# Genomic Selection for Predicting *Fusarium* Head Blight Resistance in a Wheat Breeding Program

Marcio P. Arruda, Patrick J. Brown, Alexander E. Lipka, Allison M. Krill, Carrie Thurber, and Frederic L. Kolb\*

#### Abstract

Genomic selection (GS) is a breeding method that uses markertrait models to predict unobserved phenotypes. This study developed GS models for predicting traits associated with resistance to Fusarium head blight (FHB) in wheat (Triticum aestivum L.). We used genotyping-by-sequencing (GBS) to identify 5054 singlenucleotide polymorphisms (SNPs), which were then treated as predictor variables in GS analysis. We compared how the prediction accuracy of the genomic-estimated breeding values (GE-BVs) was affected by (i) five genotypic imputation methods (random forest imputation [RFI], expectation maximization imputation [EMI], k-nearest neighbor imputation [kNNI], singular value decomposition imputation [SVDI], and the mean imputation [MNI]); (ii) three statistical models (ridge-regression best linear unbiased predictor [RR-BLUP], least absolute shrinkage and operator selector [LASSO], and elastic net]; (iii) marker density (p = 500, 1500, 3000, and 4500 SNPs); (iv) training population (TP) size ( $n_{TP}$  = 96, 144, 192, and 218); (v) marker-based and pedigree-based relationship matrices; and (vi) control for relatedness in TPs and validation populations (VPs). No discernable differences in prediction accuracy were observed among imputation methods. The RR-BLUP outperformed other models in nearly all scenarios. Accuracies decreased substantially when marker number decreased to 3000 or 1500 SNPs, depending on the trait; when sample size of the training set was less than 192; when using pedigree-based instead of marker-based matrix; or when no control for relatedness was implemented. Overall, moderate to high prediction accuracies were observed in this study, suggesting that GS is a very promising breeding strategy for FHB resistance in wheat.

Published in The Plant Genome 8 doi: 10.3835/plantgenome2015.01.0003 © Crop Science Society of America 5585 Guilford Rd., Madison, WI 53711 USA An open-access publication

All rights reserved.

RIGINALLY PROPOSED by Meuwissen et al. (2001) for animal breeding, GS predicts breeding values of individuals based on genome-wide molecular markers. It can be considered as a form of marker-assisted selection in which all markers are used to calculate GEBVs. It is assumed in GS that the quantitative trait loci (QTL) underlying the trait of interest are in linkage disequilibrium with at least one marker and that all the genetic variance can be explained by markers (Goddard and Hayes, 2007). For this reason, GS is particularly promising for predicting quantitative, complex traits where many small effect loci contribute to phenotypic variation. In GS, marker effects are estimated from a TP, for which phenotypes and genotypes are available. Marker effects are then used to predict phenotypes in a set of individuals, called the breeding population, which will only be

M.P. Arruda, P.J. Brown, A.E. Lipka, A.M. Krill, and F.L. Kolb, Dep. of Crop Sciences, Univ. of Illinois, 1102 S. Goodwin Ave., Urbana, IL 61801; C. Thurber, School of Science and Mathematics, Abraham Baldwin Agricultural College, 2802 Moore Hwy., Tifton, GA 31793. Received 16 Jan. 2015. Accepted 23 June 2015. \*Corresponding author (f-kolb@illinois.edu).

Abbreviations: BLUP, best linear unbiased prediction; DArT, diversity array technology; DON, deoxynivalenol concentration; EMI, expectation maximization imputation; FDK, Fusarium-damaged kernels; FHB, Fusarium head blight; FHBdx, Fusarium head blight index; GBS, genotyping-by-sequencing; GEBV, genomic estimated breeding value; GS, genomic selection; INC, incidence; ISK, incidence, severity, and kernel quality index; kNNI, k-nearest neighbor imputation; LASSO, least absolute shrinkage and operator selector; MNI, mean imputation; OLS, ordinary least square; PCR, polymerase chain reaction; PEBV, phenotypically estimated breeding value; QTL, quantitative trait loci; REGWQ, Ryan-Einot-Gabriel-Welch Q; RFI, random forest imputation; RR-BLUP, ridge-regression best linear unbiased prediction; SEV, severity; SNP, single-nucleotide polymorphism; SSR, simple sequence repeat; SVDI, singular value decomposition imputation; TP, training population; USWBSI, US Wheat and Barley Scab Initiative; VP, validation population.

genotyped. Based on GEBVs, individuals can be selected before being tested in field experiments, potentially speeding up the breeding cycle (Jannink et al., 2010; Heffner et al., 2009). Several factors can impact the accuracy of GS models including marker density, QTL number, sample size, and genotypic imputation methods (Zhong et al., 2009; Heffner et al., 2011; Rutkoski et al., 2013).

Fusarium head blight is the single most important wheat disease in the US Midwest, causing yield losses, reduction of grain quality (Dexter et al., 1996), and grain contamination because of mycotoxins. Fusarium graminearum Schwabe (teleomorph: Gibberella zeae Schw. Petch) is the predominant pathogen that causes FHB of wheat and barley (Hordeum vulgare L.) in North America. Classic genetic studies showed that wheat resistance to FHB is quantitatively inherited (Van Ginkel et al., 1996; Liu et al., 2005). Moreover, previous work suggests that the sources of genetic variation for FHB resistance are predominantly additive (Bai et al., 2000; Snijders, 1990), indicating that accumulation of resistance genes may be possible. Selection for FHB resistance is performed mainly using phenotypic data, and because FHB resistance is strongly influenced by the environment, inaccurate measurements of true genetic resistance of an individual breeding line frequently occur. In addition, phenotyping for FHB resistance is a laborious task requiring mist irrigation systems in the field or greenhouse capabilities. Some measurements, like mycotoxin analysis, are only obtained after harvest and are both time consuming and expensive. In this context, selection based on GEBVs instead of phenotypes could improve the breeder's ability to select individuals with superior FHB resistance. Such an approach could also increase gain per unit cost and time (Heffner et al., 2010).

The availability of genotypic information has greatly improved over the past few decades; however, phenotyping capabilities have not kept pace. As a consequence, breeders are now facing the so-called "large *p*, small *n* problem" (i.e.,  $p \gg n$ ) when applying markers to predict phenotypes. In this situation the number of predictor variables is larger than the number of observations, resulting in an infinite number of marker-effect estimates (Gianola, 2013). To address this problem, several penalized regression approaches have been proposed for GS models. Ridge-regression best linear unbiased prediction is based on an infinitesimal model in which all markers are equally shrunken toward zero. This model sets markers to be random effects with a common variance (Meuwissen et al., 2001). The LASSO performs continuous shrinkage and variable selection simultaneously. In the p >> n setting, LASSO will select at most nvariables and set the effects of the remaining predictors equal to zero (Tibshirani, 1996). The elastic net is a combination of both RR-BLUP and LASSO, where the penalty is a weighted average of the penalties from these two approaches (Zou and Hastie, 2005).

In a previous publication, Rutkoski et al. (2012) compared GS models for FHB-related traits in wheat using 2402 diversity array technology (DArT) markers and 38 single-sequence repeat (SSR) markers. This study involved germplasm from 18 different breeding programs at the United States and Canada. Heffner et al. (2011) used 1158 DArT markers to compare genomic-selection models, marker-assisted selection, and phenotypic selection for agronomic traits using germplasm from the Cornell University wheat breeding program (Ithaca, NY). Both studies showed GS as a promising strategy for wheat breeding.

With the advent of high-throughput SNP genotyping methods, high-density genome-wide marker sets are becoming viable in several crop species, making GS a practical approach for predicting complex traits. A robust, high throughput genotyping method called GBS has been applied to wheat with promising results (Poland et al., 2012b). Briefly, this technique consists of reducing genome complexity using restriction enzymes that target gene-rich regions (Elshire et al., 2011). The targeted segments are then polymerase chain reaction (PCR) amplified, barcoded, and sequenced in a multiplexed reaction. The protocol can be modified to accommodate a combination of enzymes, as performed in wheat and barley by Poland et al. (2012a). Genotyping-by-sequencing typically generates many thousands of SNPs with minimum of ascertainment bias, a common problem of SNP chips.

Resistance to FHB is an important breeding objective in most programs in the Corn Belt, including the soft red winter wheat program of the University of Illinois at Urbana–Champaign. As such, our study aimed to develop GS models for FHB-related parameters, using GBS. We compared the accuracy of three penalized models: RR-BLUP, LASSO, and elastic net. We also investigated how the prediction accuracies were affected by imputation methods, number of SNP markers, size of the training population, marker-based vs. pedigree-based relationship matrices, and controlling for relatedness.

## **Materials and Methods**

## Plant Material and Phenotypic Data

The germplasm used in this study consisted of 273 breeding lines derived from 233 different crosses. A total of 185 lines came from the University of Illinois soft red winter wheat program, and the remaining 88 lines were from 17 programs across the midwestern and eastern United States. Of these 88 lines, 50 have been used as parents in our breeding program. Hence, this collection is representative of the genetic diversity currently being explored at the University of Illinois. Measurements associated with FHB resistance were collected from a field nursery during 2011, 2013, and 2014 field seasons. Because of drought conditions, no phenotypic data were obtained in 2012. Each experiment was set up as a randomized completeblock design with two replicates. The experimental unit consisted of a 1-m-long single row. Because not all of the lines were present in all years, the experiment was analyzed as an unbalanced design. Grain spawn inoculum was prepared with a mixture of 10 different isolates of *F*.

graminearum collected throughout Illinois over several years. Agar plugs containing the isolates were placed in carboxylmethyl-cellulose liquid media and allowed to grow and sporulate (Tuite, 1969). After 7 d, a conidial suspension was obtained and mixed with autoclaved maize (Zea mays L.) grain. Colonized grains were then spread in the field at a rate of approximately 287 kg ha<sup>-1</sup> and 2 to 4 wk before anthesis. The field was mist irrigated three times per day during 6 wk starting 2 wk before the first wheat line headed. The following parameters were used to evaluate FHB resistance: incidence (INC), severity (SEV), FHB index [FHBdx; (incidence  $\times$  severity)/100], Fusarium-damaged kernels (FDK), incidence, severity, and kernel quality index [ISK;  $(0.3 \times \text{incidence})$  $+0.3 \times \text{severity} + 0.4 \times \text{FDK}$ , and deoxynivalenol concentration (DON). Incidence was estimated as the percentage of infected heads, irrespective of the amount of disease on each head. Severity was visually estimated as the percentage of infected spikelets in a wheat head. Incidence and severity are used to quantify the resistance to penetration and spread of the disease, respectively, and they are also known as type I and type II resistance (Schroeder and Christensen, 1963). Fusarium-damaged kernels are a visual estimate of the percentage of F. graminearum-damaged, tombstone kernels using known standards, and it is often classified as type III resistance (Mesterházy, 1995; Mesterházy et al., 1999). The FHBdx and ISK are disease indices routinely used by breeders when performing selection. The SEV and INC were recorded 1 m after the heading date of each breeding line, whereas FDK and DON were recorded after harvest. Resistance to toxin accumulation has been classified as type IV resistance (Miller et al., 1985) and deoxynivalenol is by far the most important mycotoxin produced by F. graminearum in wheat. Its concentration was determined by gas chromatography at the Department of Plant Pathology at the University of Minnesota.

In this study, the phenotypic data came from an unbalanced design, in which not all breeding lines were present in all years. For this reason, best linear unbiased predictions (BLUPs) were calculated instead of best linear unbiased estimators. For each line and each trait, BLUPs were obtained using PROC MIXED SAS version 9.4 (SAS Institute, 2013), according to Eq. [1]:

$$Y_{ijkl} = \mu + year_{i} + block(year)_{ij} + line_{k} + heading_{ijkl} + (year \times line)_{ik} + \varepsilon_{ijkl}$$
[1]

in which  $Y_{ijkl}$  is the observed phenotype,  $\mu$  is the overall mean, year<sub>i</sub> is the random effect of the *i*th year, block(year)<sub>ij</sub> is the random effect of *j*th block nested within the *i*th year, line<sub>k</sub> is the random effect of the *k*th line, heading<sub>ijkl</sub> is a quantitative covariate trait treated as fixed and defined as the Julian date on which heading was observed for the *l*th replicate of the *k*th line in the *j*th block within the *i*th year, (year × line)<sub>ik</sub> is the random effect of the interaction between the *i*th year and *k*th line, and  $\varepsilon_{ijkl}$  is the random error term. Heading date was included as a covariate (Miedaner et al., 2006) because it is known to have an important effect on FHB resistance measurements, especially when the inoculation method mimics natural infection (grain spawn, as used in this study) (Buerstmayr et al., 2009). Furthermore, heading date was correlated with each of the parameters used to quantify FHB resistance.

## Genotypic Data

DNA was extracted from 5-d-old leaves using a cetyltrimethyl ammonium bromide-chloroform protocol. After quantification, the GBS libraries were constructed according to Poland et al. (2012b) with a few modifications. For genomic complexity reduction, we used three two-enzyme combinations: *PstI*-HF-*MspI*, *PstI*-HF-*HinP1I*, and *PstI*-HF-*BfaI*. The enzyme *PstI*-HF (CTG-CAG) is a rare cutter, whereas *MspI* (CCGG), *HinP1I* (GCGC), and *BfaI* (CTAG) are common cutters. A different set of barcodes was used for each enzyme combination. Sequence data were obtained from 96-plex Illumina HiSeq2000 runs (W.M. Keck Center for Comparative and Functional Genomics).

The SNPs were called using the UNEAK pipeline (Lu et al., 2013) using a population-based approach similar to the one described by Poland et al. (2012a). A mismatch of 1 bp in a 64 bp tag was allowed when aligning the tag sequences using Burrows–Wheeler Alignment (Li and Durbin, 2009). Putative SNPs were identified from tags that differed by 1 bp and then filtered with the Fisher's exact test (p < 0.001), which tests whether two alleles are independent in a population of inbred lines. The minor allele frequency cutoff was set to 5%, and SNPs with more than 20% missing data were eliminated from the analysis. Additionally, SNP pairs exhibiting squared Pearson correlation coefficients exceeding  $r^2 = 0.8$  were excluded. Subsequently, a total of 5054 SNPs were used for the downstream analyses.

## Imputation Methods

Genotypic missing values were imputed using five different methods: MNI, kNNI (Troyanskaya et al., 2001), EMI (Dempster et al., 1977), RFI (Stekhoven and Bühlmann, 2011), and SVDI. These methods are described in detail by Rutkoski et al. (2013). The imputations were performed using the R scripts provided by the same authors. These analyses were performed on a Dell Precision with CORE i7 vPro 3.00GHz, 8GB RAM, 700GB HD. The four imputation methods were compared in terms of GS accuracy for all traits using the RR-BLUP model. All comparison tests were performed in SAS PROC GLM using the Ryan–Einot–Gabriel–Welch Q (REGWQ) multipletesting correction at  $\alpha = 0.05$ . The REGWQ correction conservatively controls the familywise error.

## Assessing the Level of Structure of the Population

Population structure is an important factor in GS as it can affect the prediction of breeding values (Lipka et al., 2014; Isidro et al., 2015). Biased estimates of GS accuracy can be obtained when population structure is not taken into account (Riedelsheimer et al., 2013). For these reasons, the level of structure of the germplasm was assessed before running any GS analyses. Using the 5054 SNPs obtained from the GBS protocol, a principal component analysis was performed in JMP Genomics 7 (SAS Institute, 2014) to detect population structure in this collection of lines. Additionally, the same SNPs were used to assess the relatedness among the individuals by calculating a relationship (**K**) using the A.mat function of the rrBLUP package (Endelman, 2011) in the R programming language.

## Calculation of Genomic-Estimated Breeding Values

Genomic-estimated breeding values were calculated by the following general equation (Lorenz et al., 2011):

$$g(x_i) = \sum_{k=1}^{p} x_{ik} \beta_k$$
[2]

where g() is a function relating phenotypes to genotypes,  $x_{ik}$  is the score of the *k*th SNP (coded additively as -1, 0, or 1) for the *i*th individual, and  $\beta_k$  is the effect of the *k*th SNP. For each trait, we tested three models: RR-BLUP (Hoerl and Kennard, 1970; Meuwissen et al., 2001), LASSO (Tib-shirani, 1996), and elastic net (Zou and Hastie, 2005). Shrinkage models differ from ordinary least square (OLS) regression by adding a penalty to the cost function. If in OLS regression the cost function is represented by the sum of squared residuals (i.e.,  $\sum e_i^2$ ), in shrinkage models the general cost function is given by Eq. [3]:

$$\frac{1}{2}\sum_{i}e_{i}^{2}+\frac{\lambda}{2}\sum_{k}\left|\beta_{k}\right|^{q}$$
[3]

where  $\lambda$  is the regularization parameter, and the term q is equal to 2 for RR-BLUP, 1 for LASSO, and 0 < q < 1 for elastic net (Lorenz et al., 2011). While in OLS regression the parameter estimators are calculated with  $\underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|_{l_{ln}}^{2}$ , the parameter estimators for RR-

BLÜP, LASSO, and elastic net are given by Eq. [4–6], respectively:

$$\underset{\beta}{\operatorname{argmin}} \left( \frac{\left\| Y - X\beta \right\|_{2}^{2}}{2\sigma^{2}} + \lambda \left\| \beta \right\|_{2}^{2} \right), \qquad [4]$$

$$\underset{\beta}{\operatorname{argmin}} \left( \frac{\left\| Y - X\beta \right\|_{2}^{2}}{2\sigma^{2}} + \lambda \left\| \beta \right\|_{1} \right), \quad [5] \text{ and}$$

$$(1+\lambda_{2}) \underset{\beta}{\operatorname{argmin}} \left( \frac{\left\| Y - X\beta \right\|_{2}^{2}}{2\sigma^{2}} + \lambda_{2} \left\| \beta \right\|_{2}^{2} + \lambda_{1} \left\| \beta \right\|_{1} \right), \quad [6]$$

where the notation  $\left\|\cdot\right\|_2$  is used for the  $L_2$ -norm loss function (Euclidian norm)  $\|\beta\|_2 = \left(\sum_i \beta_i^2\right)^{1/2}$ ,  $\|\|\|_1$  is used for the  $L_1$ -norm (Manhattan norm)  $\|\beta\| = \sum |\beta_i|$ , and  $\lambda$  $\geq$  0 is the tunning (penalty or regularization) parameter, which regulates the strength of shrinkage of the estimates. The  $L_1$  part is responsible for automatic variable selection, whereas the  $L_2$  part does grouped selection, encouraging grouping of highly correlated variables. The expression argmin refers to the solution of coefficients  $\beta$  that minimizes the equation inside the brackets. The terms  $\lambda_1$  and  $\lambda_2$  refer to the tunning parameters associated with the  $L_1$ - and  $L_2$ -norm penalties (Heslot et al., 2012; Ogutu et al., 2012). The rr-BLUP package (Endelman, 2011) was used to conduct RR-BLUP, while LASSO and elastic net were conducted in the glmnet R package (Friedman et al., 2010).

### **Genomic Selection Prediction Accuracies**

For each GS model and trait, the prediction accuracy was calculated as follows:

$$\frac{r(\text{GEBV:PEBV})}{\sqrt{h^2}},$$
[7]

where *r* is the Pearson's correlation between the GEBVs and the phenotypically estimated breeding values (PEBVs) in the VP (Dekkers, 2007; Albrecht et al., 2011; Zhao et al., 2012), and  $h^2$  is the broad-sense heritability on a line-mean basis (Fehr, 1991). The variance components estimated from PROC MIXED in SAS (described in Plant Material and Phenotypic Data section) were used to calculate these heritabilities.

## Number of Single-Nucleotide Polymorphism Markers

The number of markers used in GS has a strong impact on the time required to run each analysis. Thus, we wanted to assess how a reduction on the number of markers would impact the GS accuracies. Sixty randomly sampled marker datasets, each consisting of either 500, 1500, 3000, or 4500 SNPs, were drawn from the original genotypic data of 5054 SNPs. For each combination of GS model (RR, LASSO, and elastic net), marker set, and trait, GS was performed using fivefold cross validation, resulting in five values of prediction accuracy for each of the 60 runs. For each trait, the mean prediction accuracy was compared among GS models for the marker sets containing the same number of SNPs and between numbers of SNPs within the same GS model. The exact same folds were used to compare the prediction accuracy of different GS models.

## **Training Population Size**

In a cross-validation scheme, marker effects estimated from the TP are used to calculate GEBVs for individuals in the VP. As the TP size increases, the precision of the marker effects increases, thus resulting in more accurate phenotypic predictions. Therefore, we assessed the impact of the TP size ( $n_{\rm TP}$  = 96, 144, 192, and 218) on prediction accuracy for each of the parameters used to quantify FHB resistance. Ideally, multiples of 96 would have been selected for  $n_{\rm TP}$  because 96-well plates are used for genotyping; however, only 273 lines were available for this study and 55 lines were set aside for the VP.

## Pedigree-Based versus Marker-Based Relationship Matrices

Since the pedigree of the breeding lines was available, we calculated a pedigree-based relationship (**A**) matrix and compared it with a marker-based kinship (**K**) matrix in terms of genomic prediction accuracy. The **A** matrix was calculated using the coancestry coefficient (f) (Bernardo, 2010), using pedigree records going back to grandparents, and the **K** matrix was obtained using the A.mat function in the R package rrBLULP (Endelman, 2011) with all 5054 SNPs. The kin.blup function of the same package was used to obtain the GEBVs with both matrices.

## **Cross-Validation Scheme**

All analyses were performed using a fivefold cross-validation scheme, in which the 273 breeding lines are randomly divided into five groups, four groups being used in the TP for marker effect estimation and one group used in the VP. Each analysis was repeated using 60 random samples. As shown by Ly et al. (2013), overestimated genomic prediction accuracies can be obtained if closely related lines are present in both TP and VP. To avoid such situation, genetically similar lines were grouped into clusters using the *k*-means clustering algorithm (Hartigan and Wong, 1979) on marker data, as implemented by Ly et al. (2013). A total of 55 (total number of breeding lines/5) clusters were obtain in JMP Pro 12 (SAS Institute, 2015). Lines belonging to the same cluster were present in either VP or TP, not in both simultaneously.

Throughout this study, all analyses were performed with control for relatedness in both sets, except in one situation: when we compared controlling for relatedness (clusters) with a situation in which genetically similar lines were allowed to be in both TP and VP (random). Standard error of mean prediction accuracy was calculated from the averaged accuracy values across the five folds.

## Genomic-estimated Breeding Values and Phenotypically Estimated Breeding Values

After determining the best imputation method, GS model, SNP number, TP size, **A** or **K** matrix, and controlling or not for relatedness in the TP and VP, we used fivefold cross validation to estimate the mean effect of each SNP. Specifically, GEBVs were obtained for all 273 lines by multiplying the genotypic matrix by a vector containing the mean effect of each marker and adding the grand mean. As a final step, the bias of the predictions was estimated as the slope (*b*) of the linear regression of PEBVs on GEBVs, as performed by Zhang et al. (2014).

## Results

## Genotypes, Structure, and Imputation Methods

More than 30,000 SNPs were called in this collection of breeding lines using the UNEAK pipeline. After applying a Fisher's exact test cutoff (p < 0.001), a cutoff of no more than 20% missing data and a minor allele frequency cutoff of >5%, the number of SNPs was reduced to 5054 with 12.57% overall missing data.

No clear genetic structure was detected in this population, as revealed by Fig. 1. This finding is not surprising, as 86% of the lines belong to the University of Illinois breeding program or have been used in its crossing block. This low level of population structure was also detected by the principal component analysis, with the first and second principal components explaining only 4.6 and 2.8% of the total variability, respectively.

Four imputation methods were used to generate different genotypic data sets, which were then compared in terms of GS prediction accuracy (as determined from fivefold cross validation). As shown in Fig. 2, the methods performed equally well for all traits. The EMI method was numerically superior for two traits (INC and ISK), kNNI was superior for three traits (DON, FHBdx, and SEV), and SVDI resulted in numerically higher accuracies for FDK. Although no statistically significant difference was detected among the imputation methods for most traits, the methods differed considerably in terms of computation time required to impute missing genotypic data, with RFI being by far the most time demanding (>8 h), followed by EMI (1 min, 10 s), SVDI (<1 min), kNNI (<1 min), and MNI (<0.5 min). In a study involving five populations and a super-computer set, Rutkoski et al. (2013) also found the RFI method to be substantially more computationally demanding than the other methods. We decided to use the EMI approach for subsequent analysis because it gave numerically higher accuracies for two traits, including ISK, which is heavily used for selections at the University of Illinois' wheat breeding program. In addition, it was not very computationally demanding, and it has been used in other wheat GS publications (Poland et al., 2012a; Lado et al., 2013; Rutkoski et al., 2013).

## Genomic Selection Models and Single-Nucleotide Polymorphism Number

Moderate to high values of cross-validated accuracies were observed for all traits, with the highest accuracies observed for FDK, ranging from 0.67 to 0.82, and the lowest for SEV, ranging from 0.35 to 0.48 (Table 1). As a general trend, accuracies increased with the number of SNPs, albeit at different rates depending on the GS model and trait (Fig. 3). For instance, a plateau was observed for ridge regression after 1500 (INC, FDK, ISK, and DON) or 3000 SNPs (FHBdx). The LASSO and the elastic net behaved very similarly to each other, showing a different pattern from the RR-BLUP. No plateau was observed when they were used for INC and FDK. For all three models, the trait with the least increase in accuracy was ISK.



Figure 1. Heat map of the marker-based kinship (**K**) matrix for 273 wheat breeding lines. Matrix obtained from 5054 SNPs. Rows and columns represent the breeding lines.

For the same number of SNPs, RR-BLUPs outperformed LASSO and elastic net for all traits, except INC (SNP number equal to 4500). According to Zou and Hastie (2005), elastic net is expected to outperform the LASSO when p >> n. In this study, that was indeed the case in most of the cases, but the differences among these models were negligible. Considering the overall performance of the GS models, RR-BLUP was selected as the model of choice for subsequent analyses.

#### Training Population Size

Increasing the size of the TP resulted in higher accuracies for all traits (Fig. 4). For most traits, the greatest increase occurred when the TP changed from 96 to 144 breeding lines when, on average, accuracies were boosted by 11.2% across all traits. The rate of gain decreases as the TP size increases, reaching a plateau between 192 and 218 for all traits except SEV. Among all traits, FDK showed the highest accuracy values, followed by ISK and DON. Even with only 96 individuals in the TP, FDK showed mean accuracy of 0.75. The lowest accuracies were obtained for SEV, ranging from 0.33 ( $n_{\rm TP}$  = 96) to 0.48 ( $n_{\rm TP}$  = 218). In a study with agronomic and quality traits in wheat, Heffner et al. (2011) reported a decrease of the mean accuracy across all traits by 30% when  $n_{\rm TP}$ decreased from 288 to 96, which is an average reduction of 0.156% per individual removed from the TP. In our case, the mean decreased by 16.82% by changing  $n_{\rm TP}$  from 218 to 96, resulting in an average reduction of 0.138% per individual.



Figure 2. Fivefold cross-validated genomic selection accuracies for six traits associated with *Fusarium* head blight (FHB) resistance and four imputation methods. Methods receiving the same letter do not differ according to the Ryan–Einot–Gabriel–Welch multiple comparison test at  $\alpha = 0.05$  level. EMI, expectation maximization imputation; kNNI, *k*-nearest neighbor imputation; MNI, mean imputation; RFI, random forest imputation; SVDI, singular value decomposition imputation; DON, deoxynivalenol concentration; FDK, *Fusarium*-damaged kernels; FHBdx, FHB index; INC, incidence; ISK, incidence, severity and kernel quality index; SEV, severity. Genomic-estimated breeding values estimated using ridge regression best unbiased linear prediction, 218 lines in the training population, and 55 lines in the validation population. Error bars represent  $\pm 1$  standard error of the mean.

### Pedigree-Based versus Marker-Based Relationship Matrices

Genomic prediction accuracies obtained with the **K** matrix were significantly higher than accuracies obtained with the **A** matrix for all traits measured (Fig. 5). The largest difference was observed for ISK, with values of accuracy changing from 0.42 (**A**) to 0.73 (**K**). Similar differences were found for FHBdx and FDK. At the same time, DON showed the smallest difference, increasing from 0.46 (**A**) to 0.63 (**K**). Across all traits, accuracy changed from 0.36 to 0.63, a boost of 72.5%. All differences were significant according to the REGWQ multiple testing correction at  $\alpha = 0.05$ .

# Controlling for Relatedness in Training and Validation Populations

With the *k*-means clustering method, genetically similar lines were grouped in the same clusters, and the folds for cross validation were created out of the clusters. By doing so, full-sibs and half-sibs were not allowed to be in both TP and VP simultaneously. This scheme (cluster) resulted in significant reductions in genomic prediction accuracy values when compared with the situation with no control for relatedness in both sets (random) (Fig. 6). The largest reductions were observed for DON and FHBdx (15%), and the smallest were found for FDK and SEV (5% and 9.5%, respectively). Averaged across all traits, genomic prediction accuracies showed a 10% reduction, from 0.72 (random) to 0.64 (cluster). The differences were significant according to the REGWQ multiple testing correction at  $\alpha = 0.05$ .

Table 1. Fivefold cross-validated prediction accuracies  $(\pm \text{ standard error of the mean})$  for *Fusarium* head blight-related traits according to genomic selection model and marker density (single-nucleotide polymorphisms [SNPs]).

_				
Trait <sup>†</sup>	SNPs	<b>R</b> R <sup>‡</sup>	NET§	LASSO <sup>®</sup>
SEV	4500	$0.481\pm0.006~\text{A}^{\text{\#}}\text{a}^{\text{+}\text{+}}$	$0.400\pm0.007$ Ba	$0.396\pm0.007~{ m Ba}$
	3000	$0.472\pm0.006$ Aa	$0.397\pm0.010$ Ba	$0.396\pm0.010$ Ba
	1500	$0.446\pm0.08~{ m Ab}$	$0.380\pm0.011~\mathrm{Bab}$	$0.380\pm0.009~\mathrm{Bab}$
	500	$0.390\pm0.010~\text{Ac}$	$0.354\pm0.013~\text{Ab}$	$0.353\pm0.013~\mathrm{Ab}$
INC	4500	$0.599\pm0.006~\mathrm{Ba}$	$0.625\pm0.009~\mathrm{Aa}$	$0.629\pm0.009~\mathrm{Aa}$
	3000	$0.582\pm0.008$ Aa	$0.590\pm0.010~\mathrm{Aab}$	$0.593\pm0.011$ Aab
	1500	$0.573\pm0.008~\mathrm{Aa}$	$0.567\pm0.013~\text{Ab}$	$0.569\pm0.013~\text{Ab}$
	500	$0.526\pm0.009~\text{Ab}$	$0.507\pm0.015~\mathrm{Ac}$	$0.507\pm0.015~\mathrm{Ac}$
FHBdx	4500	$0.516\pm0.007~{ m Aa}$	$0.446\pm0.009~\mathrm{Ba}$	$0.446\pm0.009~\mathrm{Ba}$
	3000	$0.519\pm0.006~\mathrm{Aab}$	$0.448\pm0.008~\text{Ba}$	$0.451\pm0.009~\mathrm{Ba}$
	1500	$0.493\pm0.007~\text{Ab}$	$0.428\pm0.010~\text{Bab}$	$0.424\pm0.010~\text{Bab}$
	500	$0.425\pm0.008~{ m Ac}$	$0.399\pm0.012~\text{Ab}$	$0.401\pm0.012~\text{Ab}$
FDK	4500	$0.819\pm0.007~\mathrm{Aa}$	0.749 $\pm$ 0.009 Ba	0.749 $\pm$ 0.009 Ba
	3000	$0.812\pm0.006~{ m Aa}$	$0.702\pm0.011~\text{Bb}$	$0.705\pm0.010~\text{Bb}$
	1500	$0.805\pm0.007~\mathrm{Aa}$	$0.710\pm0.009~\text{Bb}$	$0.708\pm0.009~\text{Bb}$
	500	$0.750\pm0.008~\text{Ab}$	$0.674\pm0.013~\text{Bb}$	$0.676\pm0.013~\text{Bb}$
ISK	4500	$0.734\pm0.005~{ m Aa}$	$0.632\pm0.007~\text{Ba}$	$0.634\pm0.007~\text{Ba}$
	3000	$0.725\pm0.005~{ m Aa}$	$0.630\pm0.008~\text{Ba}$	$0.630\pm0.007~\text{Ba}$
	1500	$0.720\pm0.007~\mathrm{Aab}$	$0.631\pm0.011$ Ba	$0.631\pm0.011$ Ba
	500	$0.702\pm0.007~\mathrm{Ab}$	$0.603\pm0.009~\mathrm{Ba}$	$0.605\pm0.008~\mathrm{Ba}$
DON	4500	$0.638\pm0.005~\mathrm{Aa}$	$0.532\pm0.008~\text{Ba}$	$0.528\pm0.007~\text{Ba}$
	3000	$0.627\pm0.007~{ m Aa}$	$0.519\pm0.009~\mathrm{Ba}$	$0.514\pm0.009~\text{Ba}$
	1500	$0.622\pm0.006~{ m Aa}$	$0.522\pm0.009~\text{Ba}$	$0.525\pm0.010~\text{Ba}$
	500	$0.573\pm0.008~\text{Ab}$	$0.487\pm0.011~\text{Bb}$	$0.482\pm0.011~\text{Bb}$

<sup>+</sup> SEV, severity; INC, incidence; FHBdx, *Fusarium* head blight index; FDK, *Fusarium*-damaged kernels; ISK, incidence, severity and kernel quality index; DON, deoxynivalenol concentration.

<sup>‡</sup> RR, Ridge regression best linear unbiased prediction.

§ NET, ELASTIC-NET.

<sup>¶</sup>LASSO, least absolute shrinkage and selection operator.

<sup>#</sup>Within rows, means followed by the same capital letter are not significantly different according to the Ryan–Einot–Gabriel–Welch multiple comparison test at  $\alpha = 0.05$  level.

 $^{\dagger\dagger}$  Within columns, means followed by the same lowercase letter are not significantly different according to the Ryan–Einot–Gabriel–Welch multiple comparison test at  $\alpha=0.05$  level. Genotypic missing data imputed by the expectation maximization imputation method. Training population size was equal to 218 and validation population size equal to 55.

## Genomic-Estimated Breeding Values and Phenotypically Estimated Breeding Values

For each trait, GEBVs were calculated using the same imputation method (EMI), the best GS model (RR-BLUP), the largest training population size ( $n_{\rm TP}$  = 218), the **K** matrix, and control for relatedness in TP and VP in a fivefold cross validation scheme. When the corresponding GEBVs were plotted against the PEBVs for each trait, the values were evenly distributed around the reference line with an intercept of zero and slope of one, with higher agreement for the observations toward the middle of both distributions. The plots for ISK and DON are presented in Fig. 7. The fitted linear regression of PEBVs on GEBVs obtained estimated slope of approximately b = 1 for all traits, suggesting little to no bias in our GEBVs



Figure 3. Fivefold cross validated genomic selection accuracies for six *Fusarium* head blight (FHB)-related traits as a function of genomic selection models and single-nucleotide polymorphism (SNP) numbers. (a) SEV (severity); (b) INC (incidence); (c) FHBdx (FHB index); (d) FDK (*Fusarium*-damaged kernels); (e) ISK (incidence, severity and kernel quality index); (f) DON (deoxynivalenol concentration). RR, ridge regression best linear unbiased prediction; NET, elastic net; LASSO, least absolute shrinkage and selection operator. Genomicestimated breeding values estimated using 218 lines in the training population, 55 lines in the validation population, and genotypic missing value imputed by the expectation maximization imputation method.



Figure 4. The effect of training population (TP) size on genomic selection accuracy for six *Fusarium* head blight (FHB)-related traits. FDK, *Fusarium*-damaged kernels; ISK, incidence, severity and kernel quality index; DON, deoxynivalenol concentration; INC, incidence; FHBdx, FHB index; SEV, severity. Genomic-estimated breeding values estimated by ridge regression best-unbiased linear predictor, 55 lines in the validation population, and genotypic missing data imputed by the expectation maximization imputation method.

(Table 2). We then ranked the 273 breeding lines according to their GEBV and PEBV for ISK. When comparing the top 5% of both ranks, we observed that nine lines



Figure 5. Mean fivefold cross-validated prediction accuracies calculated using the realized marker-based kinship (**K**) and the pedigree-based (**A**) matrices for six traits associated with *Fusarium* head blight (FHB) resistance. Genomic-estimated breeding values (GEBVs) estimated using 218 lines in the training population, 55 lines in the validation population, and genotypic missing data imputed by the expectation maximization imputation method. Error bars represent  $\pm 1$  standard error of the mean.

out of 14 overlapped in the two ranks. The number of overlapping lines increased to 20 (out of 27) and 39 (out of 54) lines when the top 10 and 20% was considered, respectively. For DON, 10, 22, and 40 lines, respectively, overlapped for the top 5, 10, and 20%.



Figure 6. Mean fivefold cross-validated prediction accuracies with (cluster) and without (random) control for relatedness of individuals in the training and validation populations. Genomicestimated breeding values estimated using ridge regression best linear unbiased prediction, 218 lines in the training population, 55 lines in the validation population, and genotypic missing data imputed by the expectation maximization imputation method. Error bars represent ±1 standard error of the mean.

## Discussion

#### Genotypes and Imputation Methods

A collection of 273 breeding lines in use at the University of Illinois' soft red winter wheat breeding program was used to develop GS models for traits associated with FHB resistance. Cultivars with higher levels of resistance to this disease are urgently needed across the United States. Five thousand and fifty-four SNPs were called by GBS and the UNEAK pipeline (Lu et al., 2013). Allelic effects of each marker were estimated and used to calculate GEBVs for the lines.

Five different imputation methods were compared in terms of GS accuracy. The EMI, SVDI, kNNI, RFI, and MNI methods had roughly equivalent performance for all traits. At the same level of missing data (20%), Rutkoski et al. (2013) found similar results. Moreover, they demonstrated that these imputation methods tend to show differences with respect to GS accuracy as the level of missing data increases. Poland et al. (2012a) found no difference in GS accuracy for wheat traits when imputation methods were compared, even working with up to 80% missing data per marker. In this study, EMI was numerically superior for two traits (FDK and INC) and required only 1 min and 10 s to run a complete set of imputations for one data set. At the same time, RFI required over 8 h and did not result in higher GS accuracy for any traits. Two other methods, kNNI and SVDI, required almost 1 min to run one set of imputations, and MNI required less than 30 s, but this method resulted in the lowest accuracies.

We preferred to work with a low level of missing data for two reasons. First, the GS analysis gets computationally intensive as the number of SNP increases. Second, a total of 5054 SNPs were called with 20% missing data per markers. Genomic selection analyses with a larger number of SNPs were performed, but the increments in accuracy



Figure 7. Distribution of phenotypically estimated breeding values (PEBVs) and genomic-estimated breeding values (GEBVs) obtained from fivefold cross validation for (a) incidence, severity and kernel quality index and (b) deoxynivalenol concentration (DON). Genomic-estimated breeding values estimated using ridge regression best linear unbiased prediction, 218 lines in the training population, 55 lines in the validation population, and genotypic missing data imputed by the expectation maximization imputation method.

were not substantial enough to justify the inclusion of larger marker sets in our analysis (data not shown).

#### Genomic Selection Models and Single-Nucleotide Polymorphism Number

In this study, RR-BLUP, LASSO, and elastic net were compared for six traits that included random subsets of the 5054 SNPs that varied in size. The RR-BLUP method is based on an infinitesimal model where all predictors are maintained in the analysis, resulting in a nonparsimonious model. In contrast, the LASSO performs variable selection and keeps a subset of predictors in the model, which is a reasonable assumption in plant breeding, as

Table 2. Estimate ( $\pm$ standard error of the mean) and
99% confidence interval of unbiasedness for six Fu-
sarium head blight-related traits.

Trait <sup>†</sup>	<b>b</b> (pebv, gebv) <sup>‡</sup>	99% confidence interval	
SEV	0.958 ± 0.0517	0.9067	1.0010
INC	$1.024 \pm 0.0759$	0.9483	1.1002
FHBdx	$0.994 \pm 0.0535$	0.9405	1.0475
FDK	$1.033\pm0.0468$	0.9862	1.0799
ISK	$1.034\pm0.0468$	0.9869	1.0805
DON	$0.984 \pm 0.0615$	0.9228	1.0459

<sup>+</sup> SEV, severity; INC, incidence; FHBdx, *Fusarium* head blight index; FDK, *Fusarium*-damaged kernels; ISK, incidence, severity and kernel quality index; DON, deoxynivalenol concentration.

<sup>+</sup> Unbiasedness estimated by the slope of the linear regression of phenotypically estimated breeding values (PEBVs) on genomic-estimated breeding values (GEBVs). The GEBVs were calculated using ridge regression best-unbiased linear predictor with 5054 SNPs, genotypic missing data imputed by the expectation maximization imputation method, with 218 lines in the training population and 55 lines in the validation population. Confidence interval based on 60 values of the slope.

some SNPs are expected to not be associated with the response variable. However, this method has its shortcomings; in particular, it can be unstable with high-dimension data. In addition, the LASSO tends to be problematic when the predictor variables are highly correlated, but that should not be the case in this study, as SNPs exhibiting pairwise correlations exceeding  $r^2 \ge 0.8$  were not used. The elastic net can be seen as a combination of both methods. For most traits, we obtained higher values of accuracy when using RR-BLUP than with the other methods. In only one situation (trait = INC and SNP number = 4500) did the LASSO and elastic net provide numerically higher accuracies, although they were not significantly different from RR-BLUP. This led to the conclusion that, for the traits used in this study, a model with a very large number of small-effect SNPs is more appropriate than a reduced model. It has been suggested that the LASSO is better suited for situations when large-effect genes are present. It is possible that such genes are present in our germplasms for INC and possibly for other traits. Even so, the superiority of RR-BLUP was demonstrated.

Higher prediction accuracies were observed as the number of SNP included in the models increased; however, diminishing gains were observed in most cases after including more than 1500 or 3000 SNPs, depending on the trait. The increase in accuracy as result of higher marker density has been reported in GS studies with maize (Lorenzana and Bernardo, 2009), wheat (Heffner et al., 2011), and oat (Avena sativa L.) (Asoro et al., 2011). In these studies, diminishing gains after a certain number of markers was the overall trend, although Asoro et al. (2011) did not observe a plateau for four out of five traits. A key feature of GS is that markers covering the entire genome would potentially explain all genotypic variability (Meuwissen et al., 2001; Goddard and Hayes, 2007). Thus, if marker coverage is sufficient, and markers are in linkage disequilibrium with QTL, high prediction accuracies are expected. Our data suggested that appropriate marker coverage was obtained with at least 3000 SNPs.

## **Training Population Size**

A significant increase in accuracy was observed when more breeding lines were included in the TP. The overall rate of gain averaged over the six traits we measured was comparable to a study in wheat conducted by Heffner et al. (2011), although they worked with different traits. In contrast to their results, a plateau in accuracy was observed when  $n_{\text{TP}} > 192$  for all traits except FHBdx. In future projects, we plan to include more breeding lines in the training set. The impact of the training population size on accuracy was assessed in other crops such as barley (Hordeum vulgare L.) (Lorenz et al., 2012), oat (Asoro et al., 2011), wheat (Heffner et al., 2011), sugar beet (Beta vulgaris L. subsp. vulgaris) (Würschum et al., 2013), and maize (Crossa et al., 2014). Lorenz et al. (2012) reported diminishing returns for prediction accuracies of DON and FHB resistance when  $n_{\text{TP}} > 200$  with no real gain when  $n_{\text{TP}}$  changed from 200 to 300 barley lines. Considering that expected accuracy depends on  $n_{\rm TP}$ ,  $h^2$ , and the number of loci involved (Daetwyler et al., 2008), we would need to increase the genetic variability of the germplasm and the number of testing environments to obtain higher accuracies.

## Pedigree-Based versus Marker-Based Relationship Matrices

Significantly higher genomic prediction values were obtained with the K matrix, as compared with the A matrix. Averaged across all traits, gains in predictive ability increased 72.5% by using marker information. In animal sciences, the advantage of markers over pedigree for prediction purposes has been demonstrated (Hayes et al., 2009; Vela-Avitúa et al., 2015). In wheat, Crossa et al. (2010) compared pedigree-based models with GS models using a diverse collection of 599 historical lines from the CIMMYT Global Wheat Breeding program. Compared with pedigree-based models, GS models for grain yield using 1279 DArT markers resulted in gains ranging from 7.7 to 35.7%, depending on the model. We speculate that the striking difference in prediction accuracy observed for markers over pedigree in our study is most likely due to imprecise pedigree information. It is possible that a given wheat line or cultivar was named differently by multiple breeders. It is also likely that following pedigrees back beyond the grandparental generation would have resulted in the discovery of additional relationships. Both of these factors could cause our A matrix to underestimate the coefficient of coancestry of a pair of lines.

In a breeding context, we believe GS is likely to become the strategy of choice for prediction purposes. Although obtaining marker data involves costs and pedigree information does not, pedigrees can sometimes be difficult to obtain, especially for older or proprietary germplasm. In addition, sequencing costs have been decreasing over time (Wetterstrand, 2015), allowing even small, public breeding programs to obtain high density marker information.

# Controlling for Relatedness in the Training and Validation Populations

The germplasm used in this study consisted mainly of breeding lines from or in use at the University of Illinois. As consequence, several closely related lines were present, especially half-sibs and some full-sibs. In cross-validation schemes, overestimated prediction accuracies can be obtained if related lines are present in both TP and VP sets. To avoid such situation, the *k*-means clustering method was used on marker data to group genetically similar lines, and the folds were built out of the clusters. For two traits, DON and FHBdx, the reduction in prediction accuracy was as high as 15%, in relation to the random situation. Averaged across all traits, a 10% reduction was observed. In a cassava (*Manihot esculenta* Crantz) study, Ly et al. (2013) observed a similar level of reduction averaged across 19 traits.

## Genomic-Estimated Breeding Values and Phenotypically Estimated Breeding Values

A high agreement between PEBVs and GEBVs was observed for traits associated with FHB resistance. Consequently, wheat lines with higher levels of resistance can be selected even before being planted. At the University of Illinois' wheat breeding program, great emphasis is placed on ISK and DON when making selections. For these two traits, we showed that the most resistant lines would have been selected using both GEBV and PEBV. These results are encouraging from a breeding perspective, as selection based on GEBVs could save time and resources associated with phenotyping and still being able to identify the most resistant lines.

## Conclusions

Most of the germplasm used in this study is primarily represented by breeding lines from the University of Illinois soft red wheat program. This study showed that moderate to high genomic prediction accuracies can be achieved for FHB resistance-related traits when implemented in a breeding program. The RR-BLUP method outperformed the other models for all traits. We also demonstrated that moderate to high prediction accuracies can be obtained even with a reduced set of SNPs and as few as 96 lines in the training set. The results show that GS can indeed be a promising strategy when breeding for FHB resistance.

Historically, breeders have relied on phenotypic selection and, to some extent, on marker-assisted selection when breeding for FHB resistance. The cooperative system of nurseries and research projects supported by the USDA–ARS under the US Wheat and Barley Scab Initiative (USWBSI) has greatly improved our knowledge about the disease and has been important for incorporating resistance into cultivars. Our results support GS as a feasible breeding strategy for FHB resistance, aiding the strategies already implemented by breeders and the USWBSI. We plan to apply GS to calculate GEBVs in the  $\rm F_{4:5}$  generation to reduce the number of lines tested in replicated field trials. Additionally, GEBVs can be used to select which lines will compose the crossing blocks. Heffner et al. (2010) provide a framework for incorporating GS in a winter wheat breeding program based on single-seed descent. Finally, doubled-haploid-based programs could benefit even more from GS, as the number of lines to be field evaluated could be drastically reduced as soon as the doubled-haploid lines are available.

#### Acknowledgments

The authors are grateful to Dr. Yanhong Dong from the University of Minnesota for deoxynivalenol determination and to the reviewers and associate editor for valuable suggestions. This material is based on work supported by the USDA, under Agreement No. 59-0206-9-080. This is a cooperative project with the US Wheat & Barley Scab Initiative. The first author gratefully acknowledges the graduate fellowship from Monsanto, Inc.

#### References

- Albrecht, T., V. Wimmer, H.J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, et al. 2011. Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123:339–350. doi:10.1007/s00122-011-1587-7
- Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, and J.-L. Jannink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Gen. 4:132–144. doi:10.3835/plantgenome2011.02.0007
- Bai, G.H., G. Shaner, and H. Ohm. 2000. Inheritance of resistance to Fusarium graminearum in wheat. Theor. Appl. Genet. 100:1–8. doi:10.1007/ PL00002902
- Bernardo, R. 2010. Breeding for quantitative traits in plants. 2nd ed. Stemma Press, Woodbury.
- Buerstmayr, H., T. Ban, and J.A. Anderson. 2009. QTL mapping and marker-assisted selection for *Fusarium* head blight resistance in wheat: A review. Plant Breed. 128:1–26. doi:10.1111/j.1439-0523.2008.01550.x
- Crossa, J., G. De Los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, et al. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi:10.1534/genetics.110.118521
- Crossa, J., P. Pérez, J. Hickey, J. Burgueño, L. Ornella, J. Cerón-Rojas, et al. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112:48–60. doi:10.1038/hdy.2013.16
- Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3: E3395. doi:10.1371/journal.pone.0003395.
- Dekkers, J.C.M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124:331–341. doi:10.1111/j.1439-0388.2007.00701.x
- Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., B 39:1–38.
- Dexter, J.E., R.M. Clear, and K.R. Preston. 1996. Fusarium head blight: Effect on the milling and baking of some Canadian wheat. Cereal Chem. 73:695–701.
- Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:E19379. doi:10.1371/journal. pone.0019379
- Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Gen. 4:250–255. doi:10.3835/plantgenome2011.08.0024
- Fehr, W.R. 1991. Principles of cultivar development. V1: Theory and technique. Iowa State Univ. Press, Ames.
- Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33:1–22.
- Gianola, D. 2013. Priors in whole-genome regression: The Bayesian alphabet returns. Genetics 194:573–596. doi:10.1534/genetics.113.151753
- Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124:323–330. doi:10.1111/j.1439-0388.2007.00702.x
- Hartigan, J.A., and M.A. Wong. 1979. Algorithm AS 136: A K-means clustering algorithm. Appl. Stat. 28:100–108. doi:10.2307/2346830

- Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91:47–60. doi:10.1017/S0016672308009981
- Heffner, E.L., J.L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen. 4:65–75. doi:10.3835/plantgenome.2010.12.0029
- Heffner, E.L., A.J. Lorenz, J.L. Jannink, and M.E. Sorrells. 2010. Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci. 50:1681–1690. doi:10.2135/cropsci2009.11.0662
- Heffner, E.L., M.E. Sorrels, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12. doi:10.2135/cropsci2008.08.0512
- Heslot, N., H.P. Yang, M.E. Sorrels, and J.L. Jannink. 2012. Genomic selection in plant breeding: A comparison of models. Crop Sci. 52:14–160. doi:10.2135/cropsci2011.06.0297
- Hoerl, A.E., and R.W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55–67. doi:10.1080/004 01706.1970.10488634
- Isidro, J., J.-L. Jannink, D. Akdemir, J.A. Poland, N. Heslot, and M.E. Sorrels. 2015. Training set optimization under population structure in genomic selection. Theor. Appl. Genet. 128:145–158. doi:10.1007/ s00122-014-2418-4
- Jannink, J.L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Brief. Funct. Genomics 9:166–177. doi:10.1093/bfgp/elq001
- Lado, B., I. Matus, A. Rodríguez, L. Inostroza, J.A. Poland, F. Belzile, et al. 2013. Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 3:2105-2114.
- Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324
- Lipka, A.E., F. Lu, J.H. Cherney, E.S. Buckler, M.D. Casler, and D.E. Costich. 2014. Accelerating the switchgrass (*Panicum virgatum* L.) breeding cycle using genomic selection approaches. PLoS ONE 9:E112227. doi:10.1371/journal.pone.0112227
- Liu, S., Z.A. Abate, and A.L. McKendry. 2005. Inheritance of Fusarium head blight resistance in the soft red winter wheat Ernie. Theor. Appl. Genet. 110:454–461. doi:10.1007/s00122-004-1852-0
- Lorenz, A., S. Chao, F.G. Asoro, E.H. Heffner, T. Hayashi, H. Iwata, et al. 2011. Genomic selection in plant breeding: Knowledge and prospects. Adv. Agron. 110:77–123. doi:10.1016/B978-0-12-385531-2.00002-5
- Lorenz, A.J., K.P. Smith, and J.L. Jannink. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci. 52:1609–1621. doi:10.2135/cropsci2011.09.0503
- Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151–161. doi:10.1007/s00122-009-1166-3
- Lu, F., A.E. Lipka, J. Glaubitz, R. Elshire, J.H. Cherney, M.D. Casler, et al. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. PloS Gen. 9:E1003215. doi:10.1371/journal.pgen.1003215
- Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, Jr., et al. 2013. Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Sci. 53:1312–1325. doi:10.2135/cropsci2012.11.0653
- Mesterházy, A. 1995. Types and components of resistance to *Fusarium* head blight of wheat. Plant Breed. 114:377–386. doi:10.1111/j.1439-0523.1995. tb00816.x
- Mesterházy, A., T. Bartók, C.G. Mirocha, and R. Komoróczy. 1999. Nature of wheat resistance to *Fusarium* head blight and the role of deoxynivalenol for breeding. Plant Breed. 118:97–110. doi:10.1046/j.1439-0523.1999.118002097.x
- Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.
- Miedaner, T., F. Wilde, B. Steiner, H. Buerstmayr, V. Korzun, and E. Ebmeyer. 2006. Stacking quantitative trait loci (QTL) for Fusarium head blight resistance from non-adapter sources in an European elite spring wheat background and assessing their effects on deoxynivalenol (DON) content and disease severity. Theor. Appl. Genet. 112:562–569. doi:10.1007/s00122-005-0163-4
- Miller, J.D., J.C. Young, and D.R. Sampson. 1985. Deoxynivalenol and Fusarium head blight resistance in spring cereals. J. Phytopathol. 113:359–367. doi:10.1111/j.1439-0434.1985.tb04837.x

- Ogutu, J.O., T. Schulz-Streeck, and H.P. Piepho. 2012. Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. BMC Proceedings 6:S10.
- Poland, J.A., P.J. Brown, M.E. Sorrells, and J.L. Jannink. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:E32253. doi:10.1371/journal.pone.0032253
- Poland, J.A., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, et al. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 5:103–113. doi:10.3835/plantgenome2012.06.0006
- Riedelsheimer, C., J.B. Endelman, M. Stange, M.E. Sorrells, J.L. Jannink, and A.E. Melchinger. 2013. Genomic predictability of interconnected bi-parental maize populations. Genetics 194:493–503. doi:10.1534/ genetics.113.150227
- Rutkoski, J.E., J. Benson, Y. Jia, G. Brown-Guedira, J.L. Jannink, and M.E. Sorrells. 2012. Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. Plant Gen. 5:51–61. doi:10.3835/plantgenome2012.02.0001
- Rutkoski, J.E., J.A. Poland, J.L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 3:427-39.
- SAS Institute. 2013. The SAS System for Windows. Version 9.3. SAS Inst. Inc., Cary, NC.
- SAS Institute. 2014. JPM Genomics System for Windows. Version 7. SAS Inst. Inc., Cary, NC.
- SAS Institute. 2015. JPM Pro System for Windows. Version 12. SAS Inst. Inc., Cary, NC.
- Schroeder, H.W., and J.J. Christensen. 1963. Factors affecting resistance of wheat to scab caused by *Gibberella zeae*. Phytopathology 53:831–838.
- Snijders, C.H.A. 1990. The inheritance of resistance to head blight caused by *Fusarium culmorum* in winter wheat. Euphytica 50:11–18. doi:10.1007/BF00023155
- Stekhoven, D.J., and P. Bühlmann. 2011. MissForest—Nonparametric missing value imputation for mixed-type data. Bioinformatics 28:112–118. doi:10.1093/bioinformatics/btr597
- Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso J. R. Stat. Soc. B 58:267–288.
- Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, et al. 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525. doi:10.1093/bioinformatics/17.6.520
- Tuite, J. 1969. Plant pathological methods: Fungi and bacteria. Burgess Publishing Company, Minneapolis, MN.
- Van Ginkel, M.V., W.V.D. Schaar, and Y. Zhuping. 1996. Inheritance of resistance to scab in two wheat cultivars from Brazil and China. Plant Dis. 80:863–867. doi:10.1094/PD-80-0863
- Vela-Avitúa, S., T.H.E. Meuwissen, T. Luan, and J. Ødegård. 2015. Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. Genet. Sel. Evol. 47:9. doi:10.1186/ s12711-014-0084-2
- Wetterstrand, K.A. 2015. DNA sequencing costs: Data from the NHGRI large-scale genome sequencing program. National Human Genome Research Institute, Bethesda, MD. http://www.genome.gov/sequencingcosts (accessed 06 Jun. 2015).
- Würschum, T., J.C. Reif, T. Kraft, G. Janssen, and Y. Zhao. 2013. Genomic selection in sugar beet breeding populations. BMC Genet. 14:85. doi:10.1186/1471-2156-14-85
- Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao, J. He, et al. 2014. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS ONE 9:e93017. doi:10.1371/journal.pone.0093017
- Zhao, Y., M. Gowda, F.H. Longin, T. Würschum, N. Ranc, and J.C. Reif. 2012. Impact of selective genotyping in the training population on accuracy and bias of genomic selection. Theor. Appl. Genet. 125:707– 713. doi:10.1007/s00122-012-1862-2
- Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J.L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182:355–364. doi:10.1534/genetics.108.098277
- Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc., B 67:301–320. doi:10.1111/j.1467-9868.2005.00503.x