	PI 543846	PHG50/PHEJ8
PHP85	Pioneer	9200087	PI 559944	PHK29/PHW52
PHPR5	Pioneer	9200088	PI 559945	PHK76/PHW52
PHR03	Pioneer	9100097	PI 548803	PHT19/PHG84
PHR25	Pioneer	8800002	PI 601469	B83/207
PHR30	Pioneer	9200089	PI 559946	PHFM5/PHG47
PHR31	Pioneer	9200090	PI 559947	G50/PHRH7
PHR32	Pioneer	8800218	PI 601571	PHB82/PHG61
PHR36	Pioneer	8700017	PI 601361	(203/549)/848
PHR47	Pioneer	8800213	PI 601572	G39/PHB49
PHR55	Pioneer	9100098	PI 548804	PH005/PHG84
PHR58	Pioneer	9100099	PI 548805	PH383/PHG16
PHR62	Pioneer	8900320	PI 601786	G50/G35
PHR63	Pioneer	8900321	PI 601787	PHG29/B89
PHT10	Pioneer	8800214	PI 601573	B73/G39
PHT22	Pioneer	8900322	PI 601788	207/HD12
PHT47	Pioneer	9200091	PI 559948	PHB47/G39
PHT55	Pioneer	8800046	PI 601498	A33GB4/A34CB4
PHT60	Pioneer	8800219	PI 601574	PHW94/PHV80
PHT69	Pioneer	9200092	PI 559949	PHR73/PHJ40
PHT73	Pioneer	9200093	PI 559950	PHK05/PHG68
PHT77	Pioneer	8800038	PI 601499	814/995
PHTM9	Pioneer	9200094	PI 559951	PHG47/PHG36

continued on next page

Online Resource 1 *continued*

Inbred	**Proprietor**[a]	**PVP#**	**GRIN ID**[b]	**Pedigree**[b]
PHV07	Pioneer	9000252	PI 543847	PHG41/G21
PHV37	Pioneer	8900323	PI 601789	G27/G21
PHV53	Pioneer	9200095	PI 559952	PHB89/PHDT2
PHV57	Pioneer	9300115	PI 565111	G50/PHG72
PHV63	Pioneer	8800039	PI 601500	555/Zap¡4CB
PHV78	Pioneer	8800003	PI 601470	G42/595
PHVA9	Pioneer	9200096	PI 559953	PHK29/PHGP8
PHVJ4	Pioneer	9300103	PI 565099	PHJ40/207
PHW03	Pioneer	8900324	PI 601790	801/G48
PHW17	Pioneer	8700018	PI 601362	(1D11/B73)/(B73/051)
PHW20	Pioneer	8900325	PI 601791	(1D11/1M12)/B76
PHW30	Pioneer	9100102	PI 548808	PHG42/PHV15
PHW43	Pioneer	8900326	PI 601792	995/G35
PHW51	Pioneer	9000254	PI 543849	PHDF2/PHG41
PHW52	Pioneer	8800215	PI 601575	B73/G39
PHW53	Pioneer	9300116	PI 565112	G50/PHZ51
PHW65	Pioneer	8800040	PI 601501	861/595
PHW79	Pioneer	8800220	PI 601576	PHT90/595
PHW80	Pioneer	9300117	PI 565113	PHK76/PHN37
PHW86	Pioneer	9000255	PI 543850	PHG71/PHG72
PHWG5	Pioneer	9200097	PI 559954	PH814/PHG16
PHZ51	Pioneer	8600132	PI 601322	814/848
Q381	QRA	8500098	PI 601190	PHI3369 off-type
RS710	Rustica Semences	9000129	PI 539930	PAG1202/A641
S8324	Novartis	8800153	PI 601611	(CH593-9/B73)-S2)/B73
S8326	Novartis	8800154	PI 601612	(W117/Mo17)-S2))/Mo17
Seagull 17[d]	Rothermel	7900077	PI 600751	Mo17/Unknown
W8304	Novartis	8800032	PI 601502	B14A/B73(1)
W8555	Novartis	8900227	PI 601729	B73Ht/B84
WIL500	Wilson Hybrids	8900156	PI 601689	82C25 (Exotic Syn)
WIL900	Wilson Hybrids	8900092	PI 601684	Mo17/Tuxpeno (82C43)
WIL901	Wilson Hybrids	8900093	PI 601685	Mo17/Tuxpeno (82C232)
WIL903	Wilson Hybrids	8900094	PI 601686	Mo17/Tuxpeno (82C43)
ZS01250	Advanta	9600271	PI 595616	Unknown
ZS365	Advanta	9300304	PI 574393	PHI3358/PHI3713
ZS635	Advanta	9300305	PI 574394	PHI3358/PHI3713

[a] Proprietor names have been abbreviated; full legal company names are stated on the ex-PVP certificates. Explanation of abbreviations in the list above that are helpful for company name identification: IL Found. Seeds=Illinois Foundation Seeds, Inc., ISU RF= Iowa State University Research Foundation, and QRA=Quality Research Associates.

[b] GRIN ID=Germplasm Resource Information Network ID.

[c] All pedigrees were obtained from PVP certificates, available at ars.grin.gov.

[d] ”Seagull Seventeen” was shortened to ”Seagull 17” for formatting reasons

Epilogue

In Chapter 1, I predicted the mean, variance, and superior progeny mean of simulated biparental populations, and used these statistics to identify optimal parental combinations to produce a new inbred with improved performance in a hybrid testcross with Iodent tester PHP02. Others have predicted similar statistics (Bernardo, 2015; Mohammadi et al., 2015; Lehermeier et al., 2017; Osthushenrich et al., 2017), but all used either inbred or simulated data in the training set; none used hybrid testcross data in the training set. Some of the best predicted biparental combinations (based on both progeny mean and variance for grain yield) include: LH213 and PHR58, LH214 and PHR58, 2FACC and LH213, and 2FACC and PHR58. Most of the best biparental combinations (i.e. those pairs with the highest predicted progeny mean and predicted superior progeny mean) were those with inbreds from different proprietors. This is no surprise, as many breeders have observed that in the 3-5 years following acquisition of a new company and incorporation of its accompanying germplasm pool, there is an increase in heterotic response. In this chapter, I also show that genetic diversity between any two inbreds is a poor predictor of genetic variance within a biparental breeding populations created by those two inbreds. While not a new finding, this conclusion agrees with many previous studies on the subject (Cowen & Frey, 1987; Souza & Sorrells, 1991; Kisha et al., 1997; Manjarrez-Sandoval et al., 1997; Burkhamer et al., 1998; Bohn et al., 1999; Utz et al., 2001; Gutierrez et al., 2002; Barroso et al., 2003; Hung et al., 2012).

If I were to do this project again, I would add the inbreds with recently expired PVP certificates to the validation, or testing set. As the newly expired PVP inbreds are often descendants from older ex-PVP inbreds that are already in the training set, the genetic relationship between training set and prediction set is likely high enough to produce a useful genomic prediction. If I were a breeder at a small commercial program and were looking to leverage publicly available germplasm to jump-start my program, this genomic prediction approach would allow me to quickly identify key inbreds that that I could use to build high-performing complementary heterotic groups.

In Chapter 2, I completed a study of genomic prediction accuracy within a non-double-haploid early generation breeding population. One use of genomic selection is to grow part of a breeding population in a hybrid testcross trial, and use that data to predict the phenotypes for the remaining part of the breeding population, those individuals with genotypes but no phenotypes. By performing genomic prediction and cross-validation within a set of F4 lines, I found that gains in prediction accuracy were marginal when greater than 150 individuals were included in the training set. Therefore, it that a training set with 150 individuals will achieve the optimal balance between cost and benefit when predicting the remaining individuals in the population. This agrees with Bernardo and Yu's (2007) conclusion that a minimum of 100 to 150 lines in a training set is sufficient to ensure optimal prediction accuracy. Grain yield was a much more difficult trait to predict than test weight or grain moisture; this is likely due to the highly quantitative nature of the genetics of grain yield. I also found that RR-BLUP was the overall top performer across both testers and all traits, as it sufficiently modeled the additive genetic effect–the most useful effect for making early generation selections in plant breeding populations. Other prediction models may prove useful in other situations–such as the BayesB model for traits with a few large-effect loci, or a non-parametric model such as Random Forest model for estimation of commercial value (i.e. total genetic value) of a line. Based on the results of this project, using RR-BLUP and a minimum of 150 individuals in a training set will likely return the highest prediction accuracy when predicting hybrid testcross performance of related individuals within a similar experimental design and maize germplasm.

In retrospect, chapter 3 was the most difficult of the three research topics. Originally, the objective of this project was to use hybrid testcross data from the F3 generation to predict performance in the F4 generation. However, due to an oversight, the F4 generation was crossed to different testers than the F3 generation. Preliminary analysis showed little to no correlation of related individuals between testers. I concluded that it was not possible to execute the project as originally planned. I took inventory of what data I had, and after consulting with several advisors, I concluded that the objective as presented in Chapter 2 of this dissertation was the best course to follow. I also ended up expanding the use of statistical prediction models, and ran about a dozen more models (or variations of models) than what I report on in Chapter 2.

In Chapter 3, I report on the identification of a large number of QTL associated with inflorescence traits in maize. Maize inflorescence development is important to the seed industry, as much of the value of final product (either seed or grain) is determined by the inflorescence. Understanding the genetic architecture of the maize inflorescences will help scientists to improve seed production efficiency and create inbreds and hybrids with better dry-matter partitioning.

This project was a good opportunity for me to learn about the North American maize inbreds with expired Plant variety Protection certificates. Many of these inbreds are progenitors of today's best commercial inbreds. This collection of inbreds can be regarded as the most commercially relevant germplasm set in academia. As my goal after graduating has always been to obtain employment as a plant breeder in the industry, this project was instrumental in exposing me to the inbreds and heterotic groups that exist in today's North American dent corn germplasm.

I believe that harnessing the genetic components of performance traits in crops has great promise to helping increase food production to meet the needs of a growing population. These three topics have proved to be a solid foundation for me as I now move on to a career as a plant breeder, where I will doubtless have many opportunities to increase the potential of crop performance and production. I am grateful for my time and training at Purdue University, and look forward to contributing to the future of plant breeding!

REFERENCES

REFERENCES

Albrecht, T., Auinger, H.-J., Wimmer, V., Ogutu, J. O., Knaak, C., Ouzunova, M., ... Schön, C.-C. (2014). Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theoretical and applied genetics*, *127*(6), 1375–1386.

Albrecht, T., Wimmer, V., Auinger, H.-J., Erbe, M., Knaak, C., Ouzunova, M., ... Schön, C.-C. (2011). Genome-based prediction of testcross values in maize. *Theoretical and Applied Genetics*, *123*(2), 339.

Barroso, P. A. V., Geraldi, I. O., Vieira, M. L. C., Pulcinelli, C. E., Vencovsky, R., & Dias, C. T. d. S. (2003). Predicting performance of soybean populations using genetic distances estimated with RAPD markers. *Genetics and Molecular Biology*, *26*(3), 343–348.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Beckett, T., Morales, A., Koehler, K., & Rocheford, T. (2017). Genetic relatedness of previously Plant-Variety-Protected commercial maize inbreds. *PLOS ONE*, *12*(12), e0189277.

Beckett, T. J. (2016). *Analysis of genetic loci associated with agronomic performance in previously Plant-Variety-Protected elite commercial maize germplasm.*

Belkhir, K., Dawson, K. J., & Bonhomme, F. (2006). A comparison of rarefaction and bayesian methods for predicting the allelic richness of future samples on the basis of currently available samples. *Journal of Heredity*, *97*(5), 483–492.

Bernardo, R. (2002). *Breeding for quantitative traits in plants.* Stemma Press Woodbury.

Bernardo, R. (2014). *Essentials of plant breeding.*

Bernardo, R. (2015). Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. *Crop Science*, *54*(6), 2586–2595.

Bernardo, R. (2016). Bandwagons I, too, have known. *Theoretical and Applied Genetics*, *129*(12), 2323–2332.

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*(3), 1082–1090.

Bohn, M., Utz, H. F., & Melchinger, A. E. (1999). Genetic similarities among winter wheat cultivars determined on the basis of RFLPs, AFLPs, and SSRs and their use for predicting progeny variance. *Crop science*, *39*(1), 228–237.

Bouchet, S., Bertin, P., Presterl, T., Jamin, P., Coubriche, D., Gouesnard, B., … Charcosset, A. (2016). Association mapping for phenology and plant architecture in maize shows higher power for developmental traits compared with growth influenced traits. *Heredity*.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brown, P. J., Upadyayula, N., Mahone, G. S., Tian, F., Bradbury, P. J., Myles, S., … Buckler, E. S. (2011). Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS genetics*, *7*(11), e1002383.

Bruinsma, J., et al. (2009). The resource outlook to 2050: By how much do land, water and crop yields need to increase by 2050. In *Expert meeting on how to feed the world in* (Vol. 2050, pp. 24–26).

Burkhamer, R. L., Lanning, S. P., Martens, R. J., Martin, J. M., & Talbert, L. E. (1998). Predicting progeny variance from parental divergence in hard red spring wheat. *Crop Science*, *38*(1), 243–248.

Charcosset, A., Lefort-Buson, M., & Gallais, A. (1991). Relationship between heterosis and heterozygosity at marker loci: a theoretical computation. *Theoretical and applied genetics*, *81*(5), 571–575.

Cowen, N., & Frey, K. (1987). Relationship between genealogical distance and breeding behaviour in oats (*Avena sativa* L.). *Euphytica*, *36*(2), 413–424.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., … Beyene, Y. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in plant science*.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. (2012). Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics*, genetics–112.

Duvick, D., Smith, J., & Cooper, M. (2010). Long-term selection in a commercial hybrid maize breeding program. *Janick. I. Plant Breeding Reviews. Part*, *2*(24), 109–152.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*, *6*(5), e19379.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrblup. *Plant Genome*, *4*, 250-255.

Falconer, D. S., Mackay, T. F., & Frankham, R. (1996). Introduction to quantitative genetics. *Trends in Genetics*, *12*(7), 280.

Fehr, W. (1991). *Principles of cultivar development: theory and technique.* Macmillian Publishing Company.

Fischer, K., Edmeades, G., & Johnson, E. (1987). Recurrent selection for reduced tassel branch number and reduced leaf area density above the ear in tropical maize populations. *Crop science*, *27*(6), 1150–1156.

Flint-Garcia, S. A., Darrah, L. L., McMullen, M. D., & Hibbard, B. E. (2003). Phenotypic versus marker-assisted selection for stalk strength and second-generation european corn borer resistance in maize. *Theoretical and Applied Genetics*, *107*(7), 1331–1336.

Galli, M., Liu, Q., Moss, B. L., Malcomber, S., Li, W., Gaines, C., … Nemhauser, J. L. (2015). Auxin signaling modules regulate maize inflorescence architecture. *Proceedings of the National Academy of Sciences*, *112*(43), 13372–13377.

Gerdes, J., Tracy, W., Coors, J., Geadlemann, J., & Viney, M. K. (1993). *Compilation of North American maize breeding germplasm.* Crop Science Society of America Madison, WI.

Gutierrez, O., Basu, S., Saha, S., Jenkins, J., Shoemaker, D., Cheatham, C., & McCarty, J. (2002). Genetic distance among selected cotton genotypes and its relationship with F2 performance. *Crop Science*, *42*(6), 1841–1847.

Hayes, B., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Heffner, E. L., Lorenz, A. J., Jannink, J.-L., & Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop science*, *50*(5), 1681–1690.

Heffner, E. L., Sorrells, M. E., & Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*(1), 1–12.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 423–447.

Hung, H., Browne, C., Guill, K., Coles, N., Eller, M., Garcia, A., … Salvo, S. (2012). The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity*, *108*(5), 490.

Jacobson, A., Lian, L., Zhong, S., & Bernardo, R. (2014). General combining ability model for genomewide selection in a biparental cross. *Crop Science*, *54*(3), 895–905.

Jannink, J.-L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics*, *9*(2), 166–177.

Jia, Y., & Jannink, J.-L. (2012). Multiple trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, genetics–112.

Kadam, D. C., Potts, S. M., Bohn, M. O., Lipka, A. E., & Lorenz, A. J. (2016). Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. *G3: Genes, Genomes, Genetics*, *6*(11), 3443–3453.

Kisha, T., Sneller, C., & Diers, B. (1997). Relationship between genetic distance among parents and genetic variance in populations of soybean. *Crop Science*, *37*(4), 1317–1325.

Ko, S.-S., Li, M.-J., Ku, M. S.-B., Ho, Y.-C., Lin, Y.-J., Chuang, M.-H., ... Chang, H.-C. (2014). The bHLH142 transcription factor coordinates with TDR1 to modulate the expression of EAT1 and regulate pollen development in rice. *The Plant Cell*, *26*(6), 2486–2504.

Kruse, J. (2010). Estimating demand for agricultural commodities to 2050. *Global Harvest Initiative*.

Lambert, R., & Johnson, R. (1978). Leaf angle, tassel morphology, and the performance of maize hybrids. *Crop Science*, *18*(3), 499–502.

Leberg, P. (2002). Estimating allelic richness: effects of sample size and bottlenecks. *Molecular ecology*, *11*(11), 2445–2449.

Lehermeier, C., Teyssèdre, S., & Schön, C.-C. (2017). Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. *Genetics*, *207*(4), 1651–1661.

Leng, P.-f., Lübberstedt, T., & Xu, M.-l. (2017). Genomics-assisted breeding–a revolutionary strategy for crop improvement. *Journal of Integrative Agriculture*, *16*(12), 2674–2685.

Li, Y.-x., Li, C., Bradbury, P. J., Liu, X., Lu, F., Romay, C. M., ... Shi, Y. (2016). Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. *The Plant Journal*, *86*(5), 391–402.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18–22.

Lin, C.-S., & Poushinsky, G. (1983). A modified augmented design for an early stage of plant selection involving a large number of test lines without replication. *Biometrics*, 553–561.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., ... Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, *28*(18), 2397–2399.

Lorenz, A. J. (2013). Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. *G3: Genes, Genomes, Genetics*, *3*(3), 481–491.

Lorenzana, R. E., & Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and applied genetics*, *120*(1), 151–161.

Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B., & de Los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. *PLoS genetics*, *7*(4), e1002051.

Manjarrez-Sandoval, P., Carter, T. E., Webb, D., & Burton, J. (1997). Rflp genetic similarity estimates and coefficient of parentage as genetic variance predictors for soybean yield. *Crop science*, *37*(3), 698–703.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Chakravarti, A. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747.

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., . . . Bottoms, C. (2009). Genetic properties of the maize nested association mapping population. *Science*, *325*(5941), 737–740.

Meghji, M., Dudley, J., Lambert, R., & Sprague, G. (1984). Inbreeding depression, inbred and hybrid grain yields, and other traits of maize genotypes representing three eras. *Crop Science*, *24*(3), 545–549.

Meuwissen, T., Hayes, B., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Mohammadi, M., Tiede, T., & Smith, K. P. (2015). PopVar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. *Crop Science*, *55*(5), 2068–2077.

Morales, A. J. (2013). *Genomic approaches for improving grain yield in maize using formerly plant variety protected germplasm* (Unpublished doctoral dissertation). PURDUE UNIVERSITY.

Navarro, J. A. R., Willcox, M., Burgueño, J., Romay, C., Swarts, K., Trachsel, S., . . . Vidal, V. (2017). A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nature genetics*, *49*(3), 476–480.

Northoff, E. (2016). 2050 a third more mouths to feed. *Food and Agriculture Organization of the United Nations*.

Nyquist, W. E., & Baker, R. (1991). Estimation of heritability and prediction of selection response in plant populations. *Critical reviews in plant sciences*, *10*(3), 235–322.

Osthushenrich, T., Frisch, M., & Herzog, E. (2017). Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. *PloS one*, *12*(12), e0188839.

Peiffer, J. A., Romay, M. C., Gore, M. A., Flint-Garcia, S. A., Zhang, Z., Millard, M. J., . . . Bradbury, P. J. (2014). The genetic architecture of maize height. *Genetics*, *196*(4), 1337–1356.

Pérez, P., & de Los Campos, G. (2014). Genome-wide regression & prediction with the BGLR statistical package. *Genetics*, genetics–114.

Poland, J., & Rutkoski, J. (2016). Advances and challenges in genomic selection for disease resistance. *Annual review of phytopathology*, *54*, 79–98.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org`

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., . . . Melchinger, A. E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature genetics*, *44*(2), 217–220.

Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., Mitchell, S. E., … Buckler, E. S. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences*, *112*(12), 3823–3828.

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., … Flint-Garcia, S. A. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, *14*(6), R55.

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., … Li, Y. (2016). Genome-enabled prediction models for yield related traits in chickpea. *Frontiers in plant science*, *7*, 1666.

Schnell, F., & Utz, H. (1976). F1 leistung und elternwahl in der zuchtung von selbstbefruchtern. *Ber Arbeitstag Arbeitsgem Saatzuchtleiter*.

Sen, T. Z., Harper, L. C., Schaeffer, M. L., Andorf, C. M., Seigfried, T. E., Campbell, D. A., & Lawrence, C. J. (2010). Choosing a genome browser for a model organism database: surveying the maize community. *Database*, *2010*.

Souza, E., & Sorrells, M. (1991). Prediction of progeny variation in oat from parental genetic relationships. *Theoretical and applied Genetics*, *82*(2), 233–241.

Stromberg, L., Dudley, J., & Rufener, G. (1994). Comparing conventional early generation selection with molecular marker assisted selection in maize. *Crop Science*, *34*(5), 1221–1225.

Sun, C., Fang, J., Zhao, T., Xu, B., Zhang, F., Liu, L., … Chen, F. (2012). The histone methyltransferase SDG724 mediates H3K36me2/3 deposition at MADS50 and RFT1 and promotes flowering in rice. *The Plant Cell*, *24*(8), 3235–3247.

Tiede, T., Kumar, L., Mohammadi, M., & Smith, K. P. (2015). Predicting genetic variance in bi-parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. *Molecular breeding*, *35*(10), 199.

Truntzler, M., Ranc, N., Sawkins, M., Nicolas, S., Manicacci, D., Lespinasse, D., … Muller, C. (2012). Diversity and linkage disequilibrium features in a composite public/private dent maize panel: consequences for association genetics as evaluated from a case study using flowering time. *Theoretical and applied genetics*, *125*(4), 731–747.

Uribelarrea, M., Carcova, J., Otegui, M., & Westgate, M. (2002). Pollen production, pollination dynamics, and kernel set in maize. *Crop Science*, *42*(6), 1910–1918.

USDA. (2013a). *Germplasm resources information network-(GRIN)[online database]*. National Germplasm Resources Laboratory, Beltsville, Maryland. (USDA, ARS, National Genetic Resources Program)

USDA. (2013b). *Plant Variety Protection Act and regulations and rules of practice.* Agricultural Marketing Service, Washington, DC. (Updated July 2013)

Utz, H., Bohn, M., & Melchinger, A. (2001). Predicting progeny means and variances of winter wheat crosses from phenotypic values of their parents. *Crop Science*, *41*(5), 1470–1478.

Vollbrecht, E., & Schmidt, R. J. (2009). Development of the inflorescences. In *Handbook of maize: Its biology* (pp. 13–40). Springer.

Wallace, J., Larsson, S., & Buckler, E. (2014). Entering the second century of maize quantitative genetics. *Heredity*, *112*(1), 30–38.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Wehrens, R., & Mevik, B.-H. (2007). The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software*, *18*(2), 1–24.

Whipple, C. J., Hall, D. H., DeBlasio, S., Taguchi-Shiobara, F., Schmidt, R. J., & Jackson, D. P. (2010). A conserved mechanism of bract suppression in the grass family. *The Plant Cell*, *22*(3), 565–578.

Wright, M. N., & Ziegler, A. (2015). Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Wu, Y., San Vicente, F., Huang, K., Dhliwayo, T., Costich, D. E., Semagn, K., ... Zhang, X. (2016). Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing snps. *Theoretical and Applied Genetics*, *129*(4), 753–765.

Xavier, A., Xu, S., Muir, W. M., & Rainey, K. M. (2015). NAM: association studies in multiple populations. *Bioinformatics*, btv448.

Yu, J., Holland, J. B., McMullen, M. D., & Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, *178*(1), 539–551.

Zhao, W., Canaran, P., Jurkuta, R., Fulton, T., Glaubitz, J., Buckler, E., ... Holland, J. (2006). Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic acids research*, *34*(suppl 1), D752–D757.