Assessing Genomic Selection Prediction Accuracy in a Dynamic Barley Breeding Population

A. H. Sallam, J. B. Endelman, J.-L. Jannink, and K. P. Smith*

ABSTRACT
Prediction accuracy of genomic selection (GS) has been previously evaluated through simulation and cross-validation; however, validation based on progeny performance in a plant breeding program has not been investigated thoroughly. We evaluated several prediction models in a dynamic barley breeding population comprised of 647 six-row lines using four traits differing in genetic architecture and 1336 single nucleotide polymorphism (SNP) markers. The breeding lines were divided into six sets designated as one parent set and five consecutive progeny sets comprised of representative samples of breeding lines over a 5-yr period. We used these data sets to investigate the effect of model and training population composition on prediction accuracy over time. We found little difference in prediction accuracy among the models confirming prior studies that found the simplest model, random regression best linear unbiased prediction (RR-BLUP), to be accurate across a range of situations. In general, we found that using the parent set was sufficient to predict progeny sets with little to no gain in accuracy from generating larger training populations by combining the parent set with subsequent progeny sets. The prediction accuracy ranged from 0.03 to 0.99 across the four traits and five progeny sets. We explored characteristics of the training and validation populations (marker allele frequency, population structure, and linkage disequilibrium, LD) as well as characteristics of the trait (genetic architecture and heritability, \(H^2\)). Fixation of markers associated with a trait over time was most clearly associated with reduced prediction accuracy for the mycotoxin trait DON. Higher trait \(H^2\) in the training population and simpler trait architecture were associated with greater prediction accuracy.

GENOMIC SELECTION is touted as a marker-based breeding approach that complements traditional marker-assisted selection (MAS) and phenotypic selection. In traditional MAS, favorable alleles or genes for relatively simply inherited traits are mapped and then molecular markers linked to those alleles are used to select individuals to use as parents or to advance from segregating breeding populations (Bernardo, 2008). Marker-assisted selection is more effective than phenotypic selection if the tagged loci account for a large portion of the total genetic variation within the population of selection candidates (Collins et al., 2003; Castro et al., 2003; Xu and Crouch, 2008). The limitation of traditional MAS for highly complex traits is that it captures only a small portion of the total genetic variation because it uses a limited number of selected markers (Lande and Thompson, 1990; Bernardo, 2010). Phenotypic selection is effective on quantitative traits, but is limited to stages in breeding cycles and environments where such traits can be measured effectively, such as for advanced lines in multiple location field trials. Therefore, GS can be strategically implemented in

A.H. Sallam, Dep. of Agronomy and Plant Genetics, Univ. of Minnesota, St. Paul, MN 55108; J.B. Endelman, Dep. of Horticulture, Univ. of Wisconsin-Madison, 1575 Linden Dr., Madison, WI 53706; J.-L. Jannink, USDA-ARS, R.W. Holley Center for Agriculture and Health, Cornell Univ., Ithaca, NY 14853; K.P. Smith, Dep. of Agronomy and Plant Genetics, Univ. of Minnesota, St. Paul, MN 55108. Received 6 Jan. 2014. *Corresponding author (smith376@umn.edu).

Abbreviations: BLUEs, best linear unbiased estimations; DON, deoxynivalenol; EMMA, efficient mixed-model association; FHB, Fusarium head blight; Fst, Wright’s fixation index; GS, genomic selection; \(H^2\), heritability; GEBV, genomic estimated breeding value; LD, linkage disequilibrium; MAF, minor allele frequency; MAS, marker-assisted selection; \(r^2\), predictive ability; QTL, quantitative trait loci; REMI, restricted maximum likelihood; RKHS, Reproducing Kernel Hilbert Space; RR-BLUP, random regression best linear unbiased prediction; SNP, single nucleotide polymorphism.
breeding for quantitative traits at points in the breeding process where phenotypic selection is not feasible.

Genomic selection uses trait predictions based on estimates of all marker effects distributed across the genome (Meuwissen et al., 2001). Based on simulation studies, GS should improve gain from selection, reduce costs associated with phenotyping, and accelerate development of new cultivars by reducing the length of the breeding cycle (Heffner et al., 2009, 2010). Implementing GS is accomplished by first estimating marker effects in a training population and then using those estimates to predict the performance of selection candidates. The predicted value of a selection candidate based on marker effects is referred to as the genomic estimated breeding value (GEBV; Meuwissen et al., 2001).

A key component to the effectiveness of GS is prediction accuracy. Prediction accuracy is defined as the correlation between the GEBV and the true breeding value divided by the square root of \( H^2 \), which is estimated by measuring phenotypic performance (Goddard and Hayes, 2007; Zhong et al., 2009). There are three general methods to assess prediction accuracy using real data: (i) subset validation, (ii) interset validation, and (iii) progeny validation (Figure 1). Subset validation is implemented by randomly dividing a single population of individuals into equal subsamples; one subsample is used as a validation set to be predicted using the remaining subsamples as the training set. Subset validation has been used to assess prediction accuracy in cattle, wheat \((Triticum aestivum L.)\), and barley \((Hordeum vulgare L.)\), among many other livestock and crop species (Luan et al., 2009; Heffner et al., 2010; Lorenz et al., 2012; Poland et al., 2012). In interset validation, predefined sets of genotypes are designated as training and validation populations. These sets could be the same genotypes from independent environments as training and validation data sets or sets of breeding lines chronologically defined where older lines are used to predict newer lines from either the same or independent environments (Asoro et al., 2011; Lorenz et al., 2012). Progeny validation implies that the training population includes parents (or grandparents, and so forth) of progeny lines that comprise the validation population. A simulation study in animals has shown that decreases in prediction accuracy are associated with decay of LD between markers and quantitative trait loci (QTL) resulting from recombination in progeny generations (Habier et al., 2007). Therefore, meaningful assessment of prediction accuracy should include progeny validation. In plants, we are aware of only a single study that assesses accuracy by progeny validation using empirical phenotypic and genotypic information (Hoheinz et al., 2012).

To assess the potential of GS, researchers have explored various factors that affect prediction accuracy, including prediction models. These models include RR-BLUP, Bayes A, Bayes B, Bayes C\(\pi\), Bayes LASSO, and Reproducing Kernel Hilbert Space (RKHS; Meuwissen et al., 2001; Kizilkaya et al., 2010; de los Campos et al., 2009; Gianola and van Kaam, 2008). These models differ in the assumptions made for marker variances associated with markers and/or types of gene action (reviewed by Lorenz et al., 2011). RR-BLUP assumes that all markers have equal variance, whereas Bayes A, Bayes B, Bayes C\(\pi\), and Bayes Lasso models do not impose this constraint (Meuwissen et al., 2001; de los Campos et al., 2009; Kizilkaya et al., 2010). The RKHS regression model can capture both the additive and nonadditive interactions among loci by creating a kernel matrix that includes interactions among marker covariates (Gianola and van Kaam, 2008). Results of empirical studies have shown variable performance of prediction models on different traits (Crossa et al., 2010; Lorenz et al., 2012; Rutkoski et al., 2012).

Other factors shown to affect prediction accuracy include: (i) the LD between markers and QTL in the training and the validation populations, (ii) the size of the training population \((N)\), (iii) the \(H^2\) of the trait under investigation, and (iv) the genetic architecture of the trait. Increasing marker density will improve prediction accuracy by increasing the number of QTL that are in LD with markers and capturing more of the genetic variation (de Roos et al., 2009; Asoro et al., 2011; Heffner et al., 2011; Zhao et al., 2012). The successful application of GS across generations relies on the persistence of LD phase between markers and QTL (de Roos et al., 2008). The persistence of LD phase measured by the correlation of \(r\) among populations is likely to be a function of the genetic relationship between populations (de Roos et al., 2008; Toosi et al., 2010). Increasing \(N\) will lead to better estimation of SNP effects (Hayes et al., 2009) and therefore, increases prediction accuracy (Lorenzana and Bernardo, 2009; Asoro et al., 2011; Lorenz et al., 2012). In a simulation study, Daetwyler et al. (2010) found that prediction accuracies increased with increase in \(H^2\) of the trait regardless of the number of QTL controlling the trait or the prediction model used. In a study that manipulated \(H^2\) by introducing random error into empirical data sets, Combs and Bernardo (2013) showed that accuracy increased with increasing \(H^2\) and \(N\), and that prediction accuracies were similar for different combinations when \(H^2 \times N\) were held constant. Generally, prediction accuracy decreases with the increase of trait complexity (Hayes et al., 2010). Prediction models can vary in performance among traits with different genetic architecture. Bayes B was more accurate when a smaller number of loci control the trait whereas RR-BLUP was insensitive to genetic architecture (Daetwyler et al., 2010).

Previous studies have demonstrated the potential of GS on the basis of subset validation and interset validation. While these results are promising, additional research is needed to assess accuracy in the context of applied breeding. Specifically, validation experiments are needed to assess the accuracy of prediction on progenies (progeny validation) over time, as would occur in breeding populations. This would take into account changes in allele frequency and LD that would be expected to occur as a result of recombination and selection within a dynamic breeding program. Lorenz et al. (2012) investigated prediction accuracy for the disease Fusarium
head blight (FHB) and its associated mycotoxin deoxynivalenol (DON) using interset validation. In this study, we advance this work by using progeny validation and include additional agronomic traits. We use a set of breeding lines as a training population that include parents that were used to predict five chronological sets of progenies (2006–2010) from a breeding program. Our specific objectives were to (i) compare the accuracy of different GS prediction models on DON concentration, FHB resistance, yield, and plant height, (ii) study the effect of trait architecture on prediction accuracy, (iii) characterize changes in prediction accuracy over time, (iv) examine the relationship between prediction accuracy and training population size and composition, allele frequency, LD, and genetic distance between the training and validation populations.

**MATERIALS AND METHODS**

**Germplasm**

To explore the accuracy of genomic predictions, we utilized historical sets of breeding lines that we define as parent or progeny sets from the University of Minnesota barley breeding program. The parent set is comprised of 168 breeding lines that were developed between 1999 and 2004 and were either used as parents to develop lines in the progeny sets or were cohorts of breeding lines that were used as parents. The five progeny sets consist of five chronological sets of breeding lines evaluated between 2006 and 2010. Each progeny set consists of approximately 96 lines that were representative of the breeding lines developed that year in the breeding program. The progeny sets 2006 and 2007 are the breeding lines from the University of Minnesota barley breeding program that were included in the association mapping study conducted by Massman et al. (2011) and were referred to as CAP I and CAP II in that study. All the breeding lines in the parental and progeny sets were developed by single seed decent to at least the F_4. At that point, F_45 lines were evaluated for resistance to FHB resistance, heading date, plant height, maturity, and lodging. Lines selected as favorable for these traits are then advanced to preliminary yield trials the following year (Smith et al., 2013). The preliminary yield trial data were used to characterize progeny set lines and the year designation for the progeny set refers to the year that the breeding line entered preliminary yield trials. All pedigree, SNP markers, and phenotypic data related to these sets of breeding lines are available from the public database The Hordeum Toolbox (http://thehordeumtoolbox.org, verified 3 Oct. 2014; Blake et al., 2012).

**Phenotypic Evaluation**

The parental lines were evaluated together for agronomic traits in five experiments conducted between 2009 and 2011 at Crookston and St. Paul, MN, in an augmented block design with two replications and four incomplete blocks per replication (Supplemental Table S1). Planting density for all traits in all experiments was 300 plants m⁻². Each line was represented once per block in two-row plots 3 m in length. Six check cultivars (Drummond, Lacey, Quest, Rasmusson, Stellar-ND, and Tradition) were randomly assigned to each block (Horsley et al., 2002, 2006a; Rasmusson et al., 2001; Smith et al., 2010, 2013). We also characterized the parental lines using the historical data that was collected as part of the breeding program as these lines were entered into preliminary yield trials. Experiments for this unbalanced data set were arranged as a randomized complete block design with two replications in two-row plots 3 m in length and were conducted between 1999 and 2004. Three checks (‘Robust’, ‘Stander’, and Lacey) were common to all the experiments (Rasmusson and Wilcoxson, 1983; Rasmusson et al., 1993). For both the historic and contemporary data sets, each line was evaluated at least two times in yield trials conducted in St. Paul, Morris, and Crookston, MN. Yield was determined by harvesting each plot with a Wintersteiger small plot combine, weighing the grain, and expressing it as kg/ha. Plant height was assessed as the height in centimeters of two randomly selected samples of plants from the middle of the plot from the soil surface to the tip of the spike, excluding awns. The parental lines were evaluated for FHB resistance and DON concentration in 2009 at St. Paul and in 2010 at St. Paul and Crookston, MN, in an augmented block design with two replications in four incomplete blocks. Each line was represented one time per block in single-row plots 1.8 m in length, with 30 cm between rows. Six check cultivars (Drummond, Lacey, Quest, Rasmusson, Stellar-ND, and Tradition) were randomly assigned to each block. The parental lines were evaluated for FHB resistance and DON concentration using a previously described method (Steffenson, 2003). Briefly, in St. Paul, plants were spray inoculated with a *F. graminearum* macroconidia suspension using CO₂-pressure backpack sprayers. Plots were inoculated when at least 90% of the heads had emerged from the boot and sprayed again 3 d later (Mesfin et al., 2003). Mist irrigation was applied immediately after inoculation to promote disease infection. In Crookston, MN, plants were inoculated by grain spawn using autoclaved corn (*Zea mays* L.) colonized by five local isolates of *F. graminearum* (Horsley et al., 2006b). The colonized grain was spread on the ground 2 wk before flowering and again 1 wk later. Overhead mist irrigation started 2 wk before anthesis and continued until the hard dough stage of maturity. Fusarium head blight severity was assessed about 14 d after inoculation by estimating the percentage of infected kernels on a random sample of 10 spikes per plot using the following assessment scale 0, 1, 3, 5, 10, 15, 25, 35, 50, 75, and 100%. DON concentration was determined on a 25-g sample from the harvested grain by gas chromatography and mass spectrometry and expressed in ppm according to the procedures of Mirocha et al. (1998).

Lines included in the progeny sets were derived from crosses made between 2003 and 2007 and were evaluated in preliminary yield trials conducted from 2006 to 2010 (Supplemental Table S1). Plots were arranged in a
randomized complete block design with two replications and four check varieties (Robust, Stander, MNBrite, and Lacey). Each progeny set was evaluated for yield and plant height in Crookston, St. Paul, and Morris, MN, as described above. The progeny sets were also evaluated for FHB resistance and DON accumulation in disease nurseries as described above. Each progeny set was evaluated in three to four FHB experiments located in St. Paul and Crookston, MN, and Osnabrock and Fargo, ND. Disease inoculation, disease assessment, and DON measurements were done as previously described.

Genotypic Evaluation

DNA for genotyping was extracted from a single plant from the F_{15} bulk seed used in the phenotypic evaluation. Approximately 3-wk-old leaf tissue was harvested and freeze-dried. DNA was extracted at the USDA genotyping center in Fargo, ND, using the protocol of Slotta et al. (2008). Each DNA sample was genotyped with the 1536 SNPs referred to as BOPA1 using the Illumina GoldenGate oligonucleotide assay (Close et al., 2009). Markers were filtered in parents set based on minor allele frequency (MAF) < 0.01 and missing data frequency > 10%. Missing marker values were imputed using naïve imputation so that analytical operations could be performed.

Data Analysis

Analysis of variance was performed for DON concentration, FHB resistance, yield, and plant height using the PROC GLM procedure in SAS (v.9.3, SAS Institute, 2011). For each experiment, outlier observations with standardized residual absolute values of three or more were removed from the data set and scored as missing values. One experiment (yield in St. Paul, 2010) was removed because no significant differences were found among lines.

To avoid including common checks across experiments in variance component estimates, two-step procedures were used. For the contemporary data from the parental set, we first adjusted phenotypes for block effects by using the common checks among blocks using the PROC MIXED procedure in SAS (v.9.3, SAS Institute, 2011). The model was \( y = X\beta + Zu + e \) where \( y \) is the vector of unadjusted phenotypes, \( \beta \) is the vector of fixed block effects, and \( u \) is the vector of random check effects. \( X \) and \( Z \) are incidence matrices to relate the vector of unadjusted phenotypes to \( \beta \) and \( u \). We then adjusted phenotypes for trial effects by estimating these effects as fixed in an analysis with lines as random effects. The model was \( y^* = X\beta + Zu + e \), where \( y^* \) represents the phenotypes adjusted for block effects calculated in the first step, \( \beta \) is the vector of fixed trial effects, and \( u \) is the vector of random line effects. In the historic data for the parent set, subsets of lines were evaluated in different years, but a common set of checks was included in each trial. Similarly to the contemporary data, phenotypes were adjusted for trial effects by computing these effects in a mixed model with checks as random and trials as fixed effects. In the progeny data sets, phenotypes were adjusted for trial effects by computing these effects in a mixed model, with lines as random and trials as fixed effects. Finally, best linear unbiased estimations (BLUEs) for lines in each experiment were estimated in models with adjusted phenotypes as the response variable and lines as fixed effects. Variance components were estimated using restricted maximum likelihood (REML) in the PROC MIXED procedure in SAS by using the line BLUEs as the response variable, lines as random effects and experiments as fixed effects. Broad-sense \( H^2 \) on an entry mean was estimated for all traits using the equation \( H^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2/n) \), where \( \sigma_g^2 \) is genetic variance, \( \sigma_e^2 \) is the pooled error variance that includes G x E and residuals, and \( n \) is the number of trials.

Characterizing LD, Genetic Distance, and Parental Contribution

To assess the extent of the LD within the parental and progeny sets, the adjacent marker LD was characterized as \( r^2 \) using Haploview v.4.0 (Barrett et al., 2005). To assess the persistence of LD phase between the parental and progeny sets, the correlations of \( r \) were calculated between parental and each progeny set (de Roos et al., 2008; Toosi et al., 2010). We measured genetic distance between the parent set and each progeny set by the fixation index (Fst, Weir and Cockerham, 1984) and Nei’s genetic distance (Nei, 1987). Fst and Nei’s genetic distances measure the differentiation between two populations due to changes in allele frequencies among populations. The contribution of the parental lines to a progeny set was assessed by summing the number of parents for a progeny line that were included in the parent set over the progeny set and dividing that by twice the number of lines in that progeny set.

Association Analysis

To identify sets of markers associated with traits, association analysis was implemented using the efficient mixed-model association (EMMA) approach, which corrects for population structure using genetic relatedness (Kang et al., 2008). Association analyses were done on the parent set for DON concentration, FHB resistance, yield, and plant height using EMMA package implemented in R (Kang et al., 2008). The analysis was based on the mixed model:

\[ y = X\beta + Zu + e \]  

where \( y \) is the vector of individual phenotypes, \( X \) is an incidence matrix that relates \( \beta \) to \( y \), \( \beta \) is the vector of fixed effects that includes the overall mean and SNPs, \( Z \) is the matrix of random effects that relates \( u \) to phenotypes, \( u \) is the random effect of the genetic background of each line and is distributed as \( u \sim N(0, K\sigma^2_g) \). \( K \) is the kinship matrix derived from marker genotypes and \( \sigma^2_g \) is the genetic variance, and \( e \) is the residual where \( e \sim N(0, \sigma^2_e) \). I is the identity matrix and \( \sigma^2_e \) is the error variance (Kang et al., 2008). We used a relaxed threshold of –log \( p \)-value of 1.3 (\( p \)-value of 0.05) to identify markers.
Genomic predictions were estimated using four methods: ridge regression best linear unbiased prediction (RR-BLUP; Meuwissen et al., 2001), Gaussian kernel model (Gianola and van Kaam, 2008; Endelman, 2011), Exponential kernel model (Piepho, 2009; Endelman, 2011), and Bayes Cπ (Kizilkaya et al., 2010). RR-BLUP and Bayes Cπ can be modeled as

\[ y = 1u + \sum_{j=1}^{K} Z_j a_j \delta_j + e \]  

where \( y \) is the vector of individual phenotypes, \( u \) is the population mean, \( K \) is the number of markers, \( Z \) is the incidence matrix that links marker \( j \) genotypes to individuals, \( a \) is the effect of marker \( j \), \( \delta \) is an indicator variable that indicates the absence or the presence of marker \( j \) with probability of \( \pi \) and \( 1 - \pi \), respectively, and \( e \) is the random residual. In RR-BLUP, all markers are included (\( \delta = 1 \)) and their effects are distributed with the same variance \( \mathcal{N}(0, \sigma_z^2) \). The variance of this distribution was estimated on the basis of marker and phenotypic data using REML. A Bayesian model was used to relax the assumption of RR-BLUP to allow some marker variances to be zero. Bayes Cπ assumes common marker variance across all markers included in the model; however, it allows some markers to have no effect on the trait (Kizilkaya et al., 2010). In Bayes Cπ, it is assumed that each marker \( j \) has a zero effect with probability \( \pi \) when \( \delta_j = 0 \) and an effect \( a_j \sim \mathcal{N}(0, \sigma_z^2) \) with probability \( 1 - \pi \) when \( \delta_j = 1 \). The parameter \( \pi \) is treated as unknown and is estimated from the training data. In the Markov Chain Monte Carlo (MCMC) algorithm for Bayes Cπ, 10,000 iterations of Gibbs sampling were used and the first 2500 iterations were discarded as burn-in. We implemented Bayes Cπ analysis in R (R Development Core Team, 2012). Gaussian and Exponential kernel models were implemented to capture both the additive and nonadditive interactions between marker genotypes using the R package rrBLUP (Endelman, 2011; R Development Core Team, 2012). These models can be presented as

\[ y = 1u + Zg + e \]  

where \( y \) is the vector of individual phenotypes, \( u \) is the population mean, \( Z \) is the matrix that relates \( g \) to the adjusted phenotypes, \( g \) is the vector of genotypic values that is distributed as \( g \sim \mathcal{N}(0, \mathbf{K}\sigma_g^2) \), where \( \mathbf{K} \) is the kernel similarity matrix, and \( e \) is the residual (Endelman, 2011). The Gaussian and Exponential models do not partition the total genetic variance into additive and nonadditive variances; rather, kernel functions are used to capture these effects. Genomic predictions were calculated for all the lines in the validation population using the four prediction models. The correlation coefficient between the genomic predictions and line BLUEs was used to calculate the predictive ability (\( r_a \)). Prediction accuracy \( (r_a/H) \) of GS (Legarra et al., 2008; Chen et al., 2011) was calculated by dividing the \( r_a \) by the square root of the broad-sense \( H^2 \) derived from the validation population data.

**Training Populations**

To test the effect of training population composition on prediction accuracy, three different scenarios were implemented by varying the training data set. In the first scenario, the 168 parental lines, using either the contemporary or historic data, were used as the training set to predict the performance of lines in each of the five progeny sets. In the second scenario, we varied the training population composition by adding one or more of the progeny sets to the contemporary parent set to predict the performance of a later progeny set. In the third scenario, we used two earlier progeny sets to predict a later progeny set. For each scenario, we implemented the four prediction models described previously.

Because the experiments described above were used to assess different types of training populations that varied in population size, we also tested the effect of training population size on prediction accuracy for two out of the four traits in 2008 and 2010 progeny sets as validation populations. For DON concentration and yield, we used three scenarios. In the first scenario, we randomly sampled 25, 50, 75, 100, and 150 lines from the parent set \( (n = 168) \). For each population size, samples were drawn without replacement 500 times. In the second scenario, we combined progeny sets before the validation set (combined 2006 to 2007 when predicting 2008 and 2006 to 2009 when predicting 2010) with the parent set into a single panel from which samples were drawn to generate various training sets. We generated training sets from the larger training panels by randomly sampling 25, 50, 75, 100, 150, 168, 264, and 360 when predicting 2008 and
sampling 25, 50, 75, 100, 150, 168, 264, 360, and 456 lines when predicting 2010. For each population size, samples were drawn without replacement 500 times. In the third scenario, we combined the parental and progeny sets by sequentially adding each progeny set in chronological order to the parent set such that in each round of prediction the size of the training population was increased by 96. This represents the single case that would occur if a breeder accumulated data over time to increase the size of the training population, and thus there exists just one instance for this scenario. For each of the scenarios, we used the training populations to generate predictions of the 2008 and 2010 progeny sets for DON concentration and yield using RR-BLUP.

**RESULTS**

**Phenotypic Traits and Marker Density**

The parent set and each of the progeny sets were evaluated in multiple yield and disease experiments between 2006 and 2011. The yield experiment for the parent set in St. Paul 2010 was removed due to very severe lodging that resulted in large error variance and no significant differences among lines. For all traits and experiments, we observed significant differences among lines (p-value < 0.01) in the parent and progeny sets. Genetic variances (Table 1) decreased for DON concentration and plant height as a function of progeny set year, whereas the genetic variances for yield and FHB resistance fluctuated. The estimates of the $H^2$ were moderate for DON concentration and FHB resistance; low to moderate for yield; and high for plant height as expected based on previous studies (Boukerrou and Rasmusson, 1990; Ma et al., 2000; Mesfin et al., 2003). After filtering the 1536 BOPA1 SNPs for MAF and missing data, 984 markers remained that spanned 1085 cM of the barley genome with an average distance between adjacent markers of 1.1 cM.

**Relationship between Parent and Progeny Sets**

The average adjacent marker LD in the progeny sets were greater than the parent set and showed a slight increase over time (Fig. 2). The correlation of $r$ between parental and progeny sets ranged from 0.44 to 0.61 (Fig. 2). The parental contribution of the parent set to the progeny sets decreased continuously over time with about a 75% reduction from 2006 to 2010 (Fig. 3). Concurrent with this decrease in parental contribution was an increase in genetic distance between parent and the consecutive progeny sets over time (Fig. 3). The genetic relationship between the parents and the progeny sets can be visualized in the heatmap of the kinship matrix (Fig. 4). As lines were developed in the breeding program, their similarity to the parent set diminished over time.
Complex trait with only a few markers with $R^2$ values exceeding 0.30. Yield, on the other hand, was the most complex trait, followed by plant height, DON concentration, and FHB resistance. Number of markers ranged from 59, 24, and 17 markers had $R^2$ values greater than 0.10. Based on Fig. 5, it is clear that multiple markers are likely associated with the same QTL. Nevertheless, 83, 58, 62, and 59 markers associated with DON concentration, FHB resistance, yield, and plant height, respectively.

To characterize the possible role of selection, we examined allele frequencies of the SNP markers associated with the four traits (Fig. 6). In general, there was an increase over time with the complete set of genome-wide markers. For markers associated with individual traits this trend was most apparent for DON concentration followed by plant height and yield. No relationship was observed for FHB resistance.

To investigate the effect of trait architecture and the distribution of marker effects on prediction accuracy, we estimated the proportion of variance explained by each SNP marker in the parent set for all traits (Fig. 7). Based on association analysis using the parent set, all of the traits displayed quantitative inheritance with multiple loci distributed across the genome contributing to the traits (Fig. 5). Coincident QTLs for plant height, DON concentration, and FHB resistance on the short arm of chromosome 4H were detected in a region previously identified in a study using a similar germplasm (Massman et al., 2011). Using a relaxed $p$-value of 0.05, we identified 62, 58, 62, and 59 markers associated with DON concentration, FHB resistance, yield, and plant height, respectively.

To characterize the genetic architecture of the traits is to use the $\pi$ parameter, which is the proportion of markers with no effect, estimated from the Bayes C model. When using the parent set as a training population, the $\pi$ parameter estimates for yield over four runs ranged between 0.28 and 0.43, with a mean $\pi$ of 0.35. For DON concentration the $\pi$ parameter estimates ranged between 0.37 and 0.54, with a mean $\pi$ of 0.45. For FHB the $\pi$ parameter estimates ranged between 0.49 and 0.58, with a mean $\pi$ of 0.53. For plant height, the $\pi$ parameter estimates ranged between 0.45 and 0.80, with a mean $\pi$ of 0.63. Thus, based on $\pi$ estimates, yield was the most complex trait, followed by DON concentration, FHB resistance, and then plant height. This suggestive trend from higher to lower complexity agrees with the distribution of $R^2$ values for markers displayed in Fig. 7. Assessment of genetic architecture based on the $\pi$ parameter estimates also agrees with the results of (Lorenz et al., 2012) for DON concentration and FHB resistance.

### Marker-Trait Associations

Based on association analysis using the parent set, all of the traits displayed quantitative inheritance with multiple loci distributed across the genome contributing to the traits (Fig. 5). Coincident QTLs for plant height, DON concentration, and FHB resistance on the short arm of chromosome 4H were detected in a region previously identified in a study using a similar germplasm (Massman et al., 2011). Using a relaxed $p$-value of 0.05, we identified 62, 58, 62, and 59 markers associated with DON concentration, FHB resistance, yield, and plant height, respectively.

To characterize the genetic architecture of the traits is to use the $\pi$ parameter, which is the proportion of markers with no effect, estimated from the Bayes C model. When using the parent set as a training population, the $\pi$ parameter estimates for yield over four runs ranged between 0.28 and 0.43, with a mean $\pi$ of 0.35. For DON concentration the $\pi$ parameter estimates ranged between 0.37 and 0.54, with a mean $\pi$ of 0.45. For FHB the $\pi$ parameter estimates ranged between 0.49 and 0.58, with a mean $\pi$ of 0.53. For plant height, the $\pi$ parameter estimates ranged between 0.45 and 0.80, with a mean $\pi$ of 0.63. Thus, based on $\pi$ estimates, yield was the most complex trait, followed by DON concentration, FHB resistance, and then plant height. This suggestive trend from higher to lower complexity agrees with the distribution of $R^2$ values for markers displayed in Fig. 7. Assessment of genetic architecture based on the $\pi$ parameter estimates also agrees with the results of (Lorenz et al., 2012) for DON concentration and FHB resistance.

#### Prediction Accuracy

For the four traits investigated, all prediction models performed similarly to each other with respect to prediction accuracy (Supplemental Fig. S1). When we averaged the prediction accuracy across the five progeny set years, we found no significant differences among the models (Supplemental Table S2). There was a strong correlation among the four models for the predictions of yield for the combined set of progeny when using the parent set as a training population (Fig. 8). Consistent with other GS studies RR-BLUP, in which all marker effects are sampled from the same distribution and similarly shrunk toward zero equally, performed similarly to models that do not impose that restriction. Further comparisons of prediction accuracy are based on RR-BLUP.

Another important consideration for prediction accuracy is the need to generate new phenotypic data for model training or to use existing data sets. We estimated marker effects using available historical data for yield from the breeding program in the parent set and compared that to estimates obtained using the contemporary data set (Fig. 9). The average prediction accuracy over the 5 yr based on contemporary (0.57) and historic data (0.42) were not significantly different ($p$-value = 0.38). Combining historic and contemporary data was equal to using the contemporary data alone.

In general, when using the parent set to predict progeny sets, accuracy was highest for plant height and lowest for yield (Fig. 10). Fusarium head blight resistance and DON concentration had similar prediction accuracy. The relationship between accuracy and year of the progeny set also differed between traits (Fig. 10). For yield and plant height the prediction accuracy fluctuated over time, while for DON concentration there was an overall decrease. Accuracy for FHB resistance remained relatively constant across years.

Varying the size of the training population by adding one or more progeny sets to the parent set to predict a later progeny set generally showed the same trend observed for the parent set alone, and in several instances resulted in reduced accuracy. Using the most
recent breeding lines and environments to the test population by training a prediction model from the two progeny sets before the validation year was generally less accurate than using the parent set. The general trend was that higher trait $H^2$ in the parent training population corresponded to higher $r$, when parents predicted all the progenies using RR-BLUP (Fig. 11).

Since adding consecutive sets of lines to the parent set changed both the composition and the size of the training set, we looked at the effect of population size on prediction accuracy with the parent set only, and with the parent set combined with the progeny sets with different population sizes drawn at random. In both cases, we identified an increase in accuracy with increasing in population size for DON concentration and yield (Fig. 12). However, prediction accuracy for DON concentration seemed to plateau at a population size of 75, while yield did not appear to plateau. It was also interesting that random sampling from just the parent set often produced higher prediction accuracies than random sampling from the combined parent and progeny sets when compared at the same training population size.

**DISCUSSION**

Successful implementation of GS will involve the use of improved genotyping technology to shorten the breeding cycle and increased selection intensity by effective modeling to accurately predict breeding values. Prior studies have examined factors that affect prediction accuracy through simulation and empirically through cross-validation (e.g., Daetwyler et al., 2010; Lorenz et al., 2012). To assess prediction accuracy in a more realistic context, we used sets of parents and progenies from an active breeding program as training and validation sets. Because breeding populations are dynamic, we tested progeny sets defined chronologically over a 5-yr period. We found that prediction accuracy varied over time, and that simply adding data from breeding progenies to the training population did not improve and often reduced prediction accuracy. This suggests that careful construction of...
Changes in Breeding Populations Over Time

Breeding populations are dynamic and as such approaches to using prediction methods should be informed by changes in prediction accuracy that may occur over time. Breeding value predictions are influenced by allele frequencies, LD level, and the introgression of new alleles. These factors will change over breeding cycles due to selection, genetic drift, and unequal parental contribution to progenies. We investigated prediction accuracy in validation sets of breeding lines over a 5-yr period and observed both little to no change, as well as substantial decrease in prediction accuracy over time, depending on the trait. To better understand the underlying population parameters that could be affecting prediction accuracy, we compared the parental training population to the progeny validation sets with respect to allele frequencies, parental contribution, genetic distance, and LD.

More than 35% of marker alleles that were segregating in the parent set became fixed in the 2010 progeny set. Gradual increases in allele fixation for trait specific markers are an indication of the effect of selection and/or genetic drift. A previous study of the University of Minnesota barley breeding program showed that a reduction in allelic diversity for specific simple sequence repeat markers was in some cases associated with QTL regions for traits that were under selection (Condón et al., 2008). Once a marker associated with a trait that is segregating in training population becomes fixed in subsequent progeny generations, it loses its predictive value for the purpose of selection. We observed a substantial increase in the number of fixed SNPs associated with DON that corresponded to a reduction in prediction accuracy. However, we saw a similar increase in fixed SNPs for yield and no corresponding reduction in accuracy. One possible explanation is that yield is likely conditioned by a greater number of QTL with smaller effects and therefore increases in the number SNPs that become fixed over time would have less of an effect on prediction accuracy.

Prediction accuracy should be greatest when the training population is more closely related to the validation population (Habier et al., 2007; Hayes et al., 2009; Lorenz et al., 2012). We observed in most cases that the 2006 progeny set, which was genetically more similar to and had the largest number of direct parents from the parent set, was predicted with the greatest accuracy. The increase in the genetic distances between the parental and progeny sets was most closely associated with a decline in prediction accuracy for DON but not for FHB resistance, yield, and plant height. This indicates that other factors may also contribute to changes in prediction accuracy over time.

Populations can differ in the degree of the LD between markers and QTL due to drift, selection, and/or recombination (Dekkers, 2004; Barton and Otto, 2005). Prediction accuracy should increase as LD between markers and QTL increases. Recombination in breeding populations should reduce LD between markers and QTL over the time while selection and genetic drift should increase LD (Pfaffelhuber et al., 2008). Habier et al. (2007) studied the effect of LD on prediction accuracy over many generations and found a decrease in prediction accuracy was associated with decay of LD. We found a general increase in adjacent marker LD in the progeny sets over time, while prediction accuracy generally remained constant or decreased. We also examined the persistence of adjacent marker LD between the parent set and each of the progeny sets using the correlation of r (de Roos et al., 2008; Toosi et al., 2010). The correlation of r did not decay over the window of time of this experiment, despite the fact that genetic distance between the parents and progeny sets increased over time. Asoro et al. (2011) suggested that the ability of early generations to predict later generations was due to the persistence of the LD phase between early and late generations. Thus, even if validation populations become more genetically distant from training populations, if the LD phase is consistent, prediction accuracy will be maintained.

Figure 6. Percentage of single nucleotide polymorphisms (SNPs) that are fixed in the complete marker set (genome-wide) and in the subsets of markers associated with deoxynivalenol (DON) concentration, Fusarium head blight (FHB) resistance, yield, and plant height (see Fig. 5) in each of the five progeny sets between 2006 and 2010.
How do Trait and Population Characteristics Affect Prediction Accuracy?

Ideally, GS can be applied to traits that vary in $H^2$ and genetic architecture. In our study based on estimates of $R^2$ and the $\pi$ parameter, yield was the most complex trait while plant height was the least complex. However, inference of genetic architecture based on $\pi$ should be interpreted cautiously (Gianola, 2013). We found that yield, a more complex and lower $H^2$ trait, generally had lower prediction accuracy than a simpler and higher $H^2$ trait such as plant height. This is consistent with other studies where complex traits controlled by many loci with small effects produced lower prediction accuracy than less complex traits (Hayes et al., 2010). Genomic predictions should be more accurate for traits with higher $H^2$ (Hayes et al., 2009; Daetwyler et al., 2010; Lorenz, 2013; Combs and Bernardo, 2013). Prediction accuracy for yield in the current study was higher than accuracy observed for yield in oat (Avena sativa L.; Asoro et al., 2011). They reasoned that lower accuracy for oat yield was due to the evaluation of their germplasm in a wider range of environments which reduced the genetic variance relative to the $G \times E$ variance, and thereby reduced $H^2$ for yield. The barley germplasm in the current study was evaluated in more homogenous environments, which are the target production and evaluation environments in Minnesota. Therefore, the genetic variance is expected to be higher relative to $G \times E$, leading to a higher $H^2$ estimate and an increased prediction accuracy.

In addition to the characteristics of trait, $H^2$, and genetic variance, LD (as discussed above) and population structure can affect prediction accuracy. These factors could have contributed to the striking difference in the response of accuracy to increased training population size for DON versus yield (Fig. 12). Both traits have higher than expected accuracies at very low training population sizes (e.g., 25 individuals). Windhausen et al. (2012) suggested that high accuracy at low training population size can be diagnostic of subpopulation structure affecting accuracy. In this context, we suggest that structure could reduce accuracy at high training population size from the following mechanism. Population structure is a cause of LD: two loci that both have differences in allele frequency across subpopulations will be in LD. Thus, structure can cause association between a marker and several QTL. This phenomenon has been an ongoing issue in genome-wide association studies (e.g., Pritchard et al., 2000). In the context of genomic prediction, structure-generated disequilibrium between a marker and several QTL will prevent the marker’s estimated effect from converging on the effect of a QTL to which it is actually linked, regardless of the training population size. Though they did not comment on it, Wimmer et al. (2013) observed a phenomenon like this: in their Fig. 6A, at low model complexity,
the error of marker effect estimates increases as training population size increases. This increase in error arises presumably because of the documented deep structure in rice (Zhao et al., 2011). The question remains as to why this mechanism would more strongly affect DON than yield. We hypothesize that structure in the Minnesota barley breeding program is more strongly correlated to DON than to yield, given that it has been purposefully split into a population where FHB resistance was prioritized versus one where yield and quality continued to be prioritized (Fang et al., 2013).

The availability of genome-wide markers can improve our understanding of genetic architecture and the extent to which epistasis influences complex traits. In general, the four genomic prediction models tested produced similar accuracies across the four traits investigated in this study. The four models differed in assumptions about the genetic architecture of the trait and the extent to which nonadditive effects contribute to the prediction. Lande and Thompson (1990) suggested the use of epistatic effects in addition to additive effects in MAS schemes. Liu et al. (2003) found that including epistasis improved both the response and efficiency of MAS. In some studies, including epistasis in genomic prediction models through the use of nonadditive kernels resulted in increased prediction accuracy over RR-BLUP (Crossa et al., 2010; Wang et al., 2012). In our study, we found that simple additive models (RR-BLUP and Bayes C\textsubscript{p}) performed similarly to those that account for both additive and nonadditive effects (Exponential and Gaussian). These results are similar to a recent study of barley breeding lines evaluated for DON concentration and FHB resistance that showed that both Bayes C\textsubscript{p} and RR-BLUP produced the same level of accuracy (Lorenz et al., 2012).

**Practical Implications for Breeding**

The increasing ease and rapidly declining cost of genotyping means that assembling phenotype data to train prediction models will be the limiting step in implementing GS. We found that using the contemporary parent data was slightly, but not significantly, better than using historic parent data to train a prediction model. The contemporary data was balanced and we corrected for field spatial variability using the common checks, whereas the historic data was unbalanced and no correction for field variability was made. Nevertheless, the prediction accuracy from historic data was, in most cases, around 0.50 for each of the 5 yr, and this level of accuracy suggests GS would be effective if the breeding cycle time is half of what is done in phenotypic selection (Asoro et al., 2011). These results suggest that breeders could reduce time and costs by using unbalanced historical data after proper adjustment for spatial variability and trial effects to train prediction models. Historical unbalanced phenotypic data were also used to assess the use of GS in oat (Asoro et al., 2011). Initiating GS with existing data and later incorporating contemporary data sets should allow breeders to realize benefits of GS sooner and improve effectiveness over time.

The size and composition of the training population are important factors to manipulate prediction accuracy. Breeders may consider combining training data sets to maximize the use of the available phenotypic and genotypic information and generate larger population sizes (Hayes et al., 2009; de Roos et al., 2009; Asoro et al., 2011; Lorenz et al., 2012; Technow et al., 2013). Lorenz et al. (2012) found little to no improvement in prediction accuracy for FHB resistance and DON concentration when increasing the size of the training population by combining different barley genotypes.
breeding populations. Conversely for maize, when combining both flint and dent heterotic groups together, prediction accuracies increased by 10 and 13% when predicting dent and flint heterotic groups, respectively, for Northern corn leaf blight resistance (Technow et al., 2013). In our study, we found only a slight improvement or a reduction in accuracy when increasing the population size by adding progeny sets from the same breeding program to the parent set. However, when adding progeny to the parent set, both the size and the composition of the training populations were altered. Therefore, we separated these two factors by generating training populations by randomly sampling from the combined data set. Interestingly, the prediction accuracy for DON concentration plateaued at a much smaller population size compared with yield (Fig. 12). Prediction accuracy for yield did not level off, suggesting that the benefit from increasing training population size may depend on the trait.

In addition to optimizing prediction accuracy, the effectiveness of GS will increase by shortening the breeding cycle time and reducing the cost of selection (Heffner et al., 2010; Jannink et al., 2010). In the University of Minnesota’s barley breeding program, GS is implemented at the F₃ stage for FHB resistance, DON concentration, and yield. This is 1 yr after crossing parents, compared with a 4-yr breeding cycle that is typical for phenotypic selection. The prediction accuracies that we observed based on progeny validation always exceeded 0.25, indicating that GS should exceed phenotypic selection in gain per unit time. Combined with rapidly decreasing genotyping costs, this suggests that GS should improve breeding efficiency substantially.

Supplemental Information Available
Supplemental information is included with this article. 

Supplemental Figure S1. Prediction accuracy for (A) DON accumulation, (B) FHB resistance, (C) yield, and (D) plant height using RR-BLUP, Exponential kernel method (EXP), Gaussian kernel method (GAUSS), and Bayes Cτ when using the parent set as a training population to predict the five progeny sets.

Supplemental Table S1. Number of experimental trials for the parent and five progeny sets for deoxynivalenol (DON) accumulation, Fusarium head blight (FHB) resistance, yield, and plant height. Each line was replicated twice in each experiment.
Supplemental Table S2. Mean prediction accuracy and p-values when using parent set as the training population and five progeny sets as the validation populations for deoxynivalenol (DON) accumulation, Fusarium head blight (FHB) resistance, yield, and plant height using RR-BLUP, Exponential kernel method, Gaussian kernel method, and Bayes Cπ.

Acknowledgments
We thank Ed Schiefelbein, Guillermo Velasquez, Karen Beaubien, Richard Horsley, the Minnesota Agricultural Experiment Station NW and WC Research and Outreach Centers, and the University of Minnesota Small Grains Pathology Lab for their contributions to conducting field trials and collecting data. In addition, we thank Shiaoman Chao and Yanhong Dong for SNP genotyping and toxin analysis, respectively. Additional thanks to Yang Da for helpful suggestions on an earlier draft of this manuscript. Funding for this work was supported by grants from the National Institute of Food and Agriculture USDA Award Number 2009-65300-05661, the U.S. Wheat and Barley Scab Initiative USDA–ARS Agreement No. 59-0206-9-072, USDA HATCH project MIN-13-030, and the Rahr Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the USDA.

Figure 11. Relationship between the predictive ability (correlation between genomic estimated breeding value and phenotypic performance) when using a contemporary parent data set to predict progeny sets using ridge regression best linear unbiased prediction (RR-BLUP), and heritability of the contemporary parent training population for plant height, deoxynivalenol (DON) concentration, Fusarium head blight (FHB) resistance, and yield.

Figure 12. Relationship between population size and prediction accuracy for deoxynivalenol (DON) concentration and yield. Three scenarios are presented: (A) Using the contemporary parent data set as the training population to predict the 2008 and 2010 progeny sets. Each point represents a subset of the training population by random sampling 500 times. (B) Using the combined contemporary parent data set and progeny sets before the validation set to predict the 2008 and 2010 sets. Each point represents a subset of the training population by random sampling 500 times. (C) The sequential addition of a progeny set to the contemporary parent data set as a training population to predict the 2008 and 2010 progeny sets. The training population sizes are 168 (parent set), 264 parent set + 2006 (n = 96), and so on.