Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program

Elliot L. Heffner, Jean-Luc Jannink, and Mark E. Sorrells*

Abstract
Genomic selection (GS) uses genome-wide molecular marker data to predict the genetic value of selection candidates in breeding programs. In plant breeding, the ability to produce large numbers of progeny per cross allows GS to be conducted within each family. However, this approach requires phenotypes of lines from each cross before conducting GS. This will prolong the selection cycle and may result in lower gains per year than approaches that estimate marker-effects with multiple families from previous selection cycles. In this study, phenotypic selection (PS), conventional marker-assisted selection (MAS), and GS prediction accuracy were compared for 13 agronomic traits in a population of 374 winter wheat (Triticum aestivum L.) advanced-cycle breeding lines. A cross-validation approach that trained and validated prediction accuracy across years was used to evaluate effects of model selection, training population size, and marker density in the presence of genotype × environment interactions (G×E). The average prediction accuracies using GS were 28% greater than with MAS and were 95% as accurate as PS. For net merit, the average accuracy across six selection indices for GS was 14% greater than for PS. These results provide empirical evidence that multifamily GS could increase genetic gain per unit time and cost in plant breeding.

Quantitative traits such as grain yield have proven difficult to improve with marker-assisted selection (MAS). The main limitations are (i) small population sizes and conventional statistical methods that have inadequate power to detect and accurately estimate effects of small-effect quantitative trait loci (QTL) and (ii) gene × gene interactions (epistasis) and genotype × environment interactions (G×E) that have limited the transferability of QTL effect estimates across populations and environments (reviewed by Bernardo [2008]; Xu and Crouch, 2008). These limitations can be mitigated in plant breeding with improved marker-based breeding methods such as genomic selection (GS) (Meuwissen et al., 2001) and with “mapping-as-you-go” approaches that continually re-estimate marker effects in breeding populations and target...
environments in parallel with the selection process (Podlich et al., 2004; Breseghello and Sorrells, 2006).

Genomic selection addresses the first limitation by using a random-effects approach to jointly estimate all marker effects without significance testing to capture small-effect QTL that are excluded by conventional MAS (Meuwissen et al., 2001). Marker estimates for GS are derived from a “training population” (TP), composed of breeding material with both phenotypic and genome-wide marker data. Marker estimates are then used to calculate genomic estimated breeding values (GEBVs) of new breeding lines in the “selection population” (SP). The combination of affordable, high-throughput genotyping and GS prediction methods has resulted in marker-based prediction accuracies that are revolutionizing cattle breeding (reviewed by Hayes et al. [2009a] and Calus [2010]) and show great promise for increasing gains from selection in plant breeding (reviewed by Heffner et al. [2009] and Jannink et al. [2010]).

A “mapping-as-you-go” approach addresses the second limitation by re-estimating marker-effects in new breeding populations across target environments to capture changes in epistasis and G×E that will result from shifts in genetic backgrounds caused by selection (Podlich et al., 2004; Breseghello and Sorrells, 2006). Well-funded breeding programs are able to maximize the context specificity of marker effect estimates by conducting MAS within each new cross (e.g., a biparental population), as they have the resources to generate large progeny numbers for each cross and have extensive multi-environment testing regimens (Sebastian et al., 2010). Unfortunately, this approach will also require the phenotyping of a subset of progeny from each cross before performing marker-based selection and will fail to leverage data generated from previous breeding cycles. Alternatively, MAS cycle times can be reduced to increase gains by avoiding this phenotyping step through estimating marker effects with data across multiple families in a breeding program (Jannink et al., 2001; Rafalski, 2002; Breseghello and Sorrells, 2006; Heffner et al., 2010). While a multifamily approach will be less “population specific” and may increase error due to epistasis, this approach should reduce error due to G×E as it can leverage multiyear data thereby providing a greater sample of target environmental conditions (Podlich et al., 2004; Breseghello and Sorrells, 2006; Heffner et al., 2009).

Genomic selection within each cross, herein referred to as biparental GS, has been shown to achieve higher prediction accuracies than conventional MAS in biparental populations both in simulations (Bernardo and Yu, 2007; Wong and Bernardo, 2008) and in empirical studies (Lorenzana and Bernardo, 2009; Heffner et al., unpublished data, 2010). Biparental GS also has been found to compare favorably to phenotypic selection (PS). Lorenzana and Bernardo (2009) reported biparental GS accuracies for maize (Zea mays L.), barley (Hordeum vulgare L.), and Arabidopsis thaliana (L.) Heynh. that would approach 1.5 times more gain than PS when using year-round nurseries capable of three GS cycles per year. Heffner et al. (unpublished data, 2010) reported GS prediction accuracies (rGS) for nine quality traits in winter wheat (Triticum aestivum L.). Even when using TPs as small as 48 inbred lines and a maximum of only two GS cycles per year for winter wheat, their results also suggest that biparental GS would outperform PS.

Performance of GS in plant populations using marker effects estimated from multiple families, herein called multifamily GS, is limited. In a simulation study that used empirical marker data and simulated phenotypes for two-row barley, Zhong et al. (2009) reported a rGS of ~0.60 for a trait controlled by 80 QTL with a heritability (h2) of 0.40. Using 2 yr of phenotypic data from 1700 maize hybrids to predict 288 new hybrid combinations grown in two different years, van Eeuwijk et al. (2010) reported a rGS of ~0.70 for ear height in maize (h2 = 0.36). Finally, Crossa et al. (2010) used cross-validation to evaluate GS in 599 historical wheat lines and 284 maize inbreds from the International Maize and Wheat Improvement Center (CIMMYT). Using multiple GS models and environments, rGS for wheat grain yield ranged from 0.36 to 0.61, for maize flowering time ranged from 0.46 to 0.79, and for maize grain yield ranged from 0.42 to 0.53 (Crossa et al., 2010). Results from these studies strongly support the utility of GS in plant breeding because deterministic simulation has shown that if rGS for net merit (i.e., overall performance) exceeds 0.50, GS could greatly outperform conventional MAS in terms of gain per unit time and cost (Heffner et al., 2010).

Empirical comparisons between conventional MAS using markers identified by association mapping and multifamily GS in plant breeding programs are currently unavailable. However, association mapping of human height provides an empirical example of the difficulty of capturing genetic variance for highly polygenic traits with fixed effect models that have stringent thresholds. Despite a high h2 (~0.80) and a TP of tens of thousands of individuals, only ~5% of the phenotypic variance for human height has been accounted for with ~50 significant markers (Gudbjartsson et al., 2008; Lettre et al., 2008; Weedon et al., 2008; Visscher, 2008). In contrast, Yang et al. (2010) fit all 300,000 markers simultaneously as random effects and found that using all markers explained 45% of phenotypic variation for human height in a population of ~4,000 unrelated individuals. Thus, they concluded a large proportion of the genetic variance was explained with small-effect markers that do not pass stringent thresholds. Relaxing these thresholds can improve the amount of genetic variance explained with significant markers (Hospital et al., 1997; Moreau et al., 1998). Small population sizes and low heritability, however, will still cause small-effect markers to be below significance thresholds leaving some genetic variance unexplained. In addition, significant QTL effects will be overestimated (Beavis, 1998; Schön et al., 2004; Xu, 2003). This suggests that GS models should outperform conventional MAS models in plant populations composed of multiple families.
The objective of this study was to empirically compare phenotypic prediction accuracy ($r_p$), conventional MAS prediction accuracy ($r_{\text{MAS}}$), and $r_{\text{GS}}$ when marker effects were estimated with multifamily data from a breeding program. To meet this objective, cross-validation across years was performed in a population of 374 elite wheat breeding lines using genome-wide marker data and several marker-based prediction models to predict performance of 13 agronomic traits. For each trait, training population size ($N_{\text{TP}}$) and marker number ($N_{\text{MA}}$) were varied to investigate their effects on prediction accuracy. Finally, $r_{\text{GS}}$ and $r_p$ for net merit were compared using index selection.

**MATERIALS AND METHODS**

**Data**

A population of 374 soft winter wheat varieties and F$_2$–derived advanced breeding lines resulting from many different crosses in the Cornell University Wheat Breeding Program (Ithaca, NY) were analyzed in this study. Lines were genotyped with 5000 Diversity Array Technology (DArT) markers (Triticarte Pty. Ltd., Yaralumla, ACT, Australia), resulting in 1544 polymorphic markers. Some markers were perfectly correlated to each other due to complete linkage disequilibrium (LD) ($r^2 = 1$). Therefore, the data set was trimmed to 1158 markers by selecting the marker with the least missing data from each pair or group of markers that were in complete LD. The frequency of missing marker data was low (3.2%) and was imputed as the mean marker score for each marker as precise map position was unknown for many of the markers.

Phenotypic data for 13 traits were analyzed: grain yield, plant height, heading date (i.e., days to heading), lodging, preharvest sprouting (PHS), flour yield, flour protein, softness, sucrose solvent retention capacity (Suc-SRC), water SRC ($H_2O$-SRC), lactic acid SRC (LA-SRC), and sodium carbonate SRC (NaCO-SRC). Preharvest sprouting is the premature germination of seeds while still attached to the mother plant that decreases grain value and was measured as described by Anderson et al. (1993) and Munkvold et al. (2009). Milling and baking quality traits were measured as follows: flour yield:percentage of flour obtained from milling, softness:percentage of fine flour obtained, that is, that which can pass through a 94-mesh (180 μm) screen, protein:percent protein of flour measured using a near infrared analyzer (Unity SpectraStar 2200, Columbia, MD), and SRC:amount of soluble retained by the flour after centrifugation and draining. The four SRC tests were used to predict overall baking quality: $H_2O$-SRC for global water absorption, NaCO-SRC for damaged starch, Suc-SRC for arabinoxylan and partially hydrated gliadin content, and LA-SRC for gluten strength. The USDA-ARS Soft Wheat Quality Laboratory in Wooster, OH, performed all milling and baking quality tests as described by Guttieri et al. (2008).

Phenotypic data were collected from field trials in 2 yr, 2008 and 2009, with three locations per year near Ithaca, NY. Each year, two locations had yield plots (1.26 m by 4 m) and one location had single 1 m rows. All traits were measured in yield trials and PHS, height, and heading date were also measured in single row trials. Each location was arranged in an unreplicated augmented design (Federer, 1956) with six check varieties replicated 10 times each. A two-stage analysis was used to calculate line best linear unbiased predictions (BLUPs) because it is less computationally demanding than a one-stage analysis and has been shown to generate similar results (Mohring and Piepho, 2009). First, BLUPs were calculated for each trait in each location with a two-dimensional, first-order autoregressive (AR1 x AR1) spatial model with lines as random effects in ASReml (Gilmour et al., 2009). For PHS, an additional random effect of harvest date was included. Second, line BLUPs were calculated for each year with random effects of line and location. Only the first stage was used for $H_2O$-SRC and NaCO-SRC because they were only measured in two locations in 2008.

**Prediction Models**

Six methods were used to estimate marker effects for marker-based prediction: association analysis (AA), association analysis including kinship as a covariate (AK), ridge regression best linear unbiased prediction (RR), BayesA (BA), BayesB (BB) (Meuwissen et al., 2001), and BayesC$\pi$ (BC) (Dekkers et al., 2009). All statistical procedures herein were executed using R (R Development Core Team, 2009).

**Conventional Marker-Assisted Selection Models**

A two-stage approach using AA and multiple linear regression (MLR) was used to represent conventional MAS using multifamily data. That is, AA first reduced the number of markers (predictor variables), and MLR then selected markers to be included in the final prediction model and estimated the marker effects (regression coefficients) used to calculate GEBVs. Association analysis and AK modeled environments and markers as fixed effects, and AK had an additional random covariate, a simple identity-by-state allele sharing kinship matrix ($K$), to account for genetic covariance among individuals to reduce the number of false positive marker–trait associations caused by population structure and genetic relatedness (Zhao et al., 2007; Kang et al., 2008). Calculation of $K$ and detection of marker–trait associations were performed with the R package “emma” (efficient mixed-model association) (Kang et al., 2008). A significance threshold of 0.05 was used for AA and AK because relaxed thresholds have been shown to increase marker-based prediction accuracy (Hospital et al., 1997; Moreau et al., 1998). Relaxed thresholds were also used in MLR for forward (0.2) and backward (0.2) variable selection.

**Genomic Selection Models**

Four GS models were used in this study. Ridge regression best linear unbiased prediction assumes all markers have a common variance (Meuwissen et al., 2001; Whittaker...
et al., 2000), and thus shrinks each marker effect equally toward zero. Ridge regression best linear unbiased prediction is equivalent to estimating markers effects with a realized-relationship matrix determined from markers (Habier et al., 2007; Goddard, 2009; Piepho, 2009). The additive realized-relationship matrix was estimated in R, and the R package “emma” (Kang et al., 2008) was used to estimate the variance components to solve mixed-model equations (Henderson, 1984).

Three Bayesian models were used to address the simple, but likely unrealistic RR assumptions of all markers having nonzero effects and equal marker variances. BayesA fits all markers but allows each marker to have its own variance (Meuwissen et al., 2001). In addition to allowing for unique marker variance, BB also specifies that a portion of the markers have no effect (π) (Meuwissen et al., 2001). Thus, BB is equivalent to BA when π = 0. Finally, BC assumes common marker variances and allows for some markers to have no effect (Dekkers et al., 2009; Jannink, 2010). Additionally, BC jointly estimates π from the training data to avoid an incorrect π that can negatively affect prediction accuracy (Verbyla et al., 2010; Gianola et al., 2009). We adapted BA, BB, and BC code written by R.L. Fernando (Dekkers et al., 2009). For BC, a starting π = 0.50 was used. For BB, π = 0.90 was used because preliminary results showed that π values of 0.95 and 0.975 generally decreased accuracy. Each method was run for 2000 iterations and had a burn-in period of 200 iterations. This was considered sufficient for approximate convergence, as the average correlation of results from two independent runs of each of 40 random TPs for each trait and each model was greater than 0.99.

Prediction Accuracy and Cross-Validation

Phenotypic prediction accuracy (r_p) was the correlation of the observed phenotypes from 2008 and 2009. Marker-based r_MAP and r_GS was the correlation of GEBVs from one year and the observed phenotypes from the other year. Genomic estimated breeding values were calculated as GEBV = X ⃗g, in which GEBV was the GEBV of line i, X was the vector of the marker scores for that line, and ⃗g was the vector of marker effects obtained from the TP using a marker-based prediction model.

Three different training population sizes (N_TP = 288, 192, and 96) were used for marker-based prediction with N_m = 1158. Multiples of 96 were used to correspond with current 96- or 384-well DNA sample plates. As overall population size was 374 and maximum N_TP was 288, GEBVs were calculated for 86 lines with marker effects estimated from the TP. The observed phenotypes and GEBVs of the 86 lines from one year were correlated to observed phenotypes of the other year to obtain prediction accuracies and to avoid bias introduced by genotype × year interactions. To achieve adequate sampling of the genetic diversity both for training and validation, TP lines were randomly and proportionally sampled from six genetic clusters of size 48, 79, 95, 38, 50, and 64. Cluster assignment and selection of optimal cluster number and model using the Bayesian information criterion were done using the R package “mclust” (Fraley and Raftery, 2002, 2006).

Four marker densities (N_m = 1158, 768, 384, and 192) were used to assess the impact of N_m on prediction accuracy when using a N_TP = 288. Kinship matrix (K)-means clustering (Hartigan and Wong, 1978) was used to select informative marker subsets to minimize LD between selected markers and maximize genome coverage. Markers were selected using the function “kmeans” in R in which each marker was treated as an explanatory variable, the number of clusters equaled the N_m, the number of random starts equalled 1000, and the marker closest to the centroid of each cluster was chosen. Significance thresholds for AA and AK were relaxed from 0.05 (N_M = 1158) to 0.1, 0.2, and 0.3 for N_M = 768, 384, and 192, respectively.

All six marker-based prediction methods were evaluated for N_TP = 288 and N_m = 1158 for 100 TPs using training data in 2008 and 2009 for a total of 1200 analyses for each trait. Because the Bayesian models were computationally intensive for cross-validation and a preliminary analysis showed all GS methods produced similar trends, only two GS models (RR and BC) were used for investigation of the effects of N_TP and N_M. Therefore, four models (AA, AK, RR, and BC) were used for N_TP = 192 and 96 with N_M = 1158 and N_m = 768, 384, and 192 with N_TP = 288. As before, each scenario was analyzed with each model on 100 TPs using training data in 2008 and 2009, totaling an additional 4000 analyses for each trait. Reported r_p, r_MAP, and r_GS for each trait was the average accuracy for all 100 TPs across both years, and prediction methods were compared using a paired t test (α = 0.01).

Net Merit Prediction Accuracy

To predict net merit, trait predictions were combined using weighting determined by the “Smith-Hazel index” (Smith, 1936; Hazel, 1943) and by the “base index” (Panse, 1946; Brim et al., 1959; Williams, 1962). The estimated Smith-Hazel index is \( \mathbf{b} = \mathbf{P}^{-1} \mathbf{G} \), in which \( \mathbf{b} \) is the vector of estimated trait weights, a is the vector of relative economic trait weights, \( \mathbf{P} \) is the estimated phenotypic covariance matrix, and \( \mathbf{G} \) is the estimated additive-genetic covariance matrix. The base index ignores phenotypic and genetic covariances; thus, trait predictions are weighted simply by their relative economic trait weights.

Three economic weighting indices were used: (i) emphasis on yield, (ii) emphasis on milling and baking quality traits, and (iii) a “balanced” index representing current breeding goals of the Cornell University Wheat Breeding Program. Water SRC and NaCO-SRC were excluded from the indices, as they were only measured in two locations in 2008. The phenotypic covariance matrix (\( \mathbf{P} \)) was estimated using line BLUPs calculated using phenotypes for each line and each trait from all four locations. The additive-genetic covariance matrix (\( \mathbf{G} \)) was estimated using GEBVs for each trait and each line that were calculated from genotypic data and trait BLUPs from
all 374 lines using RR. Phenotypic prediction accuracy and RR were used for comparing prediction accuracy for each index. For the Smith-Hazel index, \(r_p = \text{cor}(\text{Ph}_a, \text{Ph}_b)\) and the prediction accuracy of ridge regression \(r_{RR}^2\) = \text{cor}(\text{GEBV}_a, \text{Ph}_b), \) in which \(\text{Ph}_a\) is a vector of observed phenotypes from 1 yr, \(\text{GEBV}_a\) is a vector of GEBVs from the 1 yr, and \(\text{Ph}_b\) is a vector of observed phenotypes from the other year. For the base index, \(r_p = \text{cor}(\text{Ph}_a, \text{Ph}_a)\) and \(r_{RR} = \text{cor}(\text{GEBV}_a, \text{Ph}_a)\). All phenotypes were standardized to mean zero with a SD of 1 before index analyses. For interpretation of index results with the models in Heffner et al. (2010), net merit accuracies were also divided by the square root of the broad-sense heritability (H) of net merit in validation data on a line-mean basis to account for the validation phenotypes not being equal to the “true genetic value” (Dekkers, 2007). This correction was not used for each trait individually, as this was unneeded to compare \(r_p\), \(r_{MAS}\), and \(r_{GS}\) and can bias accuracies due to error in estimates of heritabilities in the validation data.

**RESULTS**

**Marker-Based and Phenotypic Prediction Accuracy**

The \(r_{GS}\) was greater than \(r_{MAS}\) for all 13 traits studied (Table 1). For the maximum \(N_m\) (1158) and \(N_{TP}\) (288), mean \(r_{GS}\) (0.58) across all methods and traits was 28% greater than \(r_{MAS}\) (0.46). A large range of \(r_{GS}\) was observed, from 0.17 (grain yield; BC) to 0.76 (LA-SRC; BA). The range for \(r_{MAS}\) was 0.18 (grain yield; AK) to 0.63 (Suc-SRC; AA). Only slight differences were detected between the GS models, with BA having the highest mean accuracy across all traits. Accuracies of BA and RR were most similar, as BA was significantly different from RR only for grain yield (BA = 0.22 versus RR = 0.20). In most cases where BB and BC were significantly different from BA and RR, their \(r_{GS}\) was marginally lower than RR and BA, but differences were quite small. The best conventional MAS method was AA, which was significantly greater than AK for 6 of the 13 traits.

A wide range of \(r_p\) was also observed, ranging from 0.21 (grain yield) to 0.89 (heading date), reflecting the wide range of broad-sense heritability (H) for the traits in this study (Supplemental Table S1). When comparing the \(r_p\) to the \(r_{GS}\) (using BA which had the highest mean \(r_{GS}\)), \(r_p\) was greater for nine traits and not significantly different for four traits (test weight, H2O-SRC, lodging, and grain yield; Table 1). The ratio of \(r_{GS}\) and \(r_p\) (\(r_{GS}/r_p\)) ranged from 0.84 (heading date) to 1.08 (lodging) with a mean ratio of 0.95. When comparing the \(r_p\) to \(r_{MAS}\) (using AA which had the highest mean \(r_{MAS}\)), \(r_p\) was greater than \(r_{MAS}\) for all traits with \(r_{MAS}/r_p\) ranging from 0.56 (height) to 0.91 (grain yield) with a mean \(r_{MAS}/r_p\) of 0.76. Finally, the slope from linear regression of \(r_{GS}\) on \(r_p\) (0.81) and \(r_{MAS}\) on \(r_p\) (0.64) showed that \(r_{GS}/r_p\) and \(r_{MAS}/r_p\) decreased as \(r_p\) increased (Fig. 1).

**Effects of Training Population Size and Marker Number**

Decreasing \(N_{TP}\) had a strong negative effect on the mean accuracy across all traits for each of the four prediction models tested (Fig. 2; Supplemental Table S2). Decreasing \(N_{TP}\) from 288 to 198 and 96 reduced the average \(r_p\) by 11 and 30% and \(r_{MAS}\) by 24 and 35%, respectively (Fig. 2). Reducing \(N_m\) from 1158 to 768 and 384 resulted in a small decrease in \(r_{GS}\) and a small increase \(r_{MAS}\) (Fig. 3; Supplemental Table S3), while reducing \(N_m\) from 1158 to 192 reduced the average \(r_{GS}\) by 10% and \(r_{MAS}\) by less than 4%.

**Table 1. Phenotypic and marker-based prediction accuracy for 13 traits. Marker-based prediction was based on training population size \((N_{TP}) = 288.**

<table>
<thead>
<tr>
<th>Trait†</th>
<th>(r_p)</th>
<th>(r_{AA})</th>
<th>(r_{AX})</th>
<th>(r_{BX})</th>
<th>(r_{BA})</th>
<th>(r_{BY})</th>
<th>(r_{BC})</th>
<th>(r_{GS}/r_p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flour yield</td>
<td>0.831</td>
<td>0.648</td>
<td>0.544</td>
<td>0.762</td>
<td>0.761</td>
<td>0.740</td>
<td>0.761</td>
<td>0.917</td>
</tr>
<tr>
<td>H2O-SRC</td>
<td>0.607</td>
<td>0.484</td>
<td>0.475</td>
<td>0.577</td>
<td>0.585</td>
<td>0.570</td>
<td>0.556</td>
<td>0.963</td>
</tr>
<tr>
<td>Heading date</td>
<td>0.891</td>
<td>0.551</td>
<td>0.431</td>
<td>0.748</td>
<td>0.750</td>
<td>0.724</td>
<td>0.746</td>
<td>0.842</td>
</tr>
<tr>
<td>Height</td>
<td>0.851</td>
<td>0.521</td>
<td>0.467</td>
<td>0.740</td>
<td>0.746</td>
<td>0.719</td>
<td>0.739</td>
<td>0.877</td>
</tr>
<tr>
<td>LA-SRC</td>
<td>0.786</td>
<td>0.626</td>
<td>0.625</td>
<td>0.750</td>
<td>0.757</td>
<td>0.754</td>
<td>0.751</td>
<td>0.963</td>
</tr>
<tr>
<td>Lodging</td>
<td>0.262</td>
<td>0.190</td>
<td>0.186</td>
<td>0.278</td>
<td>0.281</td>
<td>0.284</td>
<td>0.232</td>
<td>1.073</td>
</tr>
<tr>
<td>NaCO-SRC</td>
<td>0.726</td>
<td>0.592</td>
<td>0.559</td>
<td>0.658</td>
<td>0.680</td>
<td>0.671</td>
<td>0.656</td>
<td>0.937</td>
</tr>
<tr>
<td>PHS</td>
<td>0.578</td>
<td>0.465</td>
<td>0.449</td>
<td>0.553</td>
<td>0.554</td>
<td>0.552</td>
<td>0.553</td>
<td>0.958</td>
</tr>
<tr>
<td>Protein</td>
<td>0.493</td>
<td>0.377</td>
<td>0.303</td>
<td>0.449</td>
<td>0.459</td>
<td>0.468</td>
<td>0.450</td>
<td>0.919</td>
</tr>
<tr>
<td>Softness</td>
<td>0.774</td>
<td>0.529</td>
<td>0.507</td>
<td>0.664</td>
<td>0.670</td>
<td>0.649</td>
<td>0.664</td>
<td>0.866</td>
</tr>
<tr>
<td>SucSRC</td>
<td>0.799</td>
<td>0.627</td>
<td>0.568</td>
<td>0.736</td>
<td>0.741</td>
<td>0.727</td>
<td>0.743</td>
<td>0.927</td>
</tr>
<tr>
<td>Test weight</td>
<td>0.568</td>
<td>0.470</td>
<td>0.455</td>
<td>0.560</td>
<td>0.571</td>
<td>0.560</td>
<td>0.554</td>
<td>1.005</td>
</tr>
<tr>
<td>Grain yield</td>
<td>0.205</td>
<td>0.186</td>
<td>0.180</td>
<td>0.199</td>
<td>0.223</td>
<td>0.221</td>
<td>0.174</td>
<td>1.088</td>
</tr>
<tr>
<td>Mean</td>
<td>0.644</td>
<td>0.482</td>
<td>0.442</td>
<td>0.590</td>
<td>0.598</td>
<td>0.586</td>
<td>0.582</td>
<td>0.949</td>
</tr>
</tbody>
</table>

†H2O-SRC, water solvent retention capacity; LA-SRC, lactic acid solvent retention capacity; NaCO-SRC, sodium carbonate solvent retention capacity; PHS, preharvest sprouting; SucSRC, sucrose solvent retention capacity.

\(r_p\), prediction accuracy of association analysis; \(r_{MAS}\), prediction accuracy of association analysis using kinship; \(r_{GS}\), genomic selection prediction accuracy; \(r_{GS}/r_p\), ratio of \(r_{GS}\) and \(r_p\).
Prediction Accuracy for Net Merit

Prediction accuracy using the base index and the “yield,” “balanced,” and “quality” economic weights (Table 2) was greater for GS (RR) than for PS (Table 3). For both GS and PS, the balanced weights had the lowest prediction accuracy before adjusting for error in the validation data. After adjusting the validation data for error by dividing by $H$ (Dekkers, 2007), yield and balanced weights had equivalent accuracies and the quality weights had the highest accuracy. The Smith-Hazel index resulted in higher $r_{GS}$ than $r_{PS}$ for the quality weights, equal accuracy for the balanced weights, and lower accuracy for the yield weights. The Smith-Hazel index using the yield weights had the highest accuracy while the quality weights had the lowest accuracy before and after correction using $H$. The mean $r_{GS}/H$ for all indices (0.54) was greater than the mean $r_{PS}/H$ (0.47) as only one index (Smith-Hazel – yield) resulted in a $r_p/H$ greater than $r_{GS}/H$.

DISCUSSION

Comparison of Marker-Based Prediction Methods

In this study, prediction accuracy using GS was superior to conventional MAS in a wheat breeding population composed of multiple families. Averaged across the 13 traits, the average $r_{GS}$ (0.59) was 28% higher than the average $r_{MAS}$ (0.46; Table 1). Clearly, the GS approach of jointly estimating all marker effects was able to capture more of the genetic variance than the two-stage conventional MAS approach that first selected significant markers and then estimated their effects. This result is consistent with previous empirical studies using biparental plant populations (Lorenzana and Bernardo, 2009; Heffner et al., unpublished data, 2010) and multiple-family animal populations (Moser et al., 2009) and provides additional empirical evidence that GS will increase the accuracy of marker-based selection in plant breeding.

Four GS models that each had different prior assumptions for marker effect and variance distributions were used in this study to investigate the effect of these assumptions on $r_{GS}$. Despite these model differences, the mean $r_{GS}$ across models for all traits ranged from only 0.58 to 0.60 (Table 1). This similarity in model performance is consistent with other empirical GS studies (Verbyla et al., 2010; VanRaden et al., 2009; Hayes et al., 2009a, b; Moser et al., 2009; Luan et al., 2009; Lorenzana and Bernardo, 2009; Su et al., 2010; Heffner et al., unpublished data, 2010). Some differences between GS models, however, were significant ($\alpha = 0.01$) as $r_{GS}$ standard errors were small because 100 TPs were sampled for each method. Nevertheless, differences were small in this study, and it was concluded that $r_{GS}$ was generally not influenced by GS model choice.

The observed similarity in model performance results from the interaction of two key factors: (i) effective
population size ($N_e$) and (ii) trait architecture (Daetwyler et al., 2010). A population’s $N_e$ determines the number of independent chromosomal segments ($M_e$), where different independent segments trace back to different ancestors (Hayes et al., 2009c). The trait architecture refers to the number of QTL ($N_{QTL}$) and the distribution of their effects. The Bayesian models used in this study assign a portion of the genetic variance. In contrast, when $N_{QTL} > M_e$ (or equivalently, when all QTL effects are small; i.e., the infinitesimal model holds), the ability for differential shrinkage is unneeded and RR can perform as well as or better than the Bayesian models. To consider the effect of the genetic architecture on RR, it is more useful to recognize that RR is equivalent to modeling genetic relationships with markers (Habier et al., 2007; Goddard, 2009; Hayes et al., 2009c). The accuracy of this modeling is not affected by the genetic architecture of the trait and $N_{QTL}$ does not affect RR accuracy (Daetwyler et al., 2010). In summary, the similarity in GS model performance suggests (i) $N_e$ was at least several hundred, (ii) these traits will be best predicted by GS models that capture the effects of a large number of QTL, and (iii) traits will be best predicted by GS models that capture the effects of a large number of QTL.

Model selection should also consider relatedness between the TP and SP because genetic relationships deteriorate with each generation separating the two populations. Models that rely more on marker–QTL LD, for example, Bayesian models such as those used here, should produce higher accuracies than RR in scenarios where the TP and SP are separated by multiple generations (Habier et al., 2007; Zhong et al., 2009; Meuwissen, 2009). To conduct marker-based selection on selection candidates that do not have phenotypes, the TP and SP will be separated at least one selection cycle. However, in practice, the lag between TP and SP may be greater because several GS cycles may occur while selected lines go through seed increases and/or inbreeding cycles before entering the TP to update the model (e.g., Heffner et al., 2010). Also, TPs that span several cycles of selection and have greater genetic diversity would result in larger TPs, $N_e$ and, consequently, $M_e$. In such cases, $N_{QTL}$ may be less than $M_e$ even for highly polygenic traits, suggesting that Bayesian models would be more accurate than RR. Nevertheless, updating the GS model each selection cycle should maintain genetic similarity between the TP and SP. Thus, a significant portion of $r_{GS}$ in plant breeding programs may still result from capturing genetic relationships with markers.

### Effects of Training Population Size and Marker Number

Reductions in $N_{TP}$ resulted in rapid decreases in accuracy for both conventional MAS and GS (Fig. 2; Supplemental Table S2). A decrease in accuracy was expected because it is well known that increasing $N_{TP}$ improves the estimation of marker effects (e.g., Knapp and Bridges, 1990) and accuracy (e.g., VanRaden et al., 2009; Hayes et al., 2009a; Zhong et al., 2009). Meuwissen (2009) predicted that the TP size must approach $10 \times N_e \times L$ to reach $r_{GS} = 0.90$ and

### Table 2. Economic weight indices for index selection.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Yield</th>
<th>Balanced</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flour yield</td>
<td>1.1</td>
<td>2.1</td>
<td>4.2</td>
</tr>
<tr>
<td>Heading date</td>
<td>−1.0</td>
<td>−1.0</td>
<td>−1.0</td>
</tr>
<tr>
<td>Height</td>
<td>−4.0</td>
<td>−4.0</td>
<td>−4.0</td>
</tr>
<tr>
<td>LA-SRC</td>
<td>−1.1</td>
<td>−2.1</td>
<td>−4.2</td>
</tr>
<tr>
<td>Lodging</td>
<td>−20.0</td>
<td>−20.0</td>
<td>−20.0</td>
</tr>
<tr>
<td>PHS</td>
<td>−5.0</td>
<td>−10.0</td>
<td>−10.0</td>
</tr>
<tr>
<td>Protein</td>
<td>−0.8</td>
<td>−1.6</td>
<td>−3.2</td>
</tr>
<tr>
<td>Softness</td>
<td>1.1</td>
<td>2.1</td>
<td>4.2</td>
</tr>
<tr>
<td>Suc-SRC</td>
<td>−1.1</td>
<td>−2.1</td>
<td>−4.2</td>
</tr>
<tr>
<td>Test weight</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
</tr>
<tr>
<td>Grain yield</td>
<td>50.0</td>
<td>40.0</td>
<td>30.0</td>
</tr>
</tbody>
</table>

†LA-SRC, lactic acid solvent retention capacity; PHS, preharvest sprouting; Suc-SRC, sucrose solvent retention capacity.

### Table 3. Phenotypic and genomic selection (GS) prediction accuracy for net merit using the base and Smith-Hazel indices for each economic weight index.

<table>
<thead>
<tr>
<th>Trait</th>
<th>$r_{e}^{1}$</th>
<th>$r_{e}^{1}$</th>
<th>$H^{4}$</th>
<th>$r_{e}/H^{4}$</th>
<th>$r_{GS}/H^{4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>0.16</td>
<td>0.21</td>
<td>0.49</td>
<td>0.32</td>
<td>0.44</td>
</tr>
<tr>
<td>Balanced</td>
<td>0.14</td>
<td>0.20</td>
<td>0.45</td>
<td>0.32</td>
<td>0.44</td>
</tr>
<tr>
<td>Quality</td>
<td>0.17</td>
<td>0.21</td>
<td>0.48</td>
<td>0.38</td>
<td>0.48</td>
</tr>
<tr>
<td>Smith-Hazel</td>
<td>0.33</td>
<td>0.31</td>
<td>0.45</td>
<td>0.73</td>
<td>0.68</td>
</tr>
<tr>
<td>Balanced</td>
<td>0.29</td>
<td>0.28 NS</td>
<td>0.46</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>Quality</td>
<td>0.23</td>
<td>0.28</td>
<td>0.51</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>Mean</td>
<td>0.22</td>
<td>0.25</td>
<td>0.47</td>
<td>0.47</td>
<td>0.54</td>
</tr>
</tbody>
</table>

$^{1}$, phenotypic prediction accuracy. $^{2}$, genomic prediction accuracy. $^{3}$, square root of the broad-sense heritability ($H^2$) for net merit in validation data. $^{4}$, Accuracy divided by $H$ to adjust for error in the validation data.

NS, not significant. $r_{e}$ and $r_{GS}$ were not significantly different for $\alpha = 0.01$. 

Assuming $N_{QTL} \geq M_e$, an estimate of $M_e$ can then be used to predict a lower bound for $N_{QTL}$ underlying the traits analyzed in this study (Daetwyler et al., 2010). In this population, LD decayed below an $r^2 = 0.2$ at ~1.5 cM suggesting that the $N_e$ was ~65 individuals using $1/(1 + 4N_e c)$, in which $c$ is the recombination frequency (Sved, 1971). Using Goddard’s (2009) theoretical approximation of $M_e = 1.42N_e L/\log (4N_e L)$, in which $L$ was the genome length of wheat in Morgans (~30 Morgans), $M_e$ was estimated to be ~1,000. While this estimate of $M_e$ is a very rough approximation, this estimate along with the observed similarities of model performance suggests (i) $M_e$ was at least several hundred, (ii) the 13 traits studied were highly polygenic,
from increasing NM has been seen in other empirical realizations only if NTP scales with NM (Muir, 2007). Scaling SP will be closely related and high rGS may be achievable testing regimes (e.g., Eathington et al., 2007; Sebastian et al., 2010). In most cases, plant breeders will retrain models frequently for calculating GEBVs; thus, TP and SP will be closely related and high rGMS may be achieved with NTP, far smaller than 10 × N × L (Meuwissen, 2009). In addition to the effect of N, and L, NTP requires will be affected by trait heritability, particularly when trait heritabilities are low (i.e., h² < 0.40; Hayes et al., 2009c). In this study, many traits had low H², NTP was clearly below the requirements estimated above, and accuracy showed a near linear increase with increased NTP. Therefore, considerable improvements in rGMS should have been achieved in this study if NTP was increased.

Accuracy of Predicting Net Merit
In this study, GS was comparable to PS on an individual trait basis; however, a breeder’s primary goal is improving net merit—a single character calculated as a linear combination of all economically important traits (e.g., VanRaden, 2004). Accordingly, net merit can be estimated by an index of genetic values for all traits of interest where each trait is weighted by its relative economic importance (e.g., Lin, 1978). Three different economic weighting indices (“yield,” “quality,” and “balanced”) combining 11 traits (Table 2) and two selection index methods (“Smith-Hazel” and “base”) were used to compare rP and rGMS (Table 3). The Smith-Hazel index (Smith, 1936; Hazel, 1943) was used because it is considered an optimal index. It accounts for the genetic and phenotypic correlations between traits that would cause a simple phenotypic index to be an imperfect predictor of the actual breeding goal—additive genetic net merit (Lynch and Walsh, 2008). A base index that ignores these parameters and simply weights traits by their economic values (Panse, 1946; Brim et al., 1959; Williams, 1962) was also used. While theoretically inferior to the Smith-Hazel index, the base index can be favorable when large data sets are not available for accurate estimation of the phenotypic and genetic trait correlations (Williams, 1962; Harris, 1964). It was unknown which method was best for this population; therefore, accuracies were averaged across the six indices tested. This resulted in a major finding in this study: rGMS was 14% greater than rP for net merit, on average (Table 3). Only one index resulted in rP being greater than rGMS: the Smith-Hazel index for yield where rP = 0.33 and rGMS = 0.31. It was expected that GS would compare well with PS because rGMS was competitive with rP for low H² traits; however, this was easily the largest rGMS/rP observed. As improving net merit is the primary goal for breeders, the high relative accuracy of GS for net merit was a very interesting result, and more research on GS performance for net merit is needed.

Biparental Genomic Selection or Multifamily Genomic Selection
Empirical studies of GS in biparental populations have shown that biparental GS will likely be superior to conventional MAS and phenotypic selection in terms of gain per unit time and cost (Lorenzana and Bernardo, 2009;
Heffner et al., unpublished data, 2010). Biparental GS has two clear attributes: (i) relatively low genotyping costs because extensive LD should make genome-wide marker coverage achievable with a few hundred markers and (ii) marker effect estimates will be “population specific” and can therefore capture effects caused by epistasis and rare alleles. Biparental GS, however, requires phenotypes of lines from each cross before conducting GS, which may prolong the selection cycle and result in lower gains per year than multifamily GS. Also, biparental GS may not maximize r_{GS} because N_{TP} will likely be limited because a separate TP is created for each cross. This may explain why, for the same nine wheat quality traits, r_{GS}/H in this multifamily GS study (~0.7) was greater than r_{GS}/H in a biparental GS study (~0.5) by Heffner et al. (unpublished data, 2010). Such comparisons of accuracies between multifamily and biparental GS studies, however, should be made with caution because differences in genetic variances could make comparisons misleading.

The benefit of using multifamily GS to reduce cycle time was shown by a recent simulation study for wheat and maize by Heffner et al. (2010). Their results suggest that if r_{GS} for net merit approaches 0.5, as reported here, multifamily GS could increase gain per unit time and per unit cost by more than two to threefold in plant breeding. Of course, breeders will use GS to predict untested progeny rather than perform cross-validation on lines from the same population from which the TP was sampled. Accuracies for that former purpose will differ from what we obtained so that we cannot perfectly forecast the relative merits of multifamily and biparental GS, conventional MAS, and phenotypic selection. Nevertheless, two main features of this study should make these results relevant to actual plant breeding programs: (i) the population consisted of current advanced breeding lines of an advanced-cycle breeding program and (ii) predictions were made using training data from one year and validated using lines that were not in the TP and phenotypes from another year to avoid inflation of r_{GS} caused by common G×E deviations in the TP and SP. If the results observed here and by Heffner et al. (2010) hold true, GS will dramatically increase gains from selection in plant breeding programs.

CONCLUSIONS

Historically, plant breeders have relied on extensive replication of breeding lines across many locations and years to evaluate the genetic potential of breeding material. Undoubtedly, these intensive testing procedures will be necessary for evaluating advanced material before commercialization to ensure that farmers receive stable, high-performing varieties. However, the use of GS in plant breeding may result in major changes in the methods used to select parental lines for breeding crosses and significantly reduce the time and resources needed for the selection stages leading up to prerelease testing. As indicated by this study, GS has the ability to predict genetic value of individuals without time consuming, expensive phenotyping. Therefore, selection accuracy normally achieved by replicating each line may instead be accomplished by “replicating alleles.” Because lines are selected using predictions based on allelic effects, the unit of evaluation in GS is the allele rather than the line. Precise evaluation of multiple alleles is more effectively accomplished by evaluating many unreplicated lines rather than fewer replicated lines (Knapp and Bridges, 1990). Thus, GS schemes should leverage the phenotypic and genotypic information of as many lines as possible whether replicated or not. Furthermore, this change in selection strategy should focus the breeder’s strategy on evaluating allelic effects rather than line means to get the most out of advances in high-throughput genotyping, statistical models, and breeding methodology to increase the rate of genetic gains in plant breeding. The value of GS was strongly supported by this study, as the observed prediction accuracies suggested that GS will be superior to both conventional MAS and phenotypic selection in terms of gain per unit time and cost. Further research and software development is needed to enable widespread adoption and optimization of GS in plant breeding programs.

Supplemental Information Available

Supplemental material is available free of charge at http://www.crops.org/publications/tpg.

Acknowledgments

The authors express deep gratitude to the late Walter Federer for his advice on field designs and his many contributions to plant breeding during his career. The authors thank Edward Souza and the USDA-ARS Soft Wheat Quality Laboratory in Wooster, OH, for their careful evaluations of milling and baking quality. The authors also thank Aaron Lorenz, Jesse Poland, and Adam Famoso for their helpful suggestions and Gaël Pressoir for his valuable insights at the beginning of this project. Support for the work of Elliot Heffner was provided by USDA National Needs Graduate Fellowship Competitive Grant No. 2005-38420-15785 from the National Institute of Food and Agriculture. Support for the work of Jean-Luc Janinck was provided by USDA-NIFA grant No. 2009-85606-05701 and No. 2009-65300-05661. Additional funding for this research was provided by USDA–NIFA National Research Initiative CAP grant No. 2005-05130 and by Hatch 149-402.

References


