Genome-wide selection in soft winter wheat: Effects of training population size, number of markers, and relatedness on genomic prediction accuracy

Dennis Nicuh Lozada[1], R. Esten Mason[1*], Dylan L. Larkin[1], & Jose Martin Sarinelli[2]
[1]Crop, Soil and Environmental Sciences Department, University of Arkansas, Fayetteville, AR 72701
[2]USDA-ARS Plant Science Research, Department of Crop Science, North Carolina State University, Raleigh, NC 27695
*Corresponding author: esten@uark.edu

Abstract

Genomic selection (GS) holds the promise of achieving higher genetic gains using molecular markers as predictors of breeding values of individuals. The effects of training population (TP) size, marker number, and relatedness on the accuracy of genomic predictions for grain yield (GY), heading date (HD), and plant height (PH) in a diverse panel of soft winter wheat (*N*= 239) were evaluated applying a standard single population cross-validation scheme under a ridge regression best linear unbiased prediction (rrBLUP) and different Bayesian models. Prediction accuracies, r_{GS} ranged from -0.08 to 0.70 for measured traits and BLUP phenotypic datasets for rrBLUP. Increasing TP size resulted to an increase in r_{GS}, where optimum predictions reached when 60% of the lines were used as TP. Using subsets of markers derived from association analyses also increased r_{GS} among measured traits compared to using whole marker dataset. Relative efficiency of GS per year (RE_y) for GY increased from 0.98-3.71 to 1.60-5.90 when subsets of marker data were used. Using lines belonging to same subpopulation, *Q* to predict performance on the same group also had effects on r_{GS} values, particularly for low heritable trait such as GY, indicating the importance of relatedness between the training and validation populations to achieve optimal predictions. Additionally, using locations with high phenotypic correlations to predict line GY performance also showed effects on r_{GS}. Taken together, our results demonstrated the importance of TP size, relatedness, and marker number in the context of improving GS accuracies in soft winter wheat.

Objectives

1. Determine the effects of training population (TP) size, marker number, and relatedness on the accuracy of genomic predictions for GY, HD, and PH in soft winter wheat using a single population cross-validation (CV) scheme under a ridge regression best linear unbiased prediction (rrBLUP) model; and

2. Compare different GS models using CV in terms of prediction accuracy.

Materials and Methods

Genotypic Data

- Illumina 9K SNP chip (5,661 SNPs)
- GBS markers (92,702 SNPs)

Phenotypic data

- Grain yield (GY), plant height (PH), and heading date (HD)
- Phenotypic datasets
  - BLUP across all environments (ABLUP)
  - BLUP for 2014 season (BLUP14)
  - BLUP for 2015 season (BLUP15)
  - BLUP for northern environments (NBLUP)
  - BLUP for southern environments (SBLUP)

Cross validation

- 10 x cross validation

Genomic selection models

- rrBLUP (Ridge regression best linear unbiased prediction)
- RKHS (Reproducing Kernel Hilbert Space)
- BRR (Bayesian Ridge Regression)
- Bayes A
- Bayes B

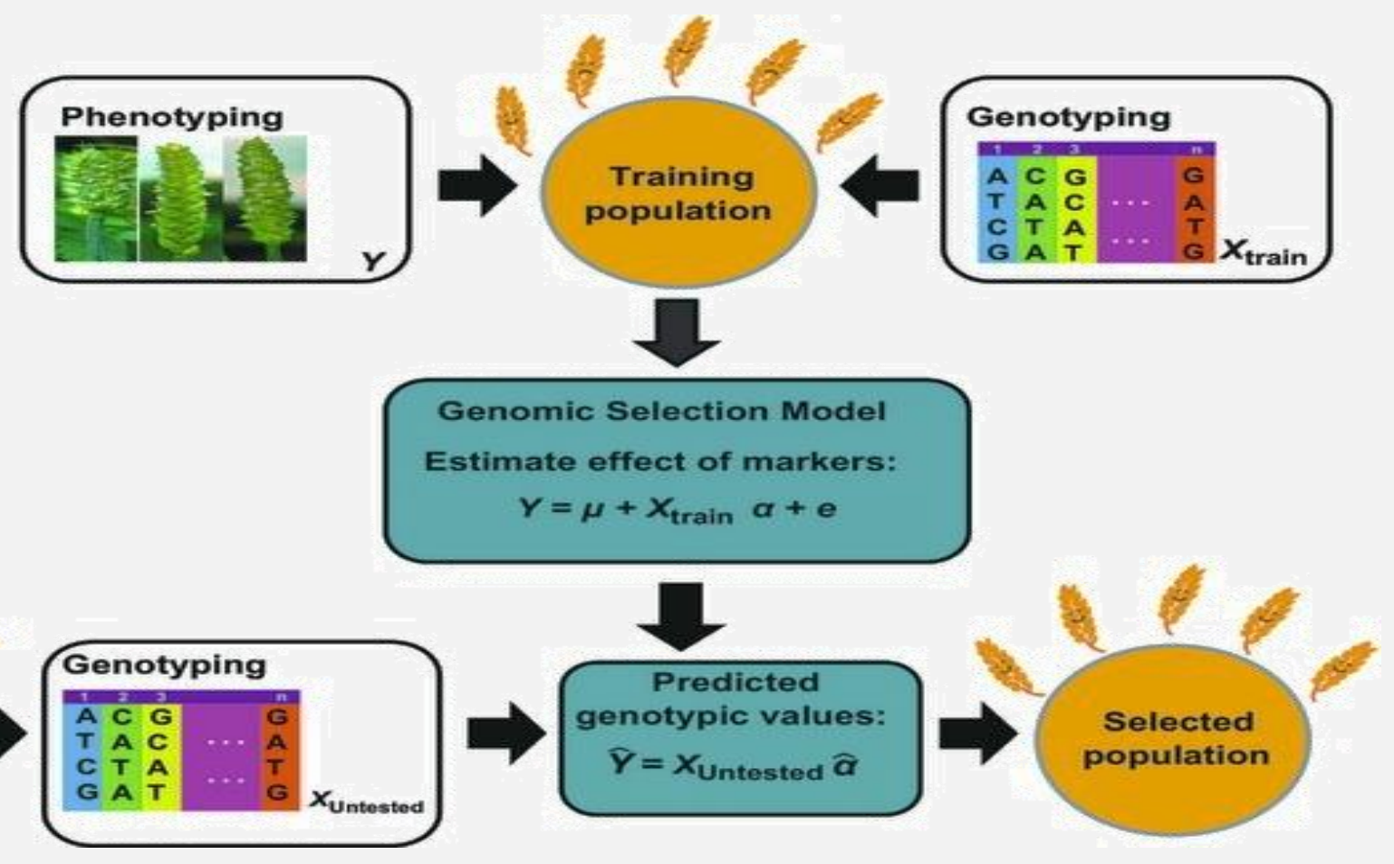

Fig. 1 Cross validation in genomic selection. Zhao et al (2015)

Results and Discussion

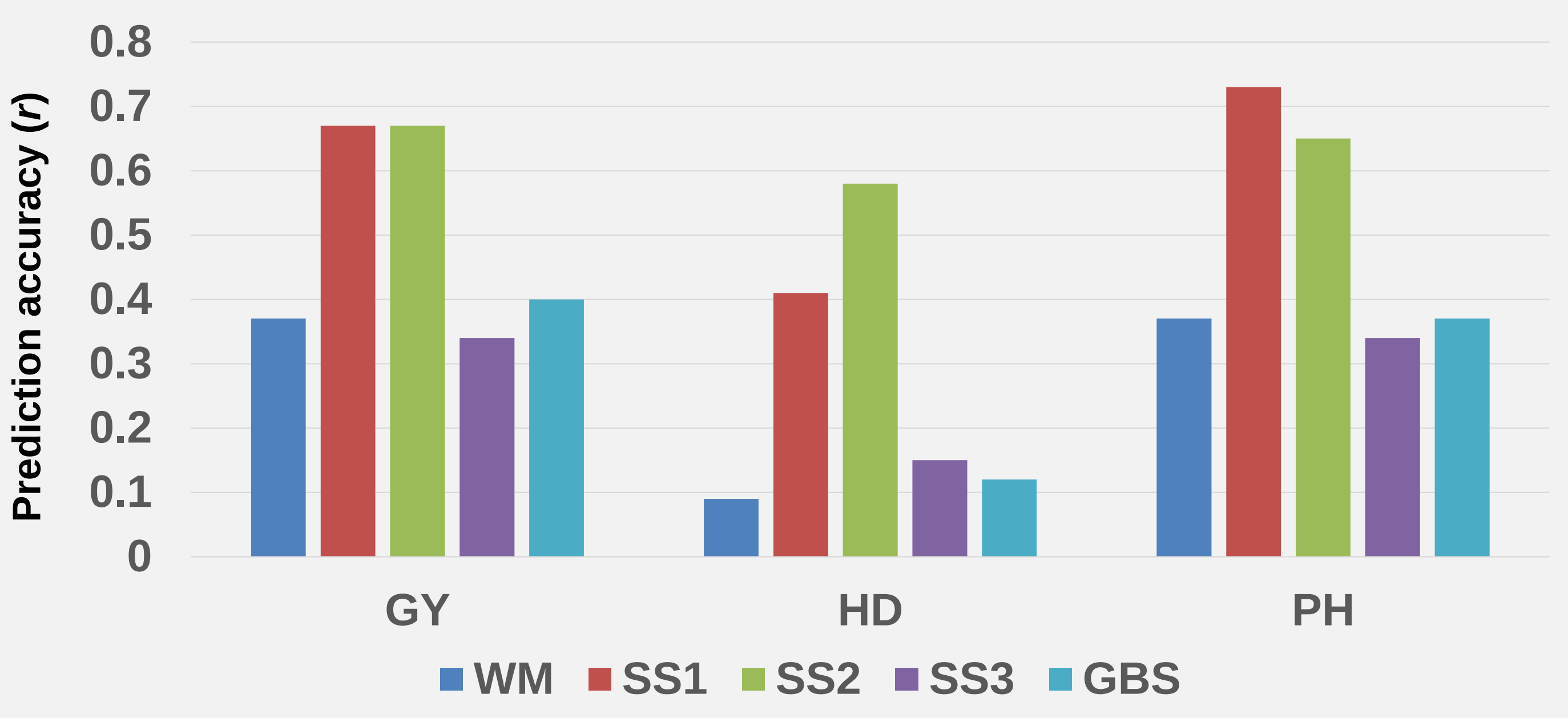

Fig. 2 Effects of using subsets of markers for genome-wide predictions using 10-fold CV in rrBLUP for GY, HD, and PH. Marker subset 1 (SS1) was based on $p < 0.10$; subset 2 (SS2) was based on $p < 0.05$; and subset 3 (SS3) was based on allele effects; all results were from association analyses. Number of markers: SS1-GY: 501 SNPs; SS1-PH: 576; SS1-HD: 613; SS2-GY: 210; SS2-PH: 206; SS2-HD: 297 ; SS3-GY: 2,599 ; SS3-PH: 2,678; SS3-HD: 2,770

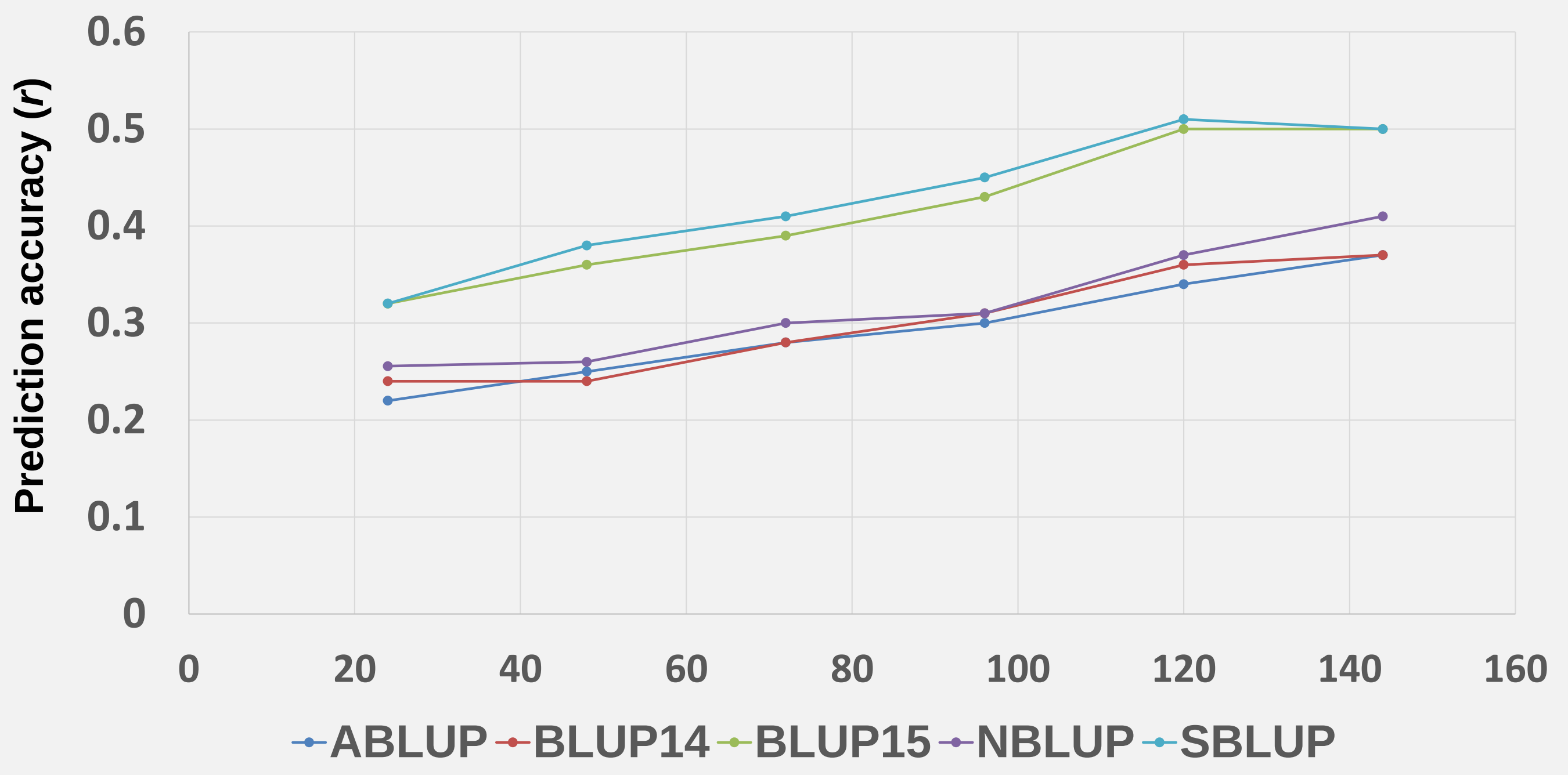

Fig. 3 Effects of training population (TP) size on GS accuracy under a 10-fold CV in rrBLUP for PH on different BLUP datasets. *ABLUP*- BLUP across all environments; *BLUP14*- BLUP values across 2014 site-years; *BLUP15*- BLUP across 2015 site-years; *NBLUP*- BLUP across northern environments; *SBLUP*- BLUP across southern environments

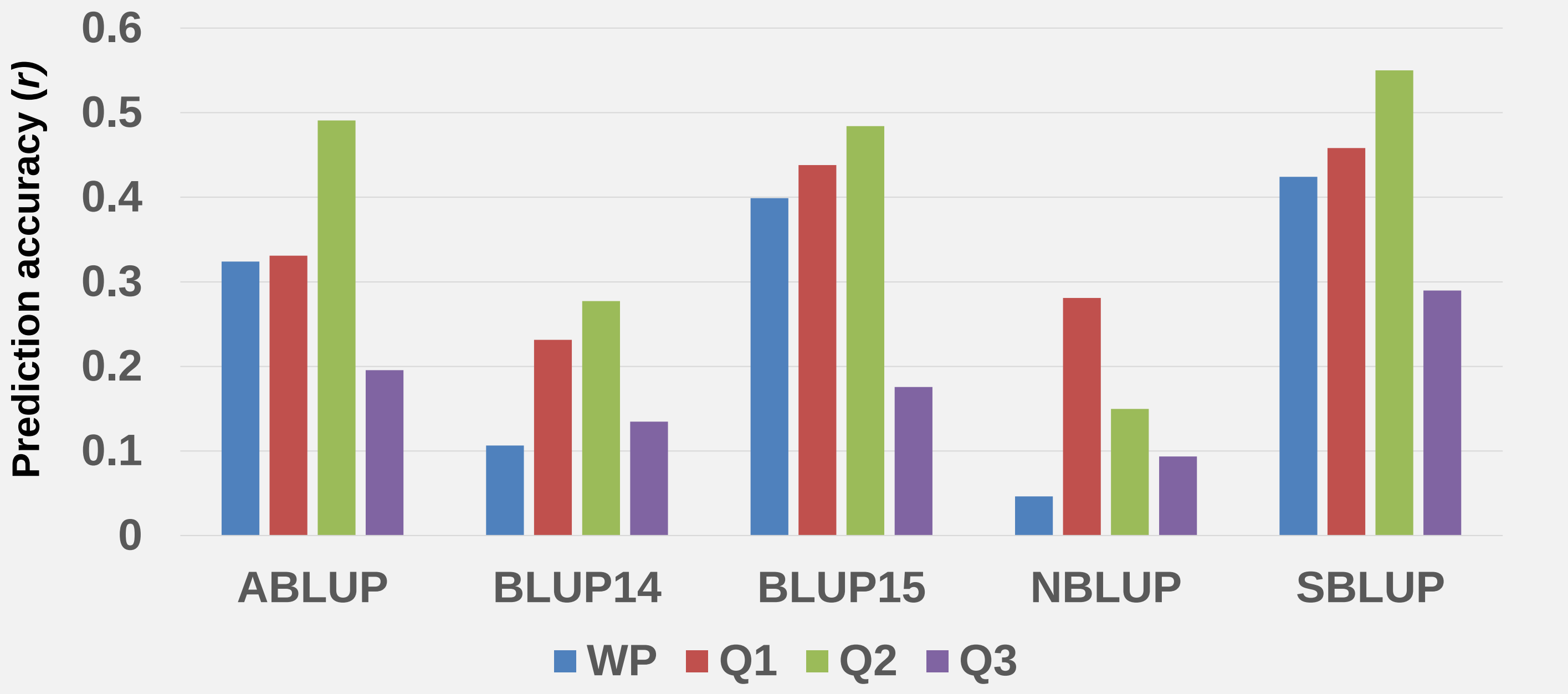

Fig. 4 Effects of relatedness of individuals on genomic prediction accuracy, 10-fold CV rrBLUP. Population structure analyses using STRUCTURE revealed three subpopulations *Q1, Q2,* and *Q3* based on genomewide marker data. WP- whole population

Table 1. Prediction accuracy (r_{GS}), relative efficiency per cycle (RE_C) compared to a cycle of phenotypic selection and relative efficiency per year (RE_Y) for grain yield under different GS models for ABLUP dataset

	Grain yield		
Model	r_{GS}	RE_C[a]	RE_Y[b]
rrBLUP	0.33	0.48	3.36
RKHS (Pedigree)	0.55	0.79	5.53
BRR	0.85	1.23	8.61
BayesA	0.78	1.13	7.91
BayesB	0.84	1.21	8.47

[a] RE_C calculated as $r/\sqrt{H}$ where *r* is the prediction accuracy and H is the heritability of the trait (H= 0.48)

[b] RE_Y obtained by multiplying RE_C with number of years to complete one cycle of phenotypic selection (e.g. for GY, estimated to be 7 years)

Results and Discussion

- Increasing TP size resulted to an increase in r_{GS}, where optimum predictions reached when 80% of the lines were used as TP
- Using subsets of markers derived from association analyses increased r_{GS} among measured traits compared to using whole marker dataset
- Relative efficiency of GS per year (RE_y) for GY increased from 0.98-3.71 to 1.60-5.90 when subsets of marker data were used
- Using lines belonging to same subpopulation, *Q* to predict GY on the same group also had effects on r_{GS} indicating the importance of relatedness between the training and validation populations to achieve optimal predictions
- Using locations with high phenotypic correlations to predict line GY performance also showed effects on r_{GS}
- Bayesian models generally showed higher prediction accuracy compared to other models

Summary

- Using subsets of more informative markers as opposed to whole genotype data with different significant levels from association analyses improved GS accuracy for GY and agronomic traits in soft winter wheat
- Relatedness and increasing TP size also showed effects in the accuracy of genomic predictions
- Results showed the importance of TP size, relatedness, and marker number improving GS accuracies in soft winter wheat

References

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4 (3):250-255

Hoffstetter A, Cabrera A, Huang M, Sneller C (2016) Optimizing Training Population Data and Validation of Genomic Selection for Economic Traits in Soft Winter Wheat. G3: Genes| Genomes| Genetics:g3. 116.032532

Lozada, D.N., Mason, E. Babar, M. A., Carver, B., Brown-Guedira, G., Merrill K., Arguello, N., Acuna, A., Vieira, L., Holder, A., Addison, C., Moon, D., Miller, R., Dreisigacker, S. (2017). Association mapping reveals loci associated with multiple traits that affect grain yield and adaptation in soft winter wheat. Euphytica 213:213-222. DOI: 10.1007/s10681-017-2005-2

Pérez P, de Los Campos G (2014) Genome-wide regression & prediction with the BGLR statistical package. Genetics: 114.164442

Zhao Y, Mette MF, Reif JC (2015) Genomic selection in hybrid breeding. Plant Breeding 134 (1):1-10

Acknowledgments

This research was supported by the Agriculture and Food Research Initiative competitive grant #2012-67013-19436 of the USDA National Institute of Food and Agriculture