The accuracy of genomic prediction between environments and populations for soft wheat traits

Mao Huang[1], Brian Ward[2], Carl Griffey[2], David Van Sanford[3], Anne McKendry[4], Gina Brown-Guedira[5], Priyanka Tyagi[6], and Clay Sneller[1*]

[1]The Ohio State University, Ohio Agriculture Research and Development Center, 1680 Madison Ave., Wooster, OH 44691; [2]University of Virginia Tech, 185 Ag Quad Lane, Blacksburg, VA 24061; [3]University of Kentucky, 1405 Veterans Drive, Lexington, KY 40546; [4]University of Missouri, Columbia, MO 65211; [5]US Department of Agriculture–Agricultural Research Service Eastern Regional Small Grains Genotyping Laboratory, Raleigh, NC, USA; [6]Dept of Crop and Soil Sciences, North Carolina State University, Raleigh, NC 27695.

[*] Corresponding author, sneller.5@osu.edu, (330) 263-3944

ABSTRACT

Genomic selection (GS) uses training population (TP) data to estimate the value of lines in a selection population (SP). In breeding, the TP and SP are often grown in different environments which can cause low prediction accuracy when the correlation of genetic effects between the environments is low. Subsets of TP data may be more predictive than using all TP data. Our objectives were 1) to evaluate the effect of using subsets of TP data on GS accuracy between environments, and 2) to assess the accuracy of models incorporating marker by environment interactions (MEI). Two wheat populations were phenotyped for 11 traits in independent environments and genotyped with SNP markers. Within each population-trait combination, environments were clustered. Data from one cluster was used as the TP to predict the value of the same lines in the other cluster(s) of environments. Models were built using all TP data or subsets of markers selected for their effect and stability. The GS accuracy using all TP data was greater than 0.25 for nine of 11 traits. The between-environment accuracy was generally greatest using a subset of stable and significant markers: accuracy increased up to 48% relative to using all TP data. We also assessed accuracy using each population as the TP and the other as the SP. Using subsets of TP data or the MEI models did not improve accuracy between populations. Using optimized subsets of markers within a population can improve GS accuracy by reducing noise in the prediction data set.

Abbreviations list:

AMMI, Additive Main effects and Multiplicative Interaction methods; AMMI10: 10% of markers subset based on AMMI stability of each marker; AHD, AMMI stability index for heading date; AHGT, AMMI stability index for plant height; ATW, AMMI stability index for test weight; AYLD, AMMI stability index for yield; BHD, ERR index for heading date; BHGT, ERR index for plant height; BLUPs, best linear unbiased prediction; BTW, ERR index for test weight; BYLD, ERR index for yield; CD, coefficient of determination; EM, expectation maximization algorithm; EP, elite panel population; ERR, Eberhart and Russell regression; FP, flour protein; FY, flour yield; GEBVs, genomic estimated breeding values; GEI, genotype by environment interactions; GS, genomic selection; HD, heading date; HGT, plant height; LA, solvent retention capacity for lactic acid; LD, linkage disequilibrium; MEI, maker by environment interactions; P0.05: subset of markers that have significant effects for the traits; PCA, principal component analysis; PEV, prediction error variance; PVAR10: 10% of markers subset based on both marker effects and the stability index; QTL, quantitative trait loci; RAN10, random set of 10% of markers; RAN40, random set of 40% of markers; rrBLUP, ridge regression best linear unbiased prediction; SO, solvent retention capacity for sodium carbonate; SE, softness equivalent; SU, solvent retention capacity for sucrose content; TP, training population; TW, test weight; VAR10, 10% of markers subset based on lowest marker by environment interactions; VAR40, 40% of markers subset based on lowest marker by environment interactions; VP, validation population; WA, solvent retention capacity for water; YLD, yield; YP, yield panel population.

INTRODUCTION

Genomic selection (GS) is a marker assisted selection tool first proposed for animal breeding by Meuwissen et al., (2001). In GS, individuals with existing phenotypic and genotypic data are used as a training population (TP) to build a prediction model. This model is used to calculate the genomic estimated breeding values (GEBVs) of individuals based solely on their genotypic data. Breeders can then select individuals with desirable GEBVs and intermate them to initiate the next breeding cycle. This can greatly reduce the duration of a breeding cycle (Meuwissen et al., 2001; Heffner et al., 2009; Jannink et al., 2010). As genotyping costs drastically decrease with the advent of new technology, GS can be cheaper than phenotypic selection (Elshire et al., 2011; Goddard and Hayes, 2007; Heffner et al., 2010; Jannink et al., 2010; Meuwissen et al., 2001) and thus facilitate the use of larger populations of related lines in selections.

Many studies have assesed prediction accuracy of GS in crops (Battenfield et al., 2016; Denis and Bouvet, 2011; Heffner et al., 2011a; Rutkoski et al., 2012). The accuracy of GS has been assessed in wheat for many traits including grain yield, test weight, thousand kernel weight, quality traits (flour yield, flour protein, grain hardness, grain length, grain width, sodium dodecyl sulfate sedimentation), resistance to Fusarium head blight and stem rust as well as trait stability for grain yield, test weight, plant height, heading date and quality traits (Crossa et al., 2016; Heffner et al., 2011a; Heffner et al., 2011b; Hoffstetter et al., 2016; Huang et al., 2016; Michel et al., 2016a; Poland et al., 2012; Rutkoski et al., 2011; Rutkoski et al., 2012; Rutkoski et al., 2014).

Most studies in wheat use subsets of TP lines as the VP and evaluate GS accuracy through cross validation (CV, Arruda et al., 2016; Crossa et al., 2010; Heffner et al., 2011b; Heslot et al., 2012; Hoffstetter et al., 2016; Huang et al., 2016; Jiang et al., 2016; Michel et al.,

2016a; Poland et al., 2012; Rutkoski et al., 2016; Zhao et al., 2014). Compared to making predictions for different populations, the accuracy of GS will be high when using CV within the same population, as lines in TP and VP are often tested in the same set of environments. When implementing GS in actual plant breeding, the TP and the VP (or selection population) will always be phenotyped in different years or seasons so the accuracy of GS between environments is a crucial issue. There are some studies that use separate TP and VP phenotyped in different environments to assess the efficacy of GS. Sallam and Smith (2016) used a VP that consisted of progeny that were directly derived from the TP and reported that GS accuracy for barley yield ranged from 0.36 to 0.66. Michel et al. (2016b) reported the GS accuracy for wheat grain yield ranged from 0.14 to 0.75 when TP and VP consisted of different lines grown in different years.

The GS accuracy between environments or populations is affected by the relatedness of the TP to the VP (Asoro et al., 2011; Heslot et al., 2013; He et al., 2016; Rutkoski et al., 2015; Zhang et al. 2016), the linkage disequilebrium (LD) patterns across the TP and VP, size of TP, marker density (Asoro et al., 2011; Habier et al., 2007; and Habier et al., 2009), and the genotype by environment interation (GEI) patterns in the TP and VP (Crossa et al., 2016; Heslot et al., 2013; Jarquin et al., 2014). In general, many studies reported that GS accuracy increases as the number of markers and the size of training population increase (Asoro et al., 2011; Bernardo and Yu, 2007; Heffner et al., 2011b; Lorenz et al., 2012; Spindel et al., 2015). Yet some studies have shown that systematically sampling TP data can yield prediction accuracies nearly equivalent to, or higher than when using all TP data (Ametz, 2015; Akdermir et al., 2015; Hoffstetter et el. 2016; Moser et al., 2010; Schulz-Streek et al., 2013). Similarily, GS accuracy can be improved when the TP consisted of an optimized subset of lines instead of all lines (Ametz, 2015; Akdermier et al., 2015; Cros et al., 2015; Hoffstetter et al., 2016; Isidro et al., 2015; Rutkoski et

al., 2015).

Oakey et al. (2016) sugested that in order to reduce genotyping costs when implementing GS, one could possibly use a smaller set of markers that could give similar or better predictive ability compared to using a large set of markers. Studies report that systematically selecting a subset of markers to build GS models can produce accuracies that are either comparable to using all markers (Moser et al., 2010; Schulz-Streek et al., 2013; Vazquez et al., 2010; and Weigel et al., 2009) or even superior to using all markers (Akdermir et al., 2015; Hoffstetter et al., 2016). Utilizing only a subset of significant set of markers improved GS accuracy in predicting wheat traits including yield, Fusarium head blight resistance and quality traits (Hoffstetter et al., 2016) and in predicting the sire breeding values in dairy cattle breeding (Weigel et al., 2009).

Genotype by environment interaction (GEI) or marker by environment interaction (MEI) effects can have a profound effect on predicting line performance. A fundamental issue for plant breeding is to obtain estimated genetic values for lines using past data that can accurately reflect genetic values that will occur in future environments: GEI diminishes the predictive value of past data. Assessing GEI patterns is critical for making selections in phenotypic selection as well as in GS (Spindel and McCouch, 2016).

Including the GEI term in GS models improved accuracy when TP and VP data were collected in different environments (Burgueno et al., 2012; Cuevas et al., 2017; Heslot et al., 2013; Jarquin et al., 2014; Lopez-Cruz et al., 2015; Tiezzi et al., 2017; Yao et al., 2016). Jarquin et al. (2014) proposed GS models that incorporate MEI effects with environment effects and/or with environment covariates that were based on climatic records or soil characteristics. They reported that GS accuracy using CV within the same set of environments for wheat grain yield improved up to 35% compared to models that only include main effects. Crossa et al., (2016)

assessed models with the effects of accessions, markers, environments, and GEI or MEI in two populations of wheat landraces. They reported that GS models with interaction effects provided better predictions for days to heading and days to maturity in wheat landraces compared to models without interaction terms. Other studies also reported that the GS accuracy was generally higher when using models with GEI effects versus without it (Cuevas et al., 2017; Tiezzi et al., 2016; Yao et al., 2016). Most of these studies only assessed GS accuracies through CV and did not predict breeding values between populations (Cuevas et al., 2017; Tiezzi et al., 2016; Yao et al., 2016). Heslot et al., (2013) used marker effects estimated within environments to cluster the environments so that similar environments could be grouped together to minimize marker by environment interaction (MEI) effects within each cluster of environments. They proposed that the TP could be optimized by deleting data from the least predictive environments, though this study did not perform predictions between clusters of environments.

Our objectives were to 1) assess the effect of optimizing marker sets on the accuracy of genomic prediction between environments within a population and 2) to evaluate the accuracy of genomic predictions between populations, and 3) to assess the accuracy of genomic prediction models that incorporate MEI effects.

MATERIALS AND METHODS

Populations

Two populations were used in this study. The first population is termed the elite population (EP) and was described in Huang et al., (2016). Briefly, it consisted of 273 soft winter wheat lines grown in the 2011-2012 and 2012-2013 seasons in 12 to 14 environments per season (Huang et al., 2016). An augmented design was used within each environment with “Branson” being the check variety. The EP was grown with high and low nitrogen (N) treatments in

Wooster Ohio and Warsaw Virginia. Both N treatments received 20 (kg ha^{-1}) of N in the fall. The low N treatment then received 45 (kg ha^{-1}) of N and high N treatment received 101 (kg ha^{-1}) in the spring (approximately Feekes stage 6). The second population was termed the yield population (YP) and it consisted of 294 lines (Table S1). No lines were shared between the EP and YP except for the check cultivar 'Branson'. In the YP, 95% of the lines came from breeding programs in five states (Kentucky, Maryland, Missouri, Ohio and Virgina). In the EP 66% of the lines came from the same five states with an additional 27% coming from Illinois and Indiana. The YP was grown on the 2013-14 and the 2014-15 seasons each at six locations: Wooster, Custar, and Fremont in Ohio; Warsaw, Virginia; Columbia, Missouri; and Lexington, Kentucky. The YP was grown with high and low N in Wooster Ohio and Columbia Missouri. The N treatments were the same as described for the EP. At each location, the low N treatment was tested with two reps, and high N treatments had one rep except for Columbia Missouri, which had two reps for high N. Fungicide was applied in both populations during flag leaf stage (approximately Feekes stage 9) in all environments to minimize disease. Each rep in each location was conducted as an augmented design. We defined an environment as a year/location/N treatment combination.

Data was collected for both populations for grain yield (YLD, kg ha^{-1}), test weight (TW, kg m^{-3}), height (HGT, cm), heading date (HD, Julian days) and seven quality traits: flour protein content (FP, %), solvent retention capacity for sucrose content (SU, %), lactic acid (LA, %), water content (WA, %), and sodium carbonate (SO, %), softness equivalent (SE, %) and flour yield (FY, %) as described by Huang et al. (2016). Data for the EP was collected in 12, 12, 12, 14, and 5 environments for YLD, TW, HGT, HD, and quality traits respectively (Table 1). For

the YP data were collected in 12, 9, 11, 11, and 4 environments for YLD, TW, HGT, HD, and quality traits respectively (Table 1).

Genotypic Data

The EP and YP were genotyped with the Illumina iSelect array for wheat having approximately 90,000 single nucleotide polymorphism (SNP) markers (Wang et al., 2014) at the USDA-ARS Biosciences Research Laboratory, Fargo, ND, U.S. All marker data was first filtered based on minor allele frequency (<10%) and missing values (>5%, Huang et al., 2016). In EP, the missing marker scores were first imputed using fastPHASE (Scheet and Stephens, 2006), which provided similar results as using the expectation maximization (EM) algorithm (Huang et al., 2016). The EM approach is less computationally intensive (Huang et al., 2016). It was designed for high dimensional marker data set with large amount of missing values and was implemented in the R package 'rrBLUP' (Endelman, 2011). Hence, missing values were imputed in the YP via EM algorithm. In order to avoid a specific chromosome region being overly represented, a SNP tagging approach was used to select markers that were relatively evenly distributed across the genome (Huang et al., 2016). A final set of 3,919 markers were retained and used in the EP analysis. Of the 3,919 EP markers, 3,537 were also scored in the YP and these were used in all YP analyses. These 3,537 were also used within the EP when comparing the EP to the YP.

Structure and linkage disequilibrium (LD) of the populations

A principal component analysis (PCA) of the genotypic data was performed within the EP (Huang et al., 2016), within the YP and with both the EP and YP. A matrix of the LD r^2 values between markers were generated within each population using the *genetics* (Warnes et al., 2013) and LDheatmap (Shin et al., 2006) packages in R. The correlation between the two LD

matrices was assessed using the R package ade4 (Dray et al., 2007). The genetic similarity/diversity between and within EP and YP was estimated by calculating a simple matching coefficient, defined as the proportion of loci that had identical genotype scores between two lines, using marker matrix, implemented in SAS using Proc IML (SAS Institute Inc. 2008).

Phenotypic data and trait stability

A two-stage approach was used to generate the best linear unbiased prediction (BLUPs) for all traits in the EP (Huang et al., 2016) and in YP. We used BLUPs of genotype effects because treating genotype effects as random (e.g. shrinkage applied) or fixed had no impact on GS accuracy despite it representing a double shrinkage in our study (Huang et al., 2016). Such results are not surprising especially for high heritability traits (Piepho et al., 2008). Data from an environment was first adjusted for block effects within each replication based on means of common check variety. This step was done in SAS version 9.2 (SAS Institute Inc. 2008). Then we used the following model to obtain across-environment BLUPs:

$$Y_{ijk} = \mu + G_i + E_j + R_k(E_j) + GE_{ij} + \varepsilon_{ijk}$$

where the effects of genotype (G_i), environment (E_j), replicate nested within environment ($R_k(E_j)$), genotype by environment interaction (GE_{ij}), and residual (ε_{ijk}) were treated as random effects. The analysis was conducted in R software (R Development Core Team, 2008) using the lme4 package (Bates et al., 2015). Heritabilities of all traits were estimated using entry means as described in Huang et al. (2016).

The BLUPs within each cluster of environments were obtained using the same model as the one being used above for across all environments: some clusters consisted of environments

that only had one replication and so the $R_k(E_j)$ term was not used for those environments (Table 1).

We placed the environments into different clusters from each population using Ward's minimum variance criteria (Betran et al., 2003; Huang et al., 2016) such that GEI variance was minimized within a cluster and maximized between clusters. In the analyses over all environments, the GEI variance was partitioned into between and within cluster components.

Trait stability indices of the genotypes were estimated with Eberhart and Russell regression (ERR, Eberhart and Russell, 1966) and Additive Main effects and Multiplicative Interaction methods (AMMI, Zobel et al., 1988), as described by Huang et al. (2016). This was done in SAS and R. For predictions between populations, stability indices were only calculated over all environments. For AMMI stability, we used the first ten, three, six, and one interaction principal components (IPC) for YLD, TW, HGT, and HD in YP, respectively, based on their screeplots: detailed methods and results can be found in Huang et al. (2016). We obtained data on quality traits from just three environments in the YP, so we did not estimate stability of these traits.

Genomic selection

For all GS analysis, the ridge regression BLUP (rrBLUP) model (Endelman 2011; Lopez-Cruz et al., 2015; Crossa et al., 2016) was used and computed in R (R development core team, 2008):

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{v} + \mathbf{e}$$

where μ is the overall mean, $\mathbf{Z}$ matrix represents the marker scores for each genotype, $\mathbf{v}$ is the vector of random marker effects, $\mathbf{e}$ is the residual matrix and $\mathbf{y}$ is a vector of the observations. In

the rrBLUP model, the fixed effect is the intercept and a penalization parameter $\lambda = \sigma^2_e / \sigma^2_u$ was used, in which σ^2_e is the residual variance and σ^2_v is the variance for **v**. This penalization parameter allows for shrinkage while equally estimating the marker effects (Endelman 2011).

The accuracy of GS was estimated 1) overall environments within a population, 2) between clusters of environments within a population, and 3) between populations across environments. This was done using 1) data from all markers (the full set of 3,919 markers after filtering and tagging), 2) subsets of markers selected for significance, or 3) markers selected for stability.

Hence, for between-cluster predictions within each population, traits were analyzed within each separate cluster. The GS models were built using data from one cluster of environments to predict performance of lines in a different cluster.

For prediction within each population, we analyzed data and performed GS both over all environments and within each cluster of environments. For predicting between the EP and YP, we utilized trait values over all environments.

Developing marker subsets within each cluster within a population

To select markers based on significance we conducted association analyses for each trait using the GAPIT package in R (Lipka et al., 2012). Both the PCA and kinship matrices were used. Subsets of markers were formed by selecting markers that were significant at $p < 0.05$ (P0.05 subsets). This was done using data only from environments within a cluster. The *p-value* of 0.05 level was chosen subjectively though Hoffstetter et al. (2016) reported that a subset of markers significant at $p < 0.05$ increased GS accuracy by up to 76% compared to using all markers.

To select subsets of markers based on stability, we first did an association analysis within

each environment. For example, there were 12 environments for YLD in the EP and so we performed 12 independent association analyses for YLD. We used the marker effects to develop a marker by environment (MEI) matrix within a cluster. The MEI was derived using the formula :

$$y_{ij} = M_i + E_j + MEI_{ij}$$

where y_{ij} is the value of the *ith* marker effect in the *jth* environment, M_i is the main effect of the *ith* marker, E_j is the main effect of the *jth* environment and MEI_{ij} is the marker by environment interaction effect. We assessed the percentage of markers to use in a subset by randomly selecting between 5% to 100% of the markers in increments of 5%, running rrBLUP and assessing accuray using CV. Accuracy did not increase with greater than 40% of the markers. Thus we chose to investigate sets with 5% or 40% of the markers. Two methods were used to select subsets of stable markers. We calculated the variance of the MEI effects for each marker and selected the 10% (VAR10 subsets) and 40% (VAR40 subsets) of markers with the lowest MEI variance (Table 2). In the second method, we conducted an AMMI analysis of the MEI matrix and selected the 10% of the markers that had low principal components scores (AMM10 subsets). The AMMI analyses involved conducting a singular value decomposition of the MEI matrix to extract the IPC scores for each marker. The stability of each marker was obtained by summing the absolute value of the first n^{th} IPC scores for that marker: a low sum indicates high stability. The number of IPC scores used for calculating the marker stability for a trait was chosen based on the scree plot and ranged from two to eight. These VAR10, VAR40, and AMM 10 subsets (Table 2) were formed using the MEI matrices based on marker effects from environments only within one cluster: thus the phenotypic data used to select stable markers within one cluster of environments was independent of the phenotypic data from the other

cluster.

For each trait and set of environments, we constructed one subet of markers based on both their significance and their stability (PVAR10 subset). For the PVAR10 subset, we first selected 15% of the markers with the largest effects (absolute value of allelic effect) from the association analysis. Then we selected the markers with the lowest MEI variance, producing a subset with 10% of the total markers. The subsets P0.05, VAR10, PVAR10, and VAR40 contained either 5%, 10% or 40% of the markers based on different selection criteria. We also generated marker subsets containing a random selection of 5% (RAN5 subset) or 40% (RAN40) of the markers to serve as controls.

Evaluating the optimization methods

Within each population and for each trait, we first used all phenotypic and genotypic data as a TP, and then set up TPs that used the different markers subsets (using RAN5, RAN40 VAR10, VAR40, AMMI10, P0.05, or PVAR10). Within a population we conducted 10-fold cross validation to assess the accuracy of GS as described by Huang et al. (2016). The cross validations were run 500 times with the RAN5 and RAN40 subsets. The difference in accuracy was minimal between these 500 iterations so we selected 10 runs for the other data subsets in order to be less computationally intensive.

Use of MEI in predictions

For the analyses between populations, we used BLUPs over all environments from each population. The ability of data from one population to predict the phenotypes in the other population was assessed. Either all data, or a subset of TP data, from one population was used to build GS model that was then used to calculate the GEBVs of lines from a different population.

Cross-validation was not used in these analyses, as each population was independent of the other. To predict between two populations, by using all data, we also assessed two additional models that incorporate either environment and/or MEI effects (Lopez‐Cruz et al., 2015; Crossa et al., 2016). The first model incorporated genomic and environment main effects:

$$y_{ij} = \mu + g_i + e_j + \varepsilon_{ij}$$

where y_{ij} is the observation for i*th* genotype at the j*th* environment, e_j is the effect of the j*th* environment, and ε_{ij} is the residual effects associated with the ij*th* observation. In this model, $g_i = \sum_{k=1}^{m} z_{ik}u_k$, in which z_{ik} is the score (-1,0,1) for the k*th* marker for the i*th* genotype and u_k is the effect for the k*th* marker (Lopez-Cruz et al., 2015). In this model, the u_k were assumed to be the main marker effects and were the same across environments. The intercept and e_j were fixed effects in this model (Lopez-Cruz et al., 2015; Crossa et al., 2016).

The second model allows borrowing information across environments for estimating marker effects across different environments and it included MEI effects (Lopez-Cruz et al., 2015; Crossa et al., 2016):

$$y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij}$$

The notation for y_{ij}, g_i, e_j, and ε_{ij} are the same as the aforementioned main effects model. The interaction term ge_{ij} is the interaction of genomic values of the i*th* individual and the j*th* environment. In this model, the marker effects consisted of two parts: the marker main effects which is a constant across all environments and an environment-specific effect for each marker (Crossa et al., 2016). The fixed effects in this model included the intercept and environment main effects. The random effects included marker effects, genomic by environment interaction effects

and the residual effects (Lopez-Cruz et al., 2015). The $g+e$ main effects model and $g+e+g\times e$ model were all computed in R using the package BGLR (de los Campos et al., 2015).

RESULTS

Phenotypic analysis

Analysis of the EP phenotypic data was reported in Huang et al. (2016). In both the EP and YP heritability was > 0.73 for all traits and the effect of environment, genotype, and GEI were significant ($p < 0.05$) for all traits (Huang et al., 2016; Table 3).

Clustering of environments within each population

We identified three clusters of environments for YLD and HD in the YP and for YLD, TW and HGT in EP. Two clusters of environments were identified for HGT in the YP and for HD in the EP (Table 1). In the YP, there was one cluster of environments and one outlier for TW. Some environments were outliers and were excluded from the analyses between clusters (Table 1). In the EP, the correlation of the phenotypic BLUPs between the most divergent clusters of environments was low for YLD ($r = 0.34$) and TW ($r = 0.33$) but was high for HGT ($r = 0.78$) and HD ($r = 0.82$); these correlations in YP were also low for YLD ($r = 0.27$) (Table 4). The percentage of GEI variance attributed to between-clusters was greater than 50% for all traits except YLD in both populations (Table S2), suggesting the clustering was least effective for YLD. The GEI variance as a percentage of genotype+GEI variance was also greater for YLD than for the other traits (Table S2).

Population structure and relationship between the EP and YP

The population structure in the YP was visually similar to that in EP, as each was characterized by one large group of lines and one smaller group that was mainly composed of

lines from the University of Missouri (Fig. 1a). Results from principal component analysis for both populations together showed lines from both populations were interspersed in all regions of the graph (Fig. 1b). The genetic similarities via simple matching coefficient between the two populations (0.60) and within the EP (0.60) or within YP (0.60) were nearly identical, suggesting that these two populations are very similar. The LD matrices of the two populations were tested via mental test and they were significantly correlated (r = 0.87, *p-value* < 0.01) indicating that they have a similar LD pattern among the common markers.

GS accuracy between clusters of environments within a population using subsets of markers

There was a variable degree of commonality between the marker subsets developed for any trait. For all traits, the scheme used to create the AMM10 and VAR10 marker subsets provided almost identical subsets of markers (results not shown). This indicates that using total MEI variance or the AMMI approach produced similar sets of markers.

The GS accuracy between clusters of environments varied by population, trait, the subset of data being used, and the particular comparison (Table 4). The accuracy of GS predictions between clusters using all TP data (e.g., all markers and all lines) was on average 7.2% lower than the phenotypic correlation, ranging from 39.0% lower to 10.9% greater (Fig. 2; Fig. S1; Table 4). For YLD, HGT and HD, the average GS accuracy between clusters of environments was 16.2%, 9.0%, and 3.1% lower than the average phenotypic correlation between clusters (Table 4). The average phenotypic correlation and GS accuracy between-cluster were nearly equal for TW (Table 4). We assessed the relationship of the correlation of phenotypic means between environment-clusters with the ability of a GS model built using data from one cluster of environments to predict the phenotypes of a different cluster. The accuracy of GS predictions

between clusters was highly associated (r = 0.97) with the correlation of phenotypes between clusters.

For most traits and between-cluster comparisons, the Ran5 subsets gave the lowest GS accuracy (Table 4). The accuracy with the Ran40 subsets was 16.9% better than the Ran5 subsets when averaged over all comparisons, and its accuracy was very similar to the GS accuracy using all markers (Table 4).

The markers in the VAR10, VAR40, and AMM10 subsets were selected based on the stability of their effects across the environments within a particular cluster. For all comparisons, GS accuracy for VAR10 and AMM10 were nearly identical: when averaged over all comparisons, both were 9% less accurate than the VAR40 subsets (Table 4). The accuracy of the VAR40 subsets was similar to the accuracy obtained with the RAN40 subsets and produced accuracies that were only 7% less accurate than using all data when averaged over all comparison.

The P0.05 and PVAR10 subsets of markers were selected considering the magnitude of marker effects (P0.05) as well as their stability (PVAR10). Averaged over all comparisons, the accuracy of the P0.05 subsets was 5.1% greater than using data from all markers, while the accuracy of the PVAR10 subsets were 7.7% more accurate that using all data (Table 4). The increase in accuracy for these two subsets was most notable for YLD where P0.05 and PVAR10 were on average 14.8% and 17.6% more accurate than using all marker data. These marker subsets had minimal impact on increasing between-cluster accuracy compared to using all data for TW, HGT, and HD (Table 4). The P0.05 and PVAR10 marker subsets had a similar number of markers as the RAN5 subset but produced greater GS accuracy than the RAN5 subsets or when using all markers (Table 4).

Between Populations: GS accuracy using all TP data or subsets of data

From the results within the EP, we determined that the P0.05 (significant markers) subset of markers was the best and thus we only used all data and this subset of data in the between-population analyses. For 13 out of the 19 traits and trait stability indices, using all data from one population to predict the phenotypes of lines in the other population gave higher accuracy than using the best subsets of marker data P0.05 approach (Table S3), thus only results from using all TP data will be discussed.

When using either all data or subsets of marker data, GS accuracy between populations was considerably lower than the accuracy between-clusters of environments within a population (Fig 2, Tables 4, 5) being 95%, 11%, 28%, and 41% lower for YLD, TW, HGT, and HD, respectively. The between-population GS accuracy for a trait was similar (correlation $r = 0.89$, calculated using values in column 4 and 5 in Table 5) whether the EP or YP were used as the TP. The average accuracy of GS between populations was negligible for YLD, ATW and AHGT and was <0.25 for AYLD, ATW, AHD, BYLD, BTW, BHD and FP (Table 5). On average, the stability estimated using the regression method gave higher between-population GS accuracy than using the AMMI method (Table 5).

The GS accuracy between populations was associated with GS accuracy within each population obtained using cross validation (Table 5, using values in column 2 or 3 correlating to values in column 6). The correlation of the average accuracy for within-population prediction and the average accuracy for between-population prediction was 0.91 (Fig. 3): if accuracy for a trait was low within a population then the data from that population had a low ability to predict phenotypes in the other population (Fig. 3). The accuracy between populations was also associated with the correlation of genetic effects of significant markers between the two

populations (Table 5): prediction accuracy was low when the correlation of marker effects was low.

Accuracy of GS using the g+e and g+e+g×e models for between population prediction

The model including both marker and environment main effects and also a model containing those effects plus the MEI effects were assessed for between-population accuracy for each trait. These two models generally produced lower between-environment accuracy for TW, HGT and HD (Table 6) than using the standard GS model rrBLUP (Table 5). Compared to using the rrBLUP model, the $g+e$ and $g+e+g\times e$ models gave the largest improvement of GS accuracy for YLD when using YP to predict EP, but these accuracies remained low ($r < 0.15$, Table 6).

DISCUSSION

One of the most fundamental issues in plant breeding is to obtain estimates of genetic values from one data set that can predict phenotypic performance in future environments. This issue remains whether using phenotypic data or GEBVs (Spindel and McCouch, 2016). We investigated whether using subsets of marker data could improve the ability of GS to predict performance in different environments. Markers were selected based on their effects, the stability of their effects, or both criterion.

Our results show that when the correlation of phenotypes between two environments was low, then the ability of data from one environment to predict the other using GS was also low, as has been reported before (Dekkers, 2007). On average, for all traits across all between-cluster comparisons within a population, randomly selecting markers (Ran5, Ran40), or subsetting markers based on their stability alone (AMMI10, Var10 and Var40) did not improve the GS accuracy compared to using all the TP data (Table 4). These results suggest that selecting markers based solely on stability of marker effects will not improve GS accuracy of between-

environment predictions. This may be because non-informative markers with no effect on the trait can be stable if they may have little to no effect in any environment, and thus included in these subsets along with predictive markers that do have an effect: we later confirmed that using significant markers could increase GS accuracy (Fig. 2; Fig. S1).

The results indicate that within a population, a GS model built using data from one set of environments can predict the performance of lines in an independent set of environments. The accuracy of GS for YLD between-environments within a population for all traits was significantly improved, relative to using all TP data, when using subsets of markers. The GS accuracy increased 8.1% in the YP and 14.7% in EP on average by using the P0.05 subsets, and by 17.6% in the EP by using the PVAR10 subset (Fig. 2; Fig. S1). Minimal increase in GS accuracy was observed for the other traits (Table 4). YLD was the trait that had the greatest GEI, suggesting that greater gains in GS accuracy from using subsets of significant and stable markers may be realized for traits with extensive GEI than for traits with low GEI. Our results were similar to the results in a previous wheat study (Hoffstetter et al., 2016) where using subsets of significant markers for YLD, FHB resistance and quality traits increased GS accuracy by 41% to 76% compared to using all markers. Hoffstetter et al. (2016) also reported GS accuracy within and between environments was greatest with a TP comprised of a subset of lines with low GEI variance and significant markers. In animal breeding, it was also reported that when using a subset of markers with largest effects, GS accuracy was either similar to or was somewhat improved compared to using all markers in the TP (Abdollahi-Arpanahi et al., 2014; Weigel et al., 2009; Vazquez et al., 2010; Moser et al., 2010). It is likely that by using only markers with stable significant effects across environments, that we are excluding markers from the least predictive portion of the genome.

Using data from one population to predict the performance of lines in a different, yet highly related, population that was phenotyped in different environments has proven to be more difficult than predicting the performance of the same lines in different environments (Beaulieu et al., 2014; Lado et al., 2016; Jarquin et al., 2014; Schulz-Streek et al., 2013). In our study, GS accuracy between the EP and YP populations was lower than GS accuracy between cluster of environments within the same population (Fig.2, Tables 5). In addition, using data subsets did not improve GS accuracy between populations compared to using all TP data. Data from one population was poor at predicting YLD, FP, and AMMI-based trait stability in the other population (Table 5).

The EP and YP were closely related to each other as shown by the PC graph (Fig. 1b) and the simple matching coefficients. The LD matrices of the two populations were highly correlated suggesting that there could also be a similar LD pattern between markers and QTL in each population. This is one requirement in order for GS to work and this consistency between populations of LD between markers and QTL appeared to occur for TW, HGT, HD, SU, LA, WA and SO. This suggests that the LD patterns between markers and QTL for YLD and FP would also likely be consistent between the two populations. A second requirement for GS to work is for QTLs to have similar effects in each population or set of environments. In our study, the ability to use data from one population to predict performance in the other was related to the correlation of the marker effects estimated in the two populations (Table 5). The correlation of markers effects between the EP and YP for YLD and FP were low. A low correlation of marker effects would lead to high GEI which would lead to a low phenotypic correlation between environments. Unsurprisingly, as shown here and by others (Dekkers, 2007), when the correlation of phenotypes between environments are low GS will not be very predictive. We also

report that GS accuracy within a population was strongly associated with GS accuracy between populations (Fig. 2; Fig. S1; Table 4): if GS accuracy was low within a population then GS accuracy between populations was also low.

In our study, low GS accuracy between populations for YLD stability and several other stability parameters indicate that stability estimated in one set of environments is not very predictive of stability in another set of lines and environments. The ERR stability indices were predicted with a higher accuracy than the AMMI-based stability indices. This could be because the ERR measures are based on environmental indices (i.e. average trait value within an environment) whose values and impact may be more repeatable across sets of environments, while the AMMI interaction PC scores may be very specific to the set of environments from which they were calculated (Eberhart and Russell, 1966; Zobel et al., 1988). Hickey et al. (2015), suggested that for closely related populations, effective GS accuracy could be obtained with a small number of markers (200-500) and perhaps 1,000 lines. Our populations are considerably smaller so maybe greater accuracy could have been attained if the size of each population was increased by adding lines that are closely related.

For between population predictions, our results are well aligned with those of others who have used a TP and VP that were composed of different lines and/or were tested in different environments. Jarquin et al. (2014) reported that GS accuracy for wheat YLD was greater within a population than between populations that consist of different lines tested in different environments. It was reported that the GS accuracy through cross validation was higher than those obtained from predicting between populations, as is also shown in this study. Jarquin et al. (2016) also compared GS accuracies for YLD by predicting through cross validation versus by predicting between locations, years and location-year combination schemes, and they have

confirmed that GS accuracy via cross validation is higher than that using the other schemes. Michel et al. (2016b) assessed GS accuracy with data from lines from five breeding cycles of winter wheat. They reported the average GS accuracy via cross-validation within a population was 0.43 for YLD and 0.57 for FP, but was just 0.29 and 0.41 for YLD and FP, when the TP and VP consisted of different lines tested in different environments. Battenfield et al. (2016) reported GS accuracy for wheat TW between populations tested in different years ranged from 0.12 to 0.35 while GS accuracy was 31.8% greater within a population. Dawson et al. (2013) assessed the ability of GS for wheat YLD using a TP of lines from a previous year to predict performance of VP lines in future years using data on 622 lines from CIMMYT that were phenotyped over 17 years. When the TP and VP were tested under different environments, the adjusted GS accuracy for YLD ranged from -0.10 to 0.80 across 17 years of testing environments (Dawson et al., 2013). Battenfield et al. (2016) reported that when the TP and VP consists of lines grown in different environments and had a few lines in common that GS accuracy ranged from 0.33 to 0.44 for FY and FP but ranged from 0.82 to 0.90 within population.

When performance of new lines needs to be predicted, Jarquin et al. (2014) recommended using GS models with GEI terms. They assessed two scenarios: the first one was to predict the value of lines that were tested in some but not all environments; and the second one was to predict the value of new lines that were not tested in any environments (Jarquin et al., 2014). Their model worked better under the first scenario, which allowed borrowing information from one set of environments for the same line to predict its performance in a different set of environments, while the later uses data from one set of lines to predict performance of a different line in different environments (Crossa et al., 2014; Jarquin et al., 2014; Jarquin et al., 2016; Saint-Pierre et al., 2016). Cuevas et al. (2017) reported that using models incorporating the g×e

term always produced higher GS accuracy than models without g×e for one maize and four wheat CIMMYT data sets. Yet in our study, utilizing the $g+e$ and $g+e+g\times e$ models did not significantly improve GS accuracy for most traits (Table 6). Similar to our results, Dawson et al. (2013) reported that modeling MEI effects did not improve GS accuracy for YLD across a wide range of environments compared to the model without GEI. They suggested that there were inconsistent marker effects among the mega environments.

The same trait being evaluated in two environments could be viewed as two traits and thus selection based on data from one environment to improve that trait in another environment could be viewed as indirect selection (Falconer, 1981). Our results show that indirect selection using GS should be effective where direct selection using phenotypes is effective. The advantage of GS then is that a cycle of GS can be completed in much less time than a cycle of phenotypic selection (Goddard and Hayes, 2007; Heffner et al., 2010; Jannink et al., 2010; Meuwissen et al., 2001). Gain per unit of time is a key measure of plant breeding efficiency. One of the major limitations to plant breeding efficiency is the amount of time required to complete acycle of phenotypic selection. This problem could be addressed and improved by using GS, as GS could rapidly improve selection efficiency per unit time and cost (Meuwissen et al., 2001; Heffner et al., 2009; Jannink et al., 2010 ; Sallam and Smith 2016). The results of this study suggest that GEI may now be one of the major factors limiting genetic gain per unit of time.

Acknowledgments

We sincerely thank the members at Dr. C. Sneller lab for their help with field data collection in Ohio. The funding support was from Triticeae Coordinated Agricultural Project (2011-68002-30029) of the USDA National Institute of Food and Agriculture.

Author contribution statement:

MH wrote the manuscript and analyzed the phenotypic and genotypic data. MH also collected the data at Ohio location. BW, CG, DS, and AM contributed to data collection in Virginia, Kentucky, and Missouri as well as editing the manuscript. GB and PT assisted the genotypic data for the yield panel population. CS led the project, edited the manuscript and helped with data analysis.

Compliance with Ethical Standards

This research complies with the current laws of the United States of America

Conflict of Interest

The authors of this study declare that there is no conflict of interest for this study.

References

Abdollahi‐Arpanahi, R., A. Pakdel, A. Nejati‐Javaremi, M. Moradi Shahrbabak, G. Morota, B. Valente, et al. 2014. Dissection of additive genetic variability for quantitative traits in chickens using SNP markers. Journal of Animal Breeding and Genetics 131: 183-193.

Akdemir, D., J.I. Sanchez and J.-L. Jannink. 2015. Optimization of genomic selection training populations with a genetic algorithm. Genetics Selection Evolution 47: 1.

Ametz, C., 2015 Genomic selection in bread wheat. PhD Dissertation, http://permalink.obvsg.at/bok/AC10777353

Arruda, M., A. Lipka, P. Brown, A. Krill, C. Thurber, G. Brown-Guedira, et al. 2016. Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (Triticum aestivum. Molecular Breeding 36: 1-11.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott and J.-L. Jannink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. The Plant Genome 4: 132-144.

Bates, D., M. Maechler, B. Bolker, S. Walker, R.H.B. Christensen, H. Singmann, et al. 2015. Package 'lme4'.

Battenfield, S.D., C. Guzmán, R.C. Gaynor, R.P. Singh, R.J. Peña, S. Dreisigacker, et al. 2016. Genomic Selection for Processing and End-Use Quality Traits in the CIMMYT Spring Bread Wheat Breeding Program. The Plant Genome.

Beaulieu, J., T.K. Doerksen, J. MacKay, A. Rainville and J. Bousquet. 2014. Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC genomics 15: 1048.

Bernardo, R. and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Science 47: 1082-1090.

Betran, F., D. Beck, M. Bänziger and G. Edmeades. 2003. Genetic analysis of inbred and hybrid grain yield under stress and nonstress environments in tropical maize. Crop Science 43: 807-817.

Burgueño, J., G. de los Campos, K. Weigel and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Science 52: 707-719.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueno, J.L. Araus, et al. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713-724.

Crossa, J., D. Jarquin, J. Franco, P. Pérez-Rodríguez, J. Burgueño, C. Saint-Pierre, et al. 2016. Genomic Prediction of Gene Bank Wheat Landraces. G3: Genes| Genomes| Genetics: g3. 116.029637.

Cuevas, J., J. Crossa, O.A. Montesinos-López, J. Burgueño, P. Pérez-Rodríguez and G. de los Campos. 2017. Bayesian Genomic Prediction with Genotype× Environment Interaction Kernel Models. G3: Genes| Genomes| Genetics 7: 41-53.

Dawson, J.C., J.B. Endelman, N. Heslot, J. Crossa, J. Poland, S. Dreisigacker, et al. 2013. The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crops Research 154: 12-22.

Dekkers, J. 2007. Prediction of response to marker ‐ assisted and genomic selection using selection index theory. Journal of Animal Breeding and Genetics 124: 331-341.

2011. Genomic selection in tree breeding: testing accuracy of prediction models including dominance effect. BMC proceedings, BioMed Central.

Denis, M. and Bouvet, J.M. 2011. Genomic selection in tree breeding: tes-ting accuracy of prediction models including dominance effect. In BMC Proceedings, 5(7):13

Dray, S., A. Du-four and J. Thiolouse. 2014. ade4. CRAN| L"(K).

Eberhart, S.t. and W. Russell. 1966. Stability parameters for comparing varieties. Crop science 6: 36-40.

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

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4: 250-255.

Falconer, D.S. and T.F. Mackay. 1996. Introduction to quantitative genetics. Harlow, UK: Longman.

Goddard, M.E. and B. Hayes. 2007. Genomic selection. Journal of Animal Breeding and Genetics 124: 323-330.

Habier, D., R. Fernando and J. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389-2397.

Habier, D., R.L. Fernando and J.C. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182: 343-353.

Hayes, B., P. Bowman, A. Chamberlain and M. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science 92: 433-443.

He, S., A.W. Schulthess, V. Mirdita, Y. Zhao, V. Korzun, R. Bothe, et al. 2016. Genomic selection in a commercial winter wheat population. Theoretical and Applied Genetics 129: 641-651.

Heffner, E.L., J.-L. Jannink, H. Iwata, E. Souza and M.E. Sorrells. 2011. Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Science 51: 2597-2606.

Heffner, E.L., J.-L. Jannink and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome 4: 65-75.

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

Heffner, E.L., M.E. Sorrells and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Science 49: 1-12.

Heslot, N., J.-L. Jannink and M.E. Sorrells. 2013. Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Science 53: 921-933.

Heslot, N., H.-P. Yang, M.E. Sorrells and J.-L. Jannink. 2012. Genomic selection in plant breeding: a comparison of models. Crop Science 52: 146-160.

Hoffstetter, A., A. Cabrera, M. Huang and C. Sneller. 2016. Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. G3: Genes| Genomes| Genetics 6: 2919-2928.

Huang, M., A. Cabrera, A. Hoffstetter, C. Griffey, D. Van Sanford, J. Costa, et al. 2016. Genomic selection for wheat traits and trait stability. Theoretical and Applied Genetics: 1-14.

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot and M.E. Sorrells. 2015. Training set optimization under population structure in genomic selection. Theoretical and applied genetics 128: 145-158.

Jannink, J.-L., A.J. Lorenz and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics 9: 166-177.

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt, J. Lorgeou, et al. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and applied genetics 127: 595-607.

Jarquin, D., J. Specht and A. Lorenz. 2016. Prospects of genomic prediction in the USDA Soybean Germplasm Collection: Historical data creates robust models for enhancing selection of accessions. G3: Genes, Genomes, Genetics 6: 2329-2341.

Jiang, Y., A.W. Schulthess, B. Rodemann, J. Ling, J. Plieske, S. Kollers, et al. 2016. Validating the prediction accuracies of marker-assisted and genomic selection of Fusarium head blight resistance in wheat using an independent sample. Theoretical and Applied Genetics: 1-12.

Lado, B., P.G. Barrios, M. Quincke, P. Silva and L. Gutiérrez. 2016. Modeling genotype× environment interaction for genomic selection with unbalanced data from a wheat breeding program. Crop Science 56: 2165-2179.

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

Lopez-Cruz, M., J. Crossa, D. Bonnett, S. Dreisigacker, J. Poland, J.-L. Jannink, et al. 2015. Increased prediction accuracy in wheat breeding trials using a marker× environment interaction genomic selection model. G3: Genes| Genomes| Genetics: g3. 114.016097.

Lorenz, A., K. Smith and J.-L. Jannink. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Science 52: 1609-1621.

Lorenz, A.J. and K.P. Smith. 2015. Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. Crop Science 55: 2657-2667.

Meuwissen, T., B. Hayes and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Michel, S., C. Ametz, H. Gungor, B. Akgöl, D. Epure, H. Grausgruber, et al. 2016. Genomic assisted selection for enhancing line breeding: merging genomic and phenotypic selection in winter wheat breeding programs with preliminary yield trials. Theoretical and Applied Genetics: 1-14.

Michel, S., C. Ametz, H. Gungor, D. Epure, H. Grausgruber, F. Löschenberger, et al. 2016. Genomic selection across multiple breeding cycles in applied bread wheat breeding. Theoretical and Applied Genetics 129: 1179-1189.

Moser, G., M.S. Khatkar, B.J. Hayes and H.W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genetics Selection Evolution 42: 37.

Oakey, H., B. Cullis, R. Thompson, J. Comadran, C. Halpin and R. Waugh. 2016. Genomic selection in multi-environment crop trials. G3: Genes| Genomes| Genetics 6: 1313-1326.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, et al. 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. The Plant Genome 5: 103-113.

R Development Core Team, 2008 R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/

Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J.-L. Jannink and M. Sorrells. 2012. Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. The Plant Genome 5: 51-61.

Rutkoski, J., J. Poland, S. Mondal, E. Autrique, L.G. Pérez, J. Crossa, et al. 2016. Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. G3: Genes| Genomes| Genetics 6: 2799-2808.

Rutkoski, J., R. Singh, J. Huerta-Espino, S. Bhavani, J. Poland, J.-L. Jannink, et al. 2015. Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. The Plant Genome 8.

Rutkoski, J.E., E.L. Heffner and M.E. Sorrells. 2011. Genomic selection for durable stem rust resistance in wheat. Euphytica 179: 161-173.

Rutkoski, J.E., J.A. Poland, R.P. Singh, J. Huerta-Espino, S. Bhavani, H. Barbier, et al. 2014. Genomic selection for quantitative adult plant stem rust resistance in wheat. The plant genome 7.

Saint Pierre, C., J. Burgueño, J. Crossa, G.F. Dávila, P.F. López, E.S. Moya, et al. 2016. Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. Scientific reports 6: 27312.

Sallam, A., J. Endelman, J.-L. Jannink and K. Smith. 2015. Assessing genomic selection prediction accuracy in a dynamic barley breeding population. The Plant Genome 8.

Sallam, A.H. and K.P. Smith. 2016. Genomic Selection Performs Similarly to Phenotypic Selection in Barley. Crop Science 56: 2871-2881.

SAS Institute Inc., 2008 SAS/STAT User's Guide, Version 9.2. SAS Institute Inc, Cary, NC

Scheet, P. and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics 78: 629-644.

Schulz‐Streeck, T., J.O. Ogutu, A. Gordillo, Z. Karaman, C. Knaak and H.P. Piepho. 2013. Genomic selection allowing for marker‐by‐environment interaction. Plant Breeding 132: 532-538.

Shin, J.-H., S. Blay, B. McNeney and J. Graham. 2006. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. Journal of Statistical Software 16: 1-10.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redoña, et al. 2015. Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS genetics 11: e1004982.

Spindel, J.E. and S.R. McCouch. 2016. When more is better: how data sharing would accelerate genomic selection of crop plants. New Phytologist 212: 814-826.

Tiezzi, F., G. de Los Campos, K.P. Gaddis and C. Maltecca. 2017. Genotype by environment (climate) interaction improves genomic prediction for production traits in US Holstein cattle. Journal of dairy science.

Vazquez, A., G. Rosa, K. Weigel, G. De los Campos, D. Gianola and D. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. Journal of dairy science 93: 5942-5949.

Wang, S., D. Wong, K. Forrest, A. Allen, S. Chao, B.E. Huang, et al. 2014. Characterization of polyploid wheat genomic diversity using a high ‐ density 90 000 single nucleotide polymorphism array. Plant biotechnology journal 12: 787-796.

Warnes, G. and F. Leisch. 2005. Genetics: population genetics. R package version 1: 1-5.

Weigel, K., G. De Los Campos, O. González-Recio, H. Naya, X. Wu, N. Long, et al. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of dairy science 92: 5248-5257.

Yao, C., G. de los Campos, M. VandeHaar, D. Spurlock, L. Armentano, M. Coffey, et al. 2016. 0307 Use of marker× environment interaction whole genome regression model to incorporate genetic heterogeneity for residual feed intake, dry matter intake, net energy in milk, and metabolic body weight in dairy cattle. Journal of Animal Science 94: 146-147.

Zhang, X., A. Sallam, L. Gao, T. Kantarski, J. Poland, L.R. DeHaan, et al. 2016. Establishment and optimization of genomic selection to accelerate the domestication and improvement of intermediate wheatgrass. The Plant Genome 9.

Zhao, Y., M. Mette, M. Gowda, C. Longin and J. Reif. 2014. Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity 112: 638-645.

Zobel, R.W., M.J. Wright and H.G. Gauch. 1988. Statistical analysis of a yield trial. Agronomy Journal 80: 388-393.

Figure Captions

Figure 1. Plot of the first two principal component scores for a) the lines within the YP and b) within the EP and YP populations using data from 3,537 SNP markers. The plot of principal component analysis for within the EP was reported by Huang et al. (2016).

Figure 2. Plot of the GS accuracy between clusters of environment within a population for grain yield, using subset of data (P0.05=set of only significant marker; PVAR10=set of significant and stable markers) compared to using all marker data within the elite panel (EP) and yield panel (YP) population. *** represents an estimate of accuracy that is significantly ($p < 0.001$) different from the accuracy obtained using all marker data. The *p-values* were adjusted to account for multiple comparisons within each population.

Figure 3. Plot of genomic selection accuracy between populations using data from all traits regressed onto the average accuracy of the trait within the elite and yield populations.

Table 1. Cluster assignment of environments using Wards minimum variance method and the matrix of genotype × environment interaction values.

Pop[†]	Trait[‡]	Cluster	Environments
EP	YLD	1	13VAL[#], 13VAM
		2	12KYM, 12MDM, 12MOM, 13MOM, 13OWM
		3	12OWL, 12OWM, 12VAM
		Outlier	12VAL, 13OWL
	TW	1	13OWL, 13OWM
		2	12KYM, 13MOM, 12MDM, 12MOM, 12VAL, 12VAM, 13VAL, 13VAM
		3	12OWL, 12OWM
		Outlier	None
	HGT	1	13VAL, 13VAM
		2	12KYM, 12MOM, 12OWL, 12OWM, 12VAM, 13MOM
		3	13OWL, 13OWM
		Outlier	12MDM, 12VAL
	HD	1	12OWL, 12OWM, 13MOM, 13ONM, 13OVM, 13OWL, 13OWM
		2	12KYM, 12MDM, 12MOM, 12VAL, 12VAM, 13VAL, 13VAM
		Outlier	None
YP	YLD	1	15OWL, 15OWM
		2	14OWL, 14OWM, 15OVM, 14VAM, 15VAM
		3	14MOL, 14MOM, 15MOL, 15MOM
		Outlier	15ONM
	TW	1	14OWL, 14VAM, 15ONM, 15OVM, 15OWL, 15OWM, 15VAM
		Outlier	14OWM, 14KYM
	HGT	1	14OWM, 15OWL, 15OWM
		2	14MOL, 14MOM, 14VAM, 15VAM
		Outlier	14OWL, 14KYM, 15MOL, 15MOM
	HD	1	15MOL, 15MOM
		2	14KYM, 14OWL, 14OWM, 14VAM, 15OWL, 15OWM, 15VAM
		3	14MOL, 14MOM
		Outlier	None

[†] Pop: refers to the population elite panel (EP) or yield panel (YP).

[‡] Traits = YLD = yield (kg ha^{-1}), TW = test weight (kg m^3), HGT = plant height (cm), HD = heading date (Julian days)

[#] The environment codes consist of the last two numbers of the year, followed by a two letter abréviation for the location (VA = Virginia, KY = Kentucky, MD = Maryland, MO = Missouri, OW = Ohio Wooster, ON = Ohio Custar, OV = Ohio Fremont Station, followed by "L" for low nitrogen and "M" for moderate nitrogen.

Table 2. Description of the subsets of training population data.

Data Subsetting Method	Description
Ran5	Randomly select 5% of the markers
Ran40	Randomly select 40% of the markers
Var40	Select 40% of the markers with the lowest MEI variance
Var10	Select 10% of the markers with the lowest MEI variance
AMM10	Select 10% of the markers with the lowest MEI variance as estimated using AMMI
P0.05	Select all markers that were significant in the association analysis at $p < 0.05$
PVar10	Select 15% markers that were most significant, and then chose the ones with the lowest MEI for a total of just 10% of all markers. based on AMMI

Table 3. Variance components, their significance, and entry mean broad-sense heritability (H) from the analysis of variance of the yield panel (YP).

Trait	Genotype	Environment	Rep(Env)	Gen*Env	Residual	H
YLD[†]	64220.0**	682555.0*	10435.0	64745.0**	179122.0	0.83
TW	175.8**	812.1	150.2	526.9**	103.0	0.73
HGT	31.9**	29.9*	0.8	4.7**	19.2	0.96
HD	2.5**	40.9*	0.0	0.9**	2.0	0.94
FP	0.1**	1.0	-	0.2**	-	0.75
SU	9.7**	0.2	-	5.9**	-	0.89
LA	92.6**	68.5	-	36.2**	-	0.93
WA	2.9**	1.1	-	1.3	-	0.91
SO	7.9**	2.0	-	2.6**	-	0.94
SE	6.9**	0.3	-	3.5**	-	0.91
FY	1.4**	0.5	-	1.0**	-	0.88

[†] YLD = yield (kg ha^{-1}), TW = test weight (kg m^3), HGT = plant height (cm), HD = heading date (Julian days), Solvent retention capacity (%) for FP= Flour protein, SU= Sucrose content, LA= Lactic acid, WA= Water content, SO= Sodium carbonate, SE= Softness equivalent, and FY= Flour yield.

* and ** indicate significance at $p < 0.05$ and $p<0.01$ levels, respectively

Table 4. Accuracy of GS between cluster of environments (data from one cluster was used to predict performance in the other cluster) within the Elite Panel (EP) and Yield Panel (YP). The phenotypic correlation (r) between clusters is shown as is accuracy using all training population data (All Data) and subsets of data as described in Table 2.

Popu-lation	Trait[†]	Data set[‡]	C1[§] to C2	C2 to C1	C1 to C3	C3 to C1	C2 to C3	C3 to C2	Avg
EP	YLD	r	0.51	0.51	0.34	0.34	0.41	0.41	0.42
	YLD	All Data	0.45	0.49	0.25	0.27	0.31	0.25	0.34
	YLD	Ran5	0.37	0.42	0.18	0.22	0.18	0.17	0.26
	YLD	Ran40	0.44	0.48	0.24	0.26	0.29	0.23	0.32
	YLD	Var40	0.44	0.46	0.24	0.28	0.23	0.24	0.32
	YLD	Var10	0.38	0.42	0.19	0.24	0.15	0.18	0.26
	YLD	AMM10	0.38	0.41	0.19	0.24	0.15	0.20	0.26
	YLD	P0.05	0.49	0.48	0.32	0.33	0.36	0.37	0.39
	YLD	PVar10	0.49	0.51	0.32	0.33	0.38	0.38	0.40
	YLD	Var40-L	0.50	0.47	0.30	0.32	0.33	0.37	0.38
	YLD	Var10-L	0.48	0.49	0.31	0.32	0.38	0.39	0.40
	YLD	P0.05-L	0.50	0.50	0.35	0.34	0.46	0.43	0.43
	YLD	PVAR10-L	0.51	0.52	0.34	0.33	0.41	0.41	0.42
YP	YLD	r	0.50	0.50	0.27	0.27	0.48	0.48	0.42
	YLD	All Data	0.47	0.44	0.24	0.21	0.46	0.43	0.37
	YLD	P0.05	0.48	0.46	0.29	0.24	0.47	0.46	0.40
	YLD	P0.05-L	0.48	0.52	0.26	0.25	0.47	0.47	0.41
EP	HGT	r	0.84	0.84	0.78	0.78	0.86	0.86	0.83
	HGT	All Data	0.77	0.72	0.67	0.66	0.75	0.80	0.73
	HGT	Ran5	0.68	0.61	0.57	0.56	0.64	0.69	0.62
	HGT	Ran40	0.76	0.71	0.66	0.65	0.74	0.79	0.72
	HGT	Var40	0.75	0.62	0.65	0.63	0.66	0.77	0.68
	HGT	Var10	0.72	0.57	0.62	0.59	0.60	0.73	0.64
	HGT	AMM10	0.72	0.58	0.62	0.59	0.61	0.73	0.64
	HGT	P0.05	0.78	0.74	0.70	0.66	0.79	0.78	0.74
	HGT	PVar10	0.79	0.77	0.71	0.70	0.80	0.82	0.76
	HGT	Var40-L	0.81	0.77	0.73	0.72	0.81	0.85	0.78
	HGT	Var10-L	0.82	0.79	0.74	0.74	0.84	0.87	0.80
	HGT	P0.05-L	0.84	0.81	0.79	0.79	0.84	0.87	0.82
	HGT	PVar10-L	0.89	0.88	0.83	0.85	0.88	0.89	0.87
YP	HGT	r	0.85	0.85					0.85
	HGT	All Data	0.82	0.83					0.83
	HGT	P0.05	0.82	0.83					0.83
	HGT	P0.05-L	0.85	0.84					0.85

EP	HD	r	0.82	0.82					0.82
	HD	All Data	0.75	0.75					0.75
	HD	Ran5	0.60	0.57					0.59
	HD	Ran40	0.73	0.72					0.73
	HD	Var40	0.58	0.64					0.61
	HD	Var10	0.54	0.55					0.55
	HD	AMM10	0.55	0.57					0.56
	HD	P0.05	0.78	0.77					0.78
	HD	PVar10	0.76	0.80					0.78
	HD	Var40-L	0.76	0.78					0.77
	HD	Var10-L	0.78	0.80					0.79
	HD	P0.05-L	0.80	0.81					0.81
	HD	PVar10-L	0.82	0.82					0.82
YP	HD	r	0.70	0.70	0.52	0.52	0.53	0.53	0.58
	HD	All Data	0.71	0.66	0.57	0.51	0.51	0.50	0.58
	HD	P0.05	0.70	0.67	0.51	0.52	0.50	0.54	0.57
	HD	P0.05-L	0.67	0.69	0.49	0.55	0.52	0.59	0.58
EP	TW	r	0.68	0.68	0.33	0.33	0.47	0.47	0.49
	TW	All Data	0.71	0.67	0.31	0.32	0.46	0.52	0.50
	TW	Ran5	0.64	0.61	0.26	0.28	0.41	0.47	0.45
	TW	Ran40	0.70	0.67	0.30	0.32	0.45	0.52	0.49
	TW	Var40	0.69	0.66	0.30	0.33	0.44	0.51	0.49
	TW	Var10	0.66	0.61	0.29	0.32	0.41	0.49	0.46
	TW	AMM10	0.66	0.61	0.29	0.32	0.41	0.49	0.46
	TW	P0.05	0.68	0.66	0.31	0.34	0.45	0.51	0.49
	TW	PVar10	0.68	0.67	0.34	0.36	0.47	0.52	0.51
	TW	Var40-L	0.72	0.65	0.36	0.32	0.49	0.47	0.50
	TW	Var10-L	0.70	0.66	0.37	0.31	0.48	0.48	0.50
	TW	P0.05-L	0.70	0.71	0.29	0.34	0.48	0.47	0.50
	TW	PVar10-L	0.73	0.68	0.37	0.34	0.48	0.47	0.51

† YLD = yield (kg ha^{-1}), TW = test weight (kg m^3), HGT = plant height (cm), HD = heading date (Julian days).

Table 5. Accuracy of genomic selection within the elite population (EP), within the yield population (YP) and between populations where all data from one population was used as the training population (TP) and the other served as the validation population (VP). All accuracies were estimated using data from all TP markers and lines. The last column shows the correlation of marker effects estimated within the EP or the YP for all markers that were significant ($p<0.05$) in at least one of the populations.

	Accuracy via CV‡ within the		TP=	EP	YP	Average Between-population Accuracy	*r* of significant marker effects
Trait	EP	YP	VP=	YP	EP		
YLD†	0.33	0.49		-0.03	0.06	0.02	0.12
TW	0.66	0.41		0.39	0.50	0.44	0.34
HGT	0.54	0.74		0.62	0.48	0.55	0.41
HD	0.56	0.68		0.41	0.32	0.37	0.21
AYLD#	0.44	0.18		0.13	0.12	0.12	-
ATW	0.17	0.22		0.00	-0.03	-0.02	-
AHGT	0.28	0.13		-0.01	-0.02	-0.02	-
AHD	0.35	0.48		0.29	0.15	0.22	-
BYLD	0.36	0.53		0.32	0.13	0.23	-
BTW	0.42	0.34		0.23	0.23	0.23	-
BHGT	0.27	0.43		0.37	0.25	0.31	-
BHD	0.65	0.37		0.23	0.33	0.28	-
FP	0.41	0.46		0.19	0.22	0.20	0.09
SU	0.6	0.58		0.39	0.42	0.41	0.35
LA	0.64	0.67		0.58	0.53	0.56	0.47
WA	0.65	0.60		0.46	0.47	0.47	0.37
SO	0.57	0.63		0.45	0.42	0.43	0.31
SE	0.37	0.59		0.31	0.26	0.28	0.34
FY	0.49	0.61		0.31	0.37	0.34	0.32

† YLD = yield (kg ha^{-1}), TW = test weight (kg m^3), HGT = plant height (cm), HD = heading date (Julian days), Solvent retention capacity (%) for FP= Flour protein, SU= Sucrose content, LA= Lactic acid, WA= Water content, SO= Sodium carbonate, SE= Softness equivalent, and FY= Flour yield.

‡ CV refers to cross validation.

A and B index refers to the AMMI index stability and the ERR index stability, respectively, for each trait.

Table 6. Accuracy of genomic selection between the elite population (EP) and the yield population (YP) using two models g + e and g + e + g×e. For each trait the EP and the YP were each used as either the training population (TP) and as the validation population (VP).

		Model: g + e		Model: g + e + g×e	
	TP=	EP	YP	EP	YP
Trait[†]	VP=	YP	EP	YP	EP
YLD		-0.03	0.10	-0.02	0.14
TW		0.37	0.45	0.38	0.40
HGT		0.59	0.45	0.55	0.42
HD		0.39	0.32	0.40	0.31
FP		0.17	0.19	0.16	0.18
SU		0.29	0.42	0.27	0.34
LA		0.58	0.50	0.56	0.45
WA		0.44	0.45	0.42	0.37
SO		0.37	0.40	0.35	0.35
SE		0.31	0.25	0.32	0.22
FY		0.27	0.37	0.24	0.34

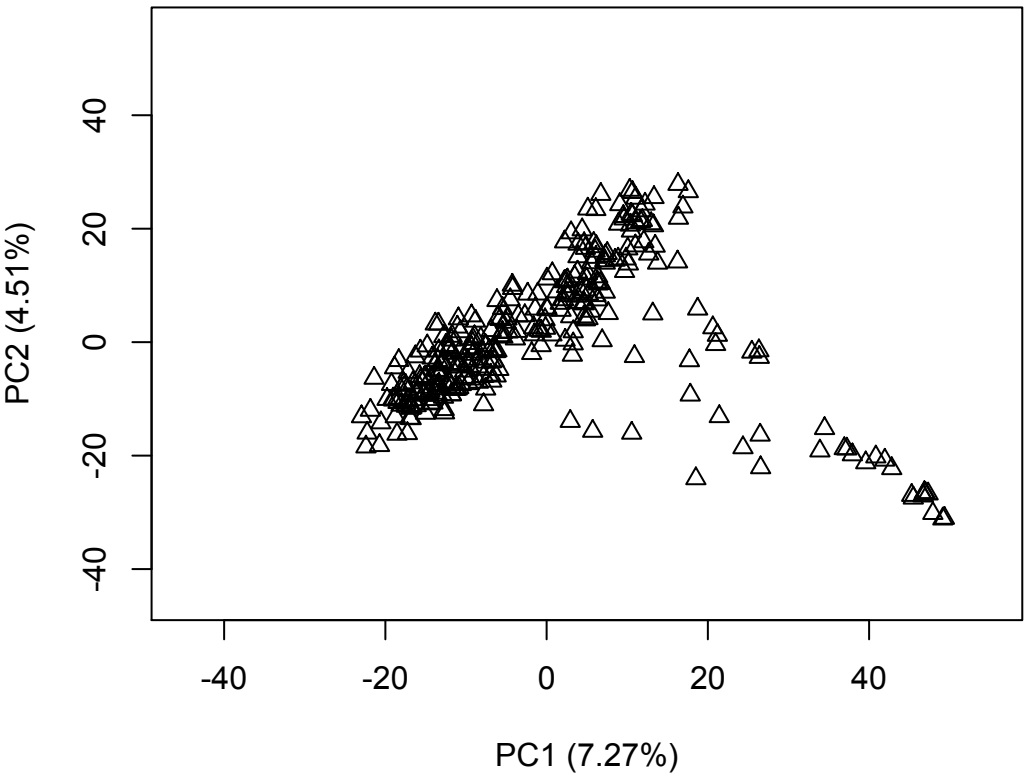
[†] YLD = yield (kg ha^{-1}), TW = test weight (kg m^3), HGT = plant height (cm), HD = heading date (Julian days), Solvent retention capacity (%) for FP= Flour protein, SU= Sucrose content, LA= Lactic acid, WA= Water content, SO= Sodium carbonate, SE= Softness equivalent, and FY= Flour yield

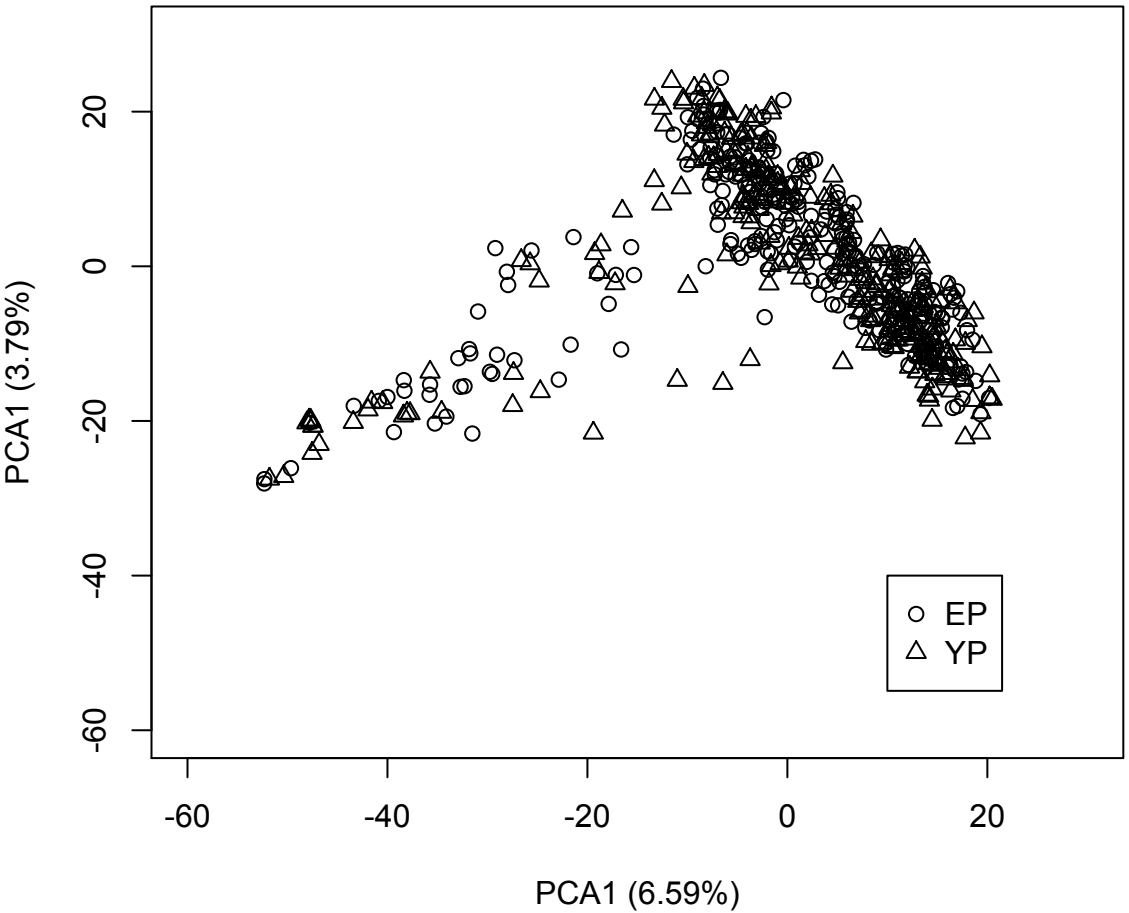
Supplemental Materials

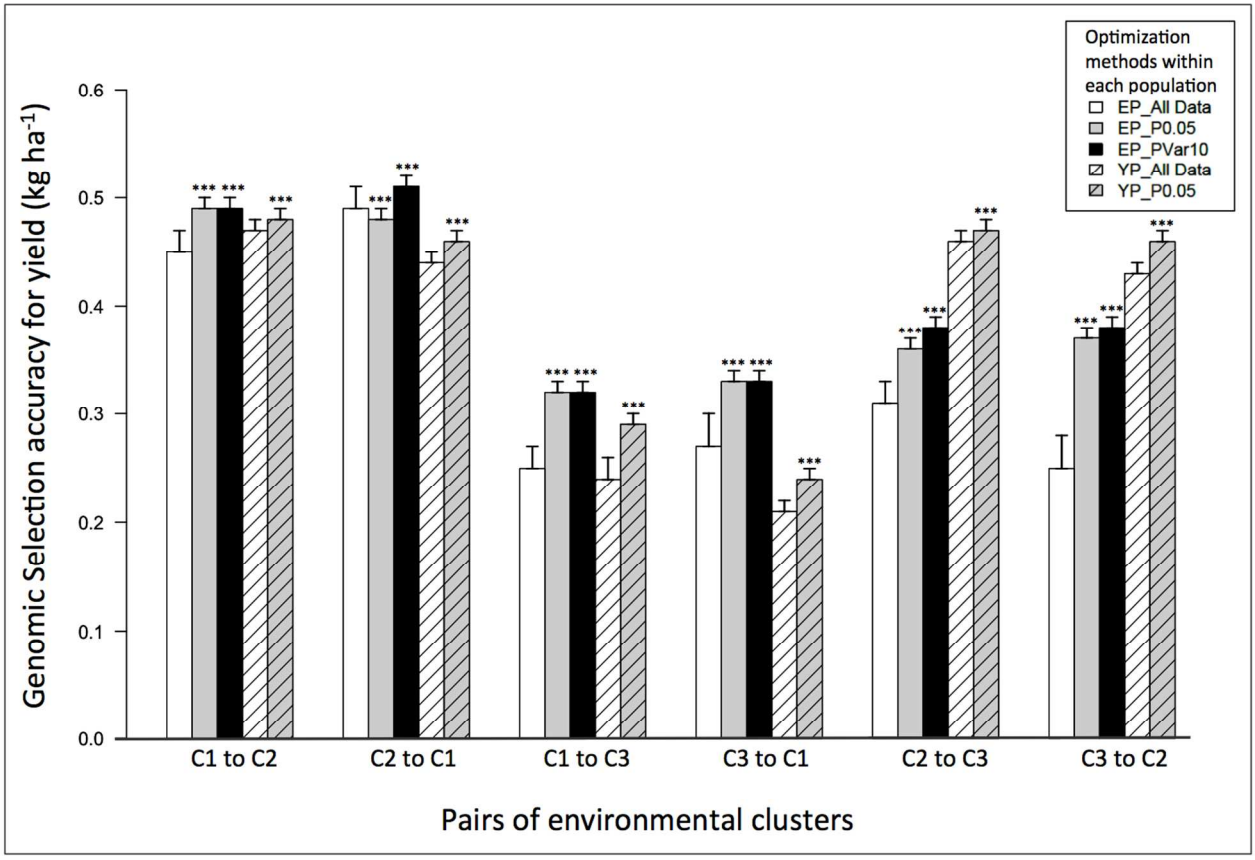
Supplemental Table 1. Entry number, line names and pedigree information for the 294 lines in the yield panel (in excel)

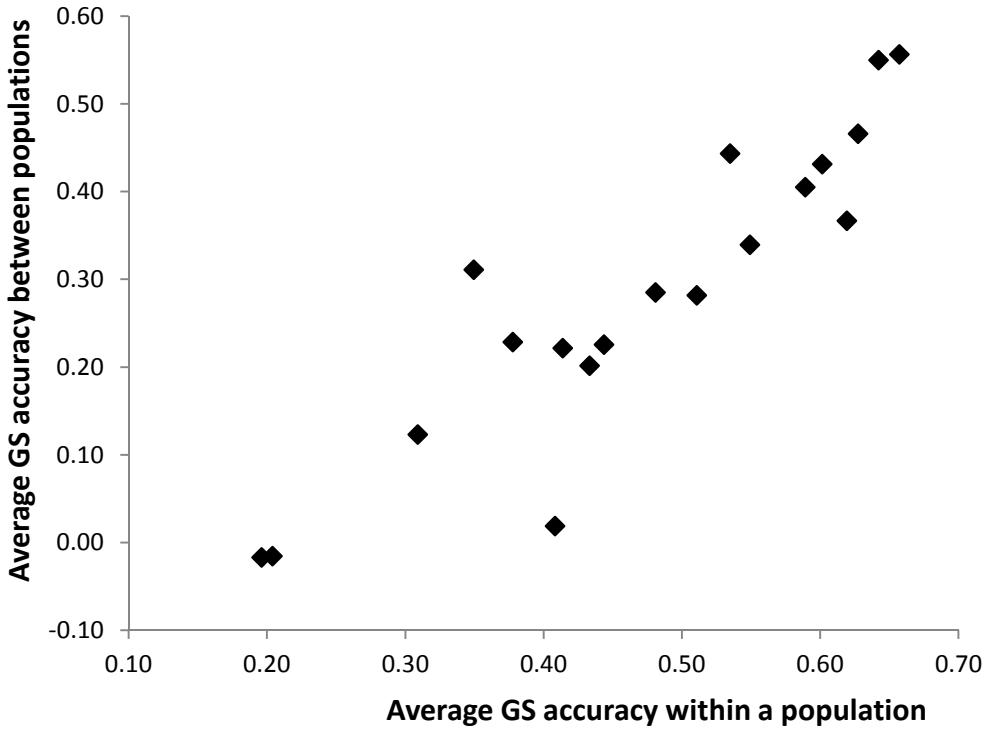
Supplemental Table 2. Accuracy of GS between clusters of environments (data from one cluster was used to predict performance in the other cluster) within the elite population (EP) and yield population (YP). The phenotypic correlation (r) between clusters is shown as is accuracy using all training population data (All Data) and subsets of data as described in Table 2.

Supplemental Figure 1. Plot of the GS accuracy between clusters of environment within a population, for a) test weight, b) plant height and c) heading date, using subset of data (P0.05=set of only significant marker; PVAR10=set of significant and stable markers) compared to using all marker data within the elite panel (EP) and yield panel (YP) population. *** represents an estimate of accuracy that is significantly ($p < 0.001$) different from the accuracy obtained using all marker data. The p-values were adjusted to account for multiple comparisons within each population.


PC2 (4.51%)
40
20
0
-20
-40
-40
-20
0
20
40
PC1 (7.27%)


PCA1 (3.79%)
20
0
-20
-40
-60
-60
-40
-20
0
20
PCA1 (6.59%)
EP
YP


Optimization methods within each population
EP_All Data
EP_P0.05
EP_PVar10
YP_All Data
YP_P0.05
Genomic Selection accuracy for yield (kg ha-1)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
C1 to C2
C2 to C1
C1 to C3
C3 to C1
C2 to C3
C3 to C2
Pairs of environmental clusters


Average GS accuracy between populations
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
0.10
0.20
0.30
0.40
0.50
0.60
0.70
Average GS accuracy within a population