Genomic Selection for Fusarium Head Blight in a soft red winter wheat (*Triticum aestivum* L.) breeding program

M.P. Arruda[1], A.M. Krill[1], P.J. Brown[1], C. Thurber[2], and F.L. Kolb[1*]

[1]University of Illinois at Urbana Champaign, Department of Crop Sciences, Urbana-IL. [2]Abraham Baldwin Agricultural College, School of Science and Mathematics, Tifton-GA. Corresponding Author: (217) 333-4256, f-kolb@uiuc.edu .

Introduction

Genomic Selection (GS) is breeding strategy aiming at selecting superior individuals based on their genomic estimated breeding values (GEBVs). The strategy is particularly promising for quantitative traits and requires dense, genome-wide marker data. Here we show how GS accuracies are affected by different parameters, for six traits associated with FHB resistance.

Objective

The aim of this work is to assess the effect of genotypic imputation methods, statistical models, marker density, and training population size on GS accuracy.

Materials and Methods

Germplasm: 273 lines from, or in use at the University of Illinois wheat breeding program.

Phenotyping: Field evaluation in a scab nursery in Urbana-IL in 2011, 2013, and 2014. BLUPs calculated for: severity (SEV), incidence (INC), FHB index (FHBdx=SEV x INC), kernels quality (FDK), ISK index (ISK=0.3xINC+0.3xSEV +0.4xFDK), and deoxynivalenol concentration (DON).

Genotyping: Genotyping-by-sequencing used with three two-enzyme combinations: *PstI-MspI*, *PstI-HinPI*, and *PstI-BfaI*. Sequence data obtained from Illumina HiSeq2000, and then analyzed with UNEAK (*maf* = 5%, missing data per marker ≤ 20%, and Fisher's exact test at 0.001 level)

Imputation methods: mean imputation (MNI), singular value decomposition (SDVI), random forest regression (RFI), and expectation maximization (EMI).

Genetic structure: Assessed through principal component analysis (PCAs) using all 5054 markers in JMP Genomics.

Statistical models: marker effects estimated with ridge regression best linear unbiased predictor – rrBLUP, least absolute shrinkage and operator selector – LASSO, and ELASTIC NET. The R packages "rr-BLUP" and "glmnet" were used.

Marker density: random samples (p = 500, 1500, 3000, and 4500 SNPs) obtained from a total 5054 SNPs.

Training population size (n_{TP}): random samples of n_{TP}=96, 144, 192, and 218 were obtained from the 273 lines.

Accuracy: Calculated as $r(GEBV:PEBV)/\sqrt{h^2}$, where r = Pearson's correlation between genomic and phenotypically breeding values (GEBVs and PEBVs); h^2 = broad-sense heritability.

Mean comparison: Each analysis was repeated 300 times and the mean accuracies were compared using the Ryan-Einot-Gabriel-Welch multiple comparison test at 0.05 level with SAS PROC GLM.

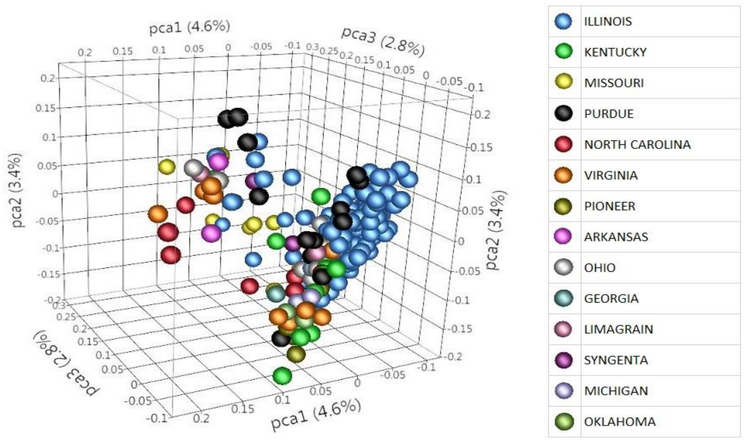

Figure 1. Principal component analysis of 273 breeding lines. Position of the lines in the coordinate system defined by first three principal components using all 5054 SNPs.

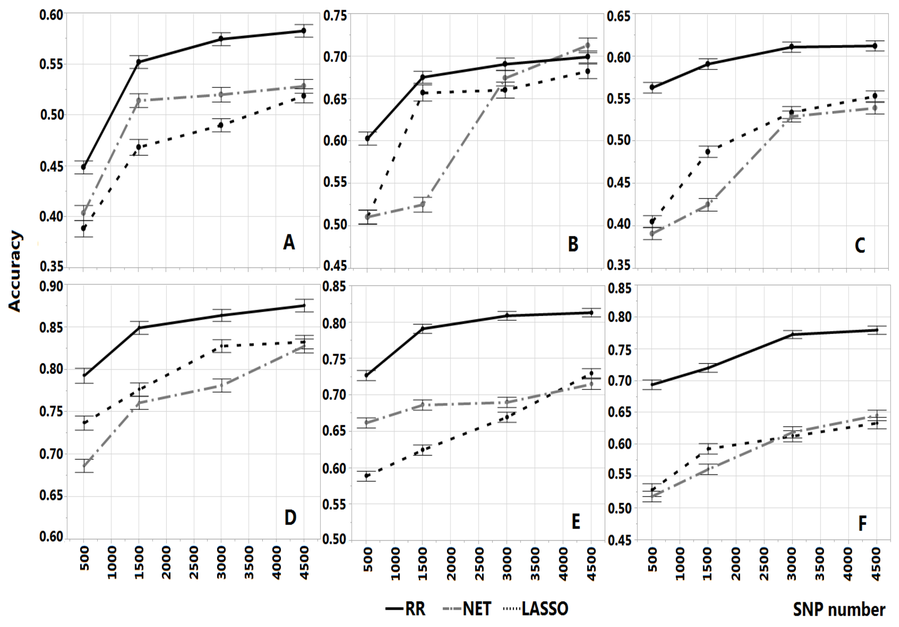

Figure 1. Five fold cross validated genomic selection accuracies for six FHB-related traits as a function of genomic selection models and SNP numbers. A = SEV (severity), B = INC (incidence), C = FHBdx (FHB index), D = FDK (Fusarium damaged kernels), E = ISK (ISK index), F = DON (deoxynivalenol concentration). RR = ridge regression best unbiased linear predictor, NET = ELASTIC-NET, LASSO = least absolute shrinkage and selection operator. Error bars represent ± one standard error of the mean.

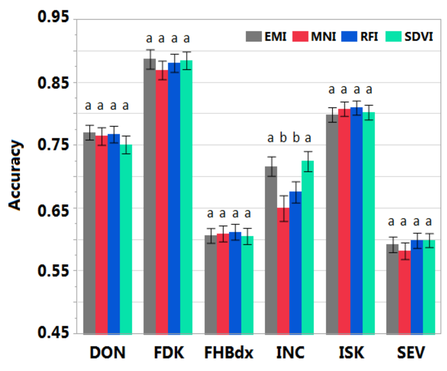

Figure 2. Five fold cross validated genomic selection accuracies for six traits associated with FHB resistance and four imputation methods. Methods receiving the same letter do not differ according to the Ryan-Einot-Gabriel-Welch multiple comparison test at 0.05 level. Error bars represent ± one standard error of the mean.

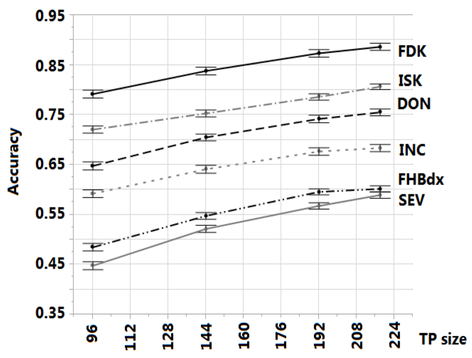

Figure 3. The effect of training population (TP) size on genomic selection accuracy for six FHB related traits. The analysis were performed using rr-BLUP and 5054 SNPs, imputed by the EMI method. Error bars represent ± one standard error of the mean.

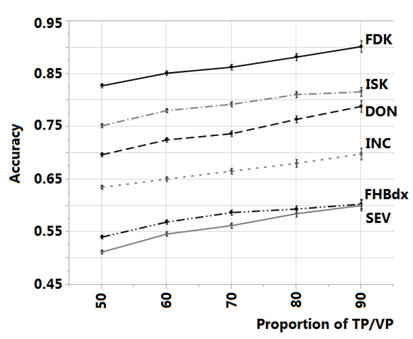

Figure 4. The effect of proportion of the training population/validation population on genomic selection accuracy for six FHB. The analysis were performed using rr-BLUP and 5054 SNPs, imputed by the EMI method. Error bars represent ± one standard error of the mean.

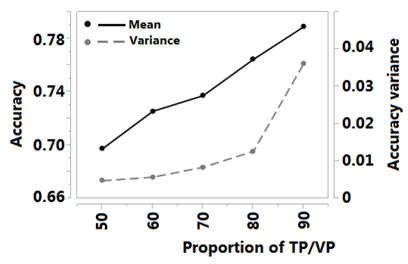

Figure 5. The effect of proportion of the training population/validation population on the mean and variance of genomic selection accuracies for DON (deoxynivalenol concentration). The analysis were performed using rr-BLUP and 5054 SNPs, imputed by the EMI method.

Results

✓ With the GBS protocol and the UNEAK pipeline, we were able to call 5054 high quality SNPs.

✓ Very low genetic structure exists in this collection, as revealed by the PC analysis. The first PC explained only 4.6% of the variability.

✓ Moderate to high accuracies were obtained for the traits evaluated in this study. The lowest accuracies were observed for SEV (ranging from 0.38 to 0.58), and the highest values were observed for FDK (ranging from 0.68 to 0.87).

✓ rr-BLUP outperformed the LASSO and ELASTIC-NET for all traits except INC (with SNP number ≥ 1500).

✓ Marker density had a significant effect on prediction accuracies, with diminishing increments after 1500 SNPs.

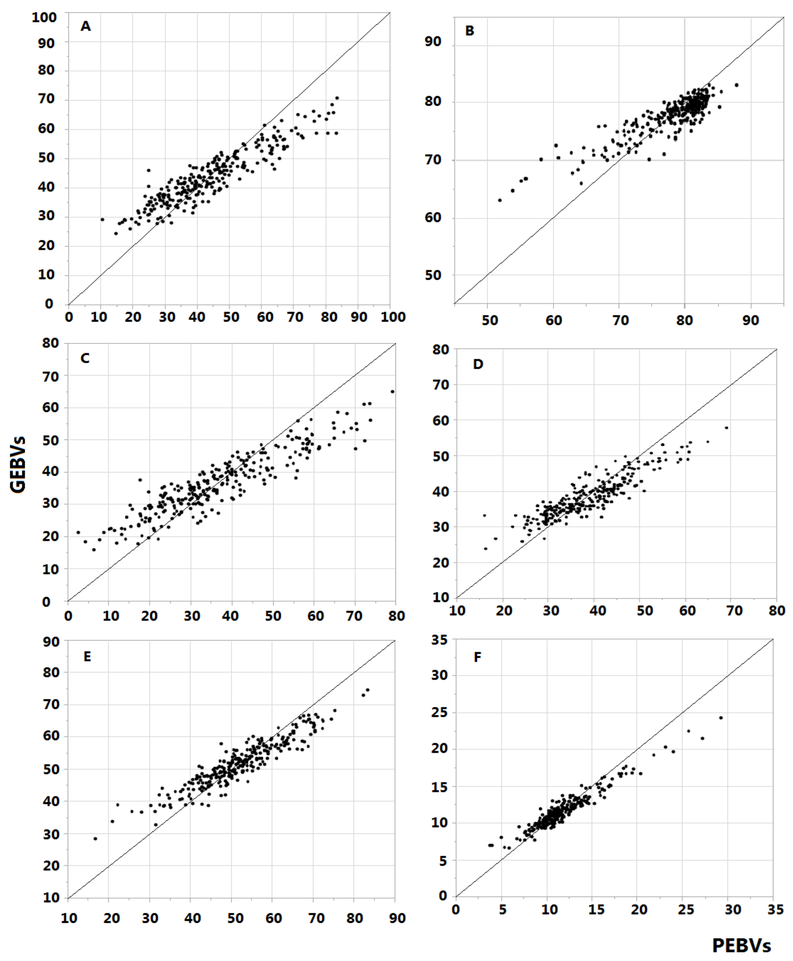

Figure 1. Distribution of phenotypically estimated breeding values (PEBVs) and genomic estimated breeding values (GEBVs) for severity (A), incidence (B), FHB index (C), Fusarium diseased kernel (D), ISK index (E), and deoxynivalenol concentration (F). The GEBVs were calculated with the mean marker effect of 5054 SNPs after 300 runs, calculated with rr-BLUP. The genotypic data was imputed with the EMI method, and the training population size was 218.

✓ The imputation methods performed equally well in terms of GS accuracy, with a numerical advantage for EMI in three out of six traits.

✓ A significant increase on accuracies was observed as the training population size increased. A plateau was observed for all traits when n_{TP}>192, except for ISK.

✓ The mean accuracy increased linearly with the ratio TP/VP, whereas the variance increases exponentially. The best combination of mean and variance was observed for TP/VP=0.8.

✓ We found a high agreement between PEBVs and GEBVs for all traits, especially for the mid values. Since the agreement is not perfect, some lines can be lost if the same threshold from PEBVs are used for GEBVs. Then, when performing selection, relaxing the GEBV threshold may be necessary in order to keep the promising lines. .

Conclusions

This study showed that moderate to high accuracies can be achieved for FHB resistance-related traits within the context of a breeding program. Ridge regression-BLUP outperformed the other models for all traits, even with its unrealistic assumption of all markers having the same variance. High accuracies can be obtained even with a reduced set of SNPs and a few hundred lines in the training set. These results are encouraging and show that genomic selection can indeed be a promising strategy when breeding for FHB resistance.

Acknowledgments

This material is based upon work supported by the U.S. Department of Agriculture, under Agreement No. 59-0206-9-057. This is a cooperative project with the U.S. Wheat & Barley Scab Initiative. The first author gratefully acknowledges Monsanto for financial support (Monsanto Fellowship).