GWAS Assisted Genomic Selection Does Not Consistently Improve Prediction Accuracy

Peter Schmuker

Abstract

The advent of affordable sequencing methods has resulted in genome wide association studies (GWAS) and genomic selection (GS) becoming increasingly common in plant breeding programs. Several studies report increases in GS prediction accuracies when GWAS results are incorporated as fixed effects to prediction models. However, many of these studies use genetic materials that are not representative of advanced breeding populations, rely on simulated phenotypic data, or improperly perform accuracy validation for incorporation of GWAS results. Here we demonstrate that incorporation of significant GWAS results to GS models do not reliably increase prediction accuracy across several agronomic traits. These results highlight the importance of tailoring GS methods to a breeding program's germplasm.

Introduction

Marker assisted selection (MAS) has proven to be a powerful tool to aid the introgression of large effect genes in spring wheat. Qualitative traits like Hessian Fly resistance or race specific rust Stem or Stripe Rust resistance can be introduced to a germplasm through tracking a single gene with a molecular marker (Prather et al. 2022). In these situations, breeding programs can use marker assisted backcrossing to quickly introduce resistance to advanced breeding lines without the need to phenotype for disease resistance. MAS can increase the speed of generation advancement, potentially improve selection accuracy, and save costs if markers are cheaper than phenotyping (Bernardo 2020).

Quantitive trats can also be rapidly improved through the usage of molecular markers (Bernardo 2014). The introgression and spread of Reduced Height genes in wheat have had a worldwide impact on food security (Ellis et al. 2002). During the Green Revolution, Rht genes were incorporated into wheat germplasm that have since been used to develop cultivars grown globally. Wheat breeding programs today may genotype potential parents from a different program’s germplasm to confirm what Rht genes are present in that population (Ellis et al. 2002). Even after the introduction and use of an Rht gene, a breeding program will still have variation present for plant height to select upon. Similarly, Adult Plant Resistance (APR) genes to Stripe Rust can significantly reduce yield loss through a generalized tolerance to the rust pathogen (Merrick et al. 2021). Introgression of a single APR gene can reduce, but not eliminate the yield penalty when a cultivar is inoculated with rust (Merrick et al. 2021). Ppd genes can alter phenology, yield potential, but most importantly maturity of a germplasm and are often tracked with molecular markers (Snape et al. 2001). Even with the introgression of a Ppd gene

and subsequent alteration of photoperiodism in wheat cultivars, breeders may still select for earlier or later maturing materials within a population fixed to be homozygous for a Ppd gene.

A large portion of the genetic variation present for traits like yield or protein content within a market class of wheat are explained by an unknown number of small effect genes (Bernardo 2020). Within advanced populations of spring wheat, breeders may struggle to find genes with large enough effects on a quantitative trait to warrant the creation of mapping populations, development of molecular markers, and regular use of these markers (Bernardo 2014). Instead, Genomic Selection models offer a way to apply selective pressure outside of the field season like MAS. Genomic Selection (GS) models vary in their assumptions about the genetic architecture of a trait. Mixed models with a kinship matrix, Genomic Best Linear Unbiased Prediction (GBLUP), or Ridge Regression on markers for BLUP (rrBLUP) fit a model that assumes many small effect genes with shared variance (Perez and Campos 2014). The prediction from the random effect term of GBLUP or summation of marker effects of genotyped individuals from rrBLUP, the breeding value, who do not have phenotypic records can be used to make selection decisions. Cultivars that have not yet been phenotyped can have performance predicted from a GS model using DNA samples. The correlation between predicted performance from a GS model and the true phenotype is a result of the heritability of the trait and relatedness of model testing and training material (Bernardo 2020).

Models built under the assumptions of many small effect genes could potentially lead to poor prediction accuracy for traits that have genetic architectures in violation of these assumptions. A trait with large effect genes segregating in the populations might have higher prediction accuracy with a model that incorporates large effect predictors to mimic the genetic architecture. Incorporation of GWAS results into GS models is hypothesized to improve prediction accuracy. Significant SNPs from GWAS analysis can be modeled as fixed effects in a GBLUP or rrBLUP (McGowan et al. 2021). SNPs that are modeled as fixed effects in GBLUP models can explain greater portions of genetic variance compared to SNPs that are only present as effects in the kinship matrix. However, performing GWAS with both training and testing genotypes contaminates the comparison made between included GWAS results or not in GS models (McGowan et al. 2021). Several studies show that incorporation of significant GWAS results as fixed effects can increase prediction accuracies in testing data (Odilbekov et al 2019, Yan et al. 2023).

Another possible reason for low prediction accuracy from GBLUP is the focus on modeling additive genetic variance. Additive by additive or epistatic genetic effects can possibly play an important role in explaining the variation of quantitative traits (Montesinos-López 2021). Machine learning methods or semi-linear regression methods can capture interactions between predictive variables (Montesinos-López 2022). Reproducing Kernel Hilbert space regression (RKHS) has been proposed as a predictive method that can capture non-additive genetic effects, in turn leading to greater prediction accuracies (Gianola and Van Kaam 2008). RKHS models can capture non-linear patterns in predictive data but could still suffer the same

issue of underestimating large genetic effects as GBLUP. Incorporation of significant GWAS results as fixed effects covariates in RKHS could improve predictive accuracy under specific genetic architectures.

Materials

Preliminary un-replicated yield trials were grown near Pullman, Washington over the 2021 and 2022 field seasons. Yield trials were planted over two locations within years, one location at Spillman Agronomy Farm and another on cooperator land nearby. Two market classes of wheat, soft and hard wheats, were grown and analyzed separately at each location. The soft market class of wheat has lower protein content and baking qualities appropriate for soba noodles, cookies, and sponge cakes. Hard wheat has greater protein content and is used for sandwich breads. All experimental entries were phenotyped for yield in bushels per acre, test weight, protein content, heading date in Julian days, and height. Each cultivar was averaged over the two locations for all phenotypes. Phenotpyic data was z score transformed using the scale function in R (R Core Team).

For the soft market classes, 627 and 522 entries were planted in 2021 and 2022 respectively. In the hard market class, 467 and 397 entries were planted in 2021 and 2022 respectively. Each year's entries in the preliminary yield trials are a distinct population of F5/F6 cultivars. Lines that show promise are selected and moved to the advanced yield trials in the subsequent years. There were no overlapping experimental entries between 2021 and 2022 in the preliminary yield trials. All entries were genotyped using Genotyping-By-Sequencing and variant calling was performed using the TASSEL GBS Version 2 Pipeline. SNPs were filtered on missingness greater than 20%, quality score of 20, heterozygosity greater than 50%, and a mean read depth of 1. NA calls were imputed using the LDKNII function in TASSEL. Genotypic data was converted from VCF format to a Numeric format using GAPIT (Wang and Zhang 2021).

Experimental Methods

Two sets of analysis were performed and presented in this paper to test the potential advantages of incorporating GWAS results into both GBLUP or RKHS regression.

The first set of analysis focuses on predictions made within individual trials. In these scenarios, a fifth of genotypes are randomly chosen as testing observations and their phenotype is excluded from model training. The remaining training genotypes have their phenotypic information used to perform GWAS. The SNPs with the lowest P values from this GWAS are noted and later used for fixed effects. GS models are also only trained with the lines randomly selected to be training genotypes. The testing genotypes are used to calculate Pearson

Correlation between predicted and observed phenotypic. The same training and testing data split was used for every model. For example, the same random fifth of lines designated as the testing materials would be treated as testing genotypes for all models. Once all models had been trained on the same training population and accuracy recorded on testing materials, a new random sampling of all materials for testing and training splits would occur. This process of randomly selecting genotypes as testing and leaving the remaining as training was repeated 50 times for every trait within each market class and year combination for four total scenarios, two market classes with two years of data, with five traits per scenario. It is important to restate that GWAS was only performed using phenotypic information from genotypes that were randomly designated for the training population.

The second set of analysis focuses on prediction accuracy within market classes but across years. In this scenario the trial in one year is considered the training population and the trial in the other year is treated as the testing population. The training population is used for both GWAS and training of GS models. This situation is more challenging as the 2021 and 2022 growing seasons were seriously different. The 2021 growing season was the worst drought in the past 100 years of Washington history (Geranios 2021). The same accuracy metric, Pearson correlation coefficient within the testing material, was used for to measure model accuracy.

Statistical Modeling

Within each of the two experimental scenarios and for all traits, thirty-one total prediction models were tested. Initially, a GWAS method was implemented using only phenotypic information from genotypes assigned to the testing pool. The GWAS method implemented in this study was Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) (Huang et al. 2017). BLINK is a multi-locus GWAS method that controls for SNPs in linkage disequilibrium and uses Bayesian Information Criteria to control pseudo QTN during model construction. BLINK was executed using a minor allele frequency of .1 and three Principal Components (PC) to control population structure. BLINK was implemented in GAPIT. SNPs were selected based on P value from BLINK results. The SNP was the lowest P value was recorded in a list, the three SNPs with the three lowest P values in another list (contains the SNP with the single lowest P value), and SNPs with five lowest P values in a final list (contains the previous three SNPs in this list).

Three fixed effect predictive models were implemented using the results from BLINK analysis. The first fixed effect model used the SNP with the lowest P value from BLINK output along with three principal components as linear predictors. Markers are coded as numeric 0, 1 ,2 values representing minor allele homozygous, heterozygous, and major allele homozygous respectively. Still only using training data the fixed effect model had regression coefficients assigned for the lowest P value SNP with three principal components in a single multiple regression model. The second and third fixed effect model was created this time using the three and five SNPs with lowest P values along with three PCs respectively. Fixed effect models were

created using the GLM method in GAPIT along with SNP Testing set to false. These fixed effect models represent a form of marker assisted selection for the improvement of quantitative traits.

Four GBLUP models were implemented as well. A standard GBLUP model with three PCs as fixed effect and the Zhang matrix was implemented for the random term. Three additional GBLUP models were implemented with different fixed effects covariates. These additional GBLUP models included list of SNPs created from BLINK along with three PCs. The one, there, and five lowest P value SNPs were included as fixed effects.

Four compressed BLUP (CBLUP) models were tested. Compressed BLUP uses a reduced kinship with groups rather than individuals to achieve greater prediction accuracy over standard GBLUP (Wang et al. 2018). CBLUP should offer greater prediction accuracy for traits with low heritability and a high number of underlying causal genes (Wang et al. 2018). A CBLUP model was implemented using only three PCs as fixed effects and then three additional CBLUP models with the same SNPs as fixed effect covariates. CBLUP with executed using GAPIT with model parameters left to the default settings.

RKHS regression can be performed using a matrix of numerically coded marker variables (n by p) or a relationship matrix like the Zhang matrix for GBLUP (n by n). Different Kernels can be used to transform the marker matrix into a relationship matrix with different properties than matrices meant to capture only additive effects. Several kernels were tested including the Gaussian Kernel (GK), Sigmoid Kernel (SK), and Polynomial Kernel (SK). The Gaussian Kernel offers great flexibility in its construction as a hyperparameter, the bandwidth parameter h, can be tuned to alter cell values in the relationship matrix. Normally, researchers perform cross-validation when using the GK to select a proper h value. Instead, three bandwidth values of .01, 1, and 2 were tested in all scenarios. Additionally, the Polynomial Kernel can be tuned with the degree number hyperparameter. For simplicity only a second-degree PK was calculated. For every Kernel discussed, a model was fit that had no fixed effect terms. Gaussian Kernels were created following De Los Campos 2010 with three different bandwidth hyperparameters. Sigmoid Kernel and Polynomial Kernels were created following code given in Montesinos-López 2021. Three additional models were fit with each Kernel using the same SNP lists with three PCs as the GBLUP and CBLUP methods. RKHS models with Kernels were fit using BGLR and a 5,000 sample burn-in period for 10,000 total iterations.

Predicted values were composed of both the random and fixed effects of every Genomic Selcetion model. For example, the GBLUP models with SNPs included as fixed effects combined the Best Linear Unbiased Estimator and Best Linear Unbiased Predictor that come from the fixed and random effect terms respectively for the final prediction. This was done so the effect of all model terms was used to evaluate prediction accuracies. This was also true of the RKHS methods implemented with BGLR, fixed effect estimates, and random effect predictors from relatedness were combined for one predicted value from the Yhat object in BGLR. The three fixed effect only models that mimicked a form of MAS only make prediction from fixed effects as the error term is not related to prediction.

Fixed Effect Models	No Fixed Effect	3 PCs	1 SNP	3 SNP	5 SNP	RKHS GK h1 Models	No Fixed Effect	3 PCs	1 SNP	3 SNP	5 SNP
QTN 1		x	x			GK 1 Bandwidth	x				
QTN 3		x		x		GK 1 Bandwidth QTN1		x	x		
QTN 5		x			x	GK 1 Bandwidth QTN3		x		x	
						GK 1 Bandwidth QTN5		x			x
GBLUP Models						RKHS GK h2 Models					
GBLUP		x				GK 2 Bandwidth	x				
GBLUP QTN1		x	x			GK 2 Bandwidth QTN1		x	x		
GBLUP QTN3		x		x		GK 2 Bandwidth QTN3		x		x	
GBLUP QTN5		x			x	GK 2 Bandwidth QTN5		x			x
CBLUP Models						RKHS 2nd Degree PK Models					
CBLUP		x				Polynomial Kernel	x				
CBLUP QTN1		x	x			Polynomial Kernel QTN1		x	x		
CBLUP QTN3		x		x		Polynomial Kernel QTN3		x		x	
CBLUP QTN5		x			x	Polynomial Kernel QTN5		x			x
RKHS GK h.01 Models						RKHS SK Models					
GK .01 Bandwidth	x					Sigmoid	x				
GK .01 Bandwidth QTN1		x	x			Sigmoid QTN1		x	x		
GK .01 Bandwidth QTN3		x		x		Sigmoid QTN3		x		x	
GK .01 Bandwidth QTN5		x			x	Sigmoid QTN5		x			x

Results for Scenario One

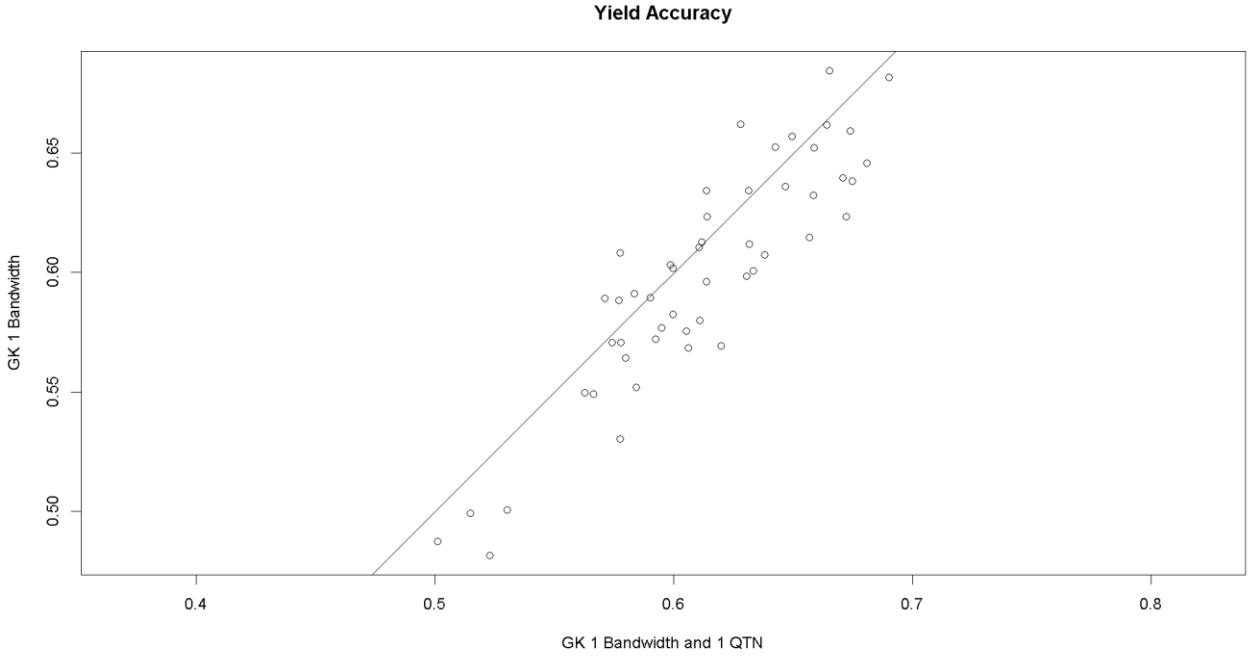
Soft 2021

Over fifty replications of randomly splitting the population into training and testing materials, RKHS regression with a Gaussian Kernel performed the best with one of the three tested bandwidth parameters. Including significant GWAS results as fixed effects along with PCs further improved the average Pearson Correlation for some traits but not all. Averaging prediction Pearson Correlation across all traits, the RKHS regression models with a 2 bandwidth hyperparameter performed the best with an r value of 0.5646, including GWAS results reduced accuracy on average with including one QTN with 3 PC reducing average accuracy to 0.5569 from 0.5646. On average, GBLUP models also had lower average accuracy when GWAS results were incorporated with a decrease from an average 0.5224 correlation to 0.5048 as QTN inclusion increased. Traits that benefitted from inclusion of significant GWAS results were Yield and Heading date. Test Weight, Protein content, and Plant Height did not improve in both GBLUP or RKHS with different Kernels when GWAS results were incorporated as additional fixed effects. The fixed effect only models performed the worst and had the lowest average correlation and highest standard deviation in correlation values.

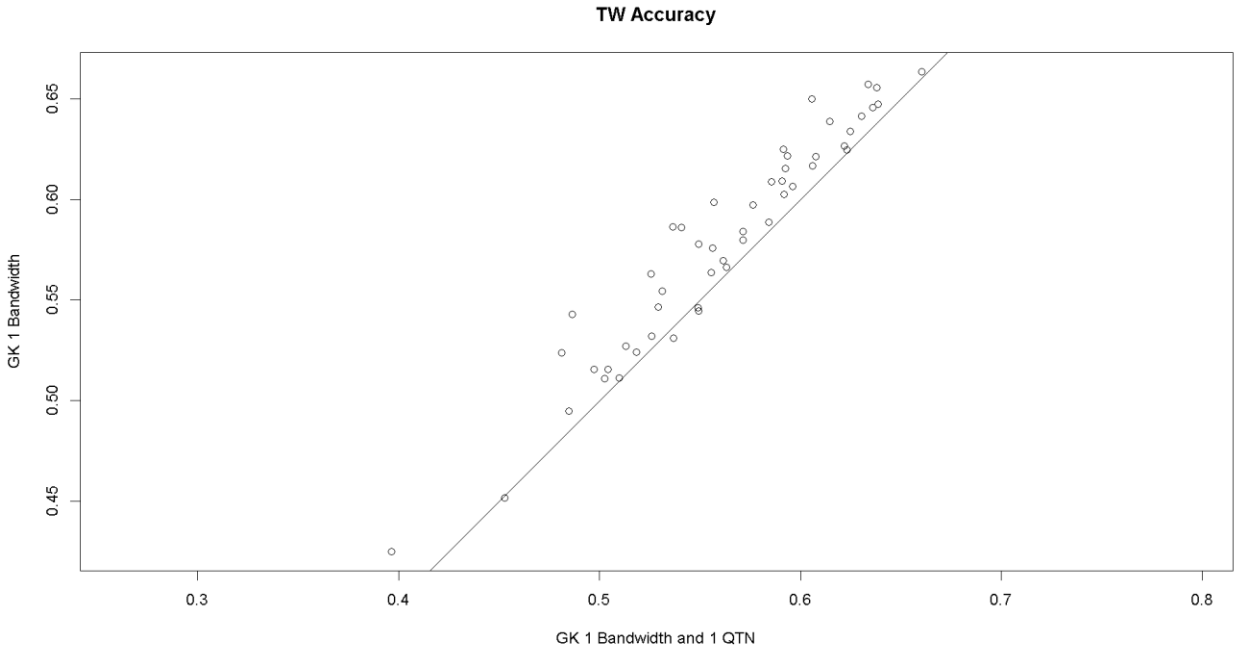
Model	Height_Mean_Cor	Height_Mean_SD	Yield_Mean_Cor	Yield_Mean_SD	Heading_Mean_Cor	Heading_Mean_SD	Protein_Mean_Cor	Protein_Mean_SD	TW_Mean_Cor	TW_Mean_SD	Grand_Mean_Cor	Grand_Mean_SD
QTN1	0.164	0.1111	0.3473	0.0677	0.2053	0.0827	0.2109	0.0967	0.3658	0.0655	0.2587	0.0847
QTN3	0.2561	0.122	0.3971	0.0604	0.3632	0.0821	0.2768	0.0804	0.3583	0.0591	0.3303	0.0808
QTN5	0.327	0.0851	0.4182	0.0602	0.4243	0.0702	0.2968	0.0769	0.3555	0.0599	0.3644	0.0705
GK h .01	0.5878	0.0525	0.5463	0.0545	0.5179	0.0619	0.4385	0.052	0.5201	0.0561	0.5221	0.0554
GK h .01 qtn1	0.5743	0.0549	0.5604	0.0497	0.5302	0.0646	0.4107	0.0538	0.4909	0.0539	0.5133	0.0554
GK h .01 qtn3	0.5526	0.0544	0.5524	0.0528	0.5656	0.0659	0.3678	0.0785	0.4422	0.0607	0.4961	0.0625
GK h .01 qtn5	0.531	0.056	0.5352	0.0556	0.5815	0.0692	0.3393	0.0671	0.4124	0.063	0.4799	0.0622
GK h 1	0.6142	0.0482	0.5984	0.0478	0.5443	0.0618	0.4773	0.0499	0.5789	0.0543	0.5626	0.0524
GK h 1 qtn1	0.6003	0.0511	0.6116	0.0444	0.555	0.0674	0.4573	0.0517	0.562	0.0546	0.5572	0.0538
GK h 1 qtn3	0.5797	0.0496	0.6031	0.0479	0.5852	0.0654	0.4407	0.0546	0.5423	0.0519	0.5502	0.0539
GK h 1 qtn5	0.5624	0.0506	0.5892	0.0487	0.5975	0.0673	0.4258	0.0515	0.5251	0.0552	0.54	0.0547
GK h 2	0.617	0.0496	0.6065	0.0472	0.5409	0.0664	0.4747	0.0517	0.5837	0.0542	0.5646	0.0538
GK h 2 qtn1	0.6013	0.0511	0.6199	0.0426	0.5537	0.0723	0.4468	0.0534	0.5628	0.0547	0.5569	0.0548
GK h 2 qtn3	0.5787	0.0506	0.6082	0.0469	0.5823	0.068	0.4329	0.0578	0.541	0.053	0.5486	0.0553
GK h 2 qtn5	0.5621	0.051	0.5941	0.0477	0.5952	0.0672	0.4226	0.0552	0.5247	0.0596	0.5397	0.0561
cblup	0.5832	0.0644	0.5557	0.0495	0.5201	0.0636	0.4243	0.0519	0.5167	0.0509	0.52	0.0561
cblup qtn1	0.5736	0.0645	0.5702	0.0466	0.5307	0.0688	0.4117	0.0521	0.5048	0.0539	0.5182	0.0572
cblup qtn3	0.5646	0.0633	0.5666	0.0502	0.5683	0.0662	0.4018	0.0626	0.4896	0.0535	0.5182	0.0592
cblup qtn5	0.5589	0.0583	0.5538	0.0515	0.582	0.0684	0.3918	0.0588	0.4775	0.0557	0.5128	0.0585
gBLUP	0.5878	0.0522	0.5457	0.0547	0.5195	0.0607	0.4396	0.0516	0.5196	0.0553	0.5224	0.0549
gBLUP qtn1	0.5754	0.0548	0.5605	0.05	0.5312	0.0629	0.4235	0.0545	0.5022	0.0558	0.5186	0.0556
gBLUP qtn3	0.5561	0.0546	0.5548	0.0528	0.5671	0.0649	0.408	0.0579	0.4818	0.0542	0.5136	0.0569
gBLUP qtn5	0.5396	0.0547	0.5438	0.0532	0.5832	0.0683	0.3949	0.0533	0.4626	0.0547	0.5048	0.0568
polynomial	0.5992	0.0511	0.5656	0.0524	0.5293	0.0601	0.462	0.0501	0.5474	0.0548	0.5407	0.0537
polynomial qtn1	0.5856	0.0541	0.579	0.0481	0.5407	0.0644	0.4429	0.0525	0.5298	0.0537	0.5356	0.0546
polynomial qtn3	0.5653	0.053	0.5712	0.0509	0.5746	0.0642	0.4229	0.0549	0.5083	0.0541	0.5285	0.0554
polynomial qtn5	0.547	0.0533	0.5586	0.0519	0.588	0.0675	0.4072	0.0522	0.4871	0.0553	0.5176	0.056
sigmoid	0.5577	0.0578	0.508	0.0587	0.4784	0.0663	0.3877	0.0567	0.4692	0.0575	0.4802	0.0594
sigmoid qtn1	0.5438	0.0595	0.5278	0.0525	0.4973	0.0693	0.3681	0.0617	0.4383	0.0577	0.4751	0.0601
sigmoid qtn3	0.5248	0.0603	0.524	0.0557	0.5447	0.0692	0.3526	0.0664	0.4096	0.0552	0.4711	0.0614
sigmoid qtn5	0.5076	0.0602	0.5113	0.0554	0.5663	0.0719	0.342	0.0657	0.393	0.063	0.464	0.0632

Within Year Soft 2021 Prediction Accuracy averages and standard deviation over 50 replicates.

Color scaling is applied per column with green for Mean Correlation showing higher average values while green for Mean Standard Deviation of Correlation indicates lower spread. Average correlation was calculated by averaging the correlation between predicted vs observed values in the testing genotypes over fifty random replicates. Standard deviation (SD) values are the population standard deviation for correlation values over the fifty replicates. The same fifty training and test splits were used for evaluating all models, this allows for pairwise comparison of accuracy. The below plots show accuracy between an RKHS GK model with no GWAS results in the model versus an RKHS GK model with one QTN and 3 PC included as fixed effects in the prediction of Yield. The line represents a 1 to 1 relationship in accuracy between the two models, dots above the line show higher accuracy for the model on the y axis, dots below show higher accuracy for the x axis model. For most of the fifty splits, the inclusion of GWAS results improved accuracy.



For Test Weight, inclusion of GWAS results reduces accuracy. Using the same GK bandwidth and the inclusion of one SNP from GWAS results, most train/test splits have higher accuracy without GWAS results included. This is shown with most dots being above the one-to-one accuracy line, meaning the model on the y axis has more random replicates of train/test splits with higher accuracy.



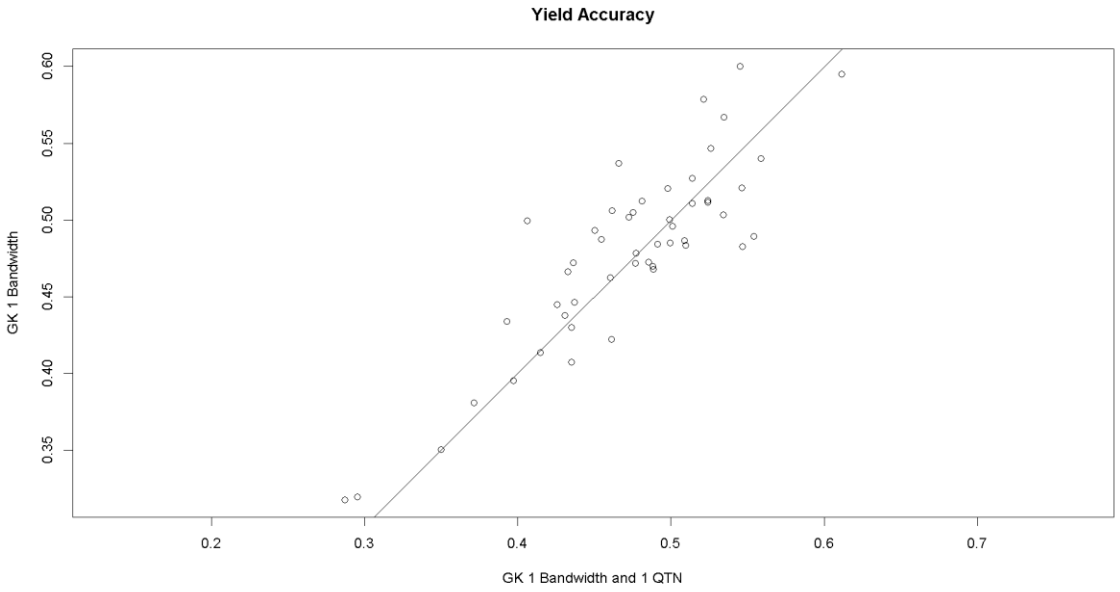
Soft 2022

Averaging the fifty correlation values within model and per trait in the Soft 2022 scenario also shows that the RKHS with GK perform the best, this time with the 1 bandwidth hyperparameter performing better than 2 as in the Soft 2021 data. In both cases, these two Kernels performed very closely on average. The best model in this scenario, RKHS GK h 1, never had greater average prediction accuracy from the inclusion of GWAS results as fixed effects. Additionally, the standard deviation of correlation values increases as GWAS results are incorporated. Prediction accuracy with GBLUP is also not improved as GWAS results are incorporated for all five traits. On average, GBLUP accuracy becomes more variable shown by higher SD values as GWAS results are incorporated. CBLUP performed the worst on average, even performing less accurately in the testing data than the three fixed effect models. CBLUP did benefit from the inclusion of GWAS results, but accruacy was still worse than GBLUP or RKHS with no significant SNP included.

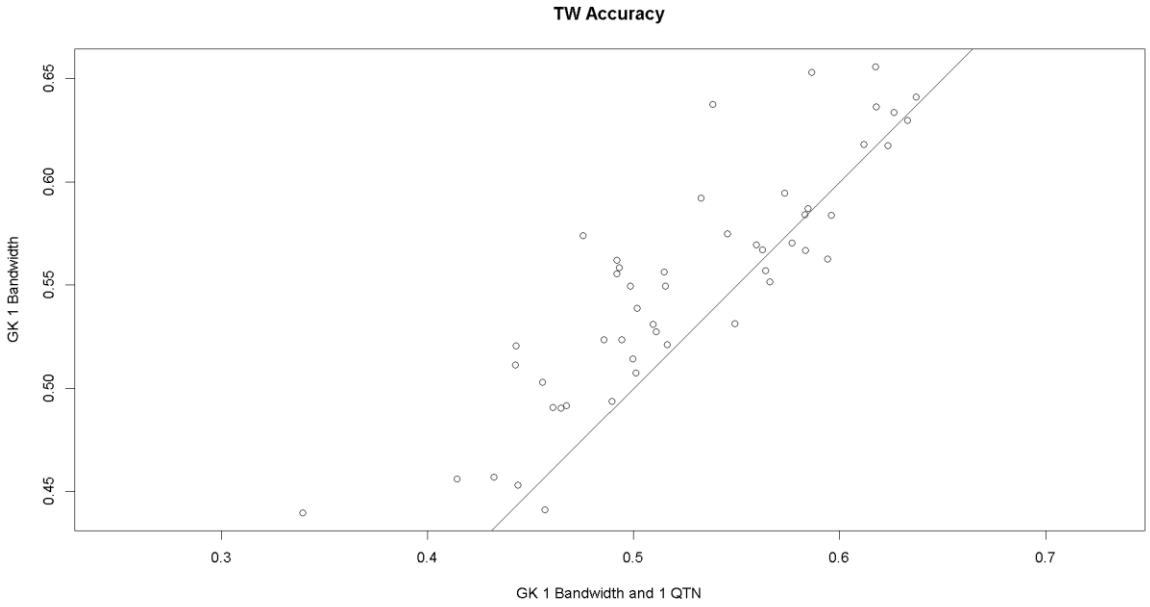
Model	Height_Mean_Cor	Height_Mean_SD	Yield_Mean_Cor	Yield_Mean_SD	Heading_Mean_Cor	Heading_Mean_SD	Protein_Mean_Cor	Protein_Mean_SD	TW_Mean_Cor	TW_Mean_SD	Grand_Mean_Cor	Grand_Mean_SD
QTN1	0.1983	0.0835	0.2347	0.0991	0.3852	0.1254	0.2329	0.0752	0.3019	0.1068	0.2706	0.098
QTN3	0.271	0.0909	0.2859	0.101	0.4342	0.1252	0.2543	0.07	0.3393	0.0931	0.31694	0.09604
QTN5	0.3045	0.0946	0.3231	0.0969	0.4728	0.1257	0.279	0.0726	0.3618	0.0857	0.34824	0.0951
GK h .01	0.493	0.0706	0.4496	0.0601	0.553	0.0623	0.4945	0.054	0.5343	0.0585	0.50488	0.0611
GK h .01 qtn1	0.476	0.0744	0.4375	0.0648	0.5386	0.0829	0.4606	0.066	0.4948	0.0685	0.4815	0.07132
GK h .01 qtn3	0.4587	0.0735	0.3948	0.0776	0.516	0.1016	0.3976	0.0721	0.4512	0.076	0.44366	0.08016
GK h .01 qtn5	0.4269	0.0803	0.3752	0.0791	0.5055	0.1171	0.3423	0.0734	0.4185	0.0822	0.41368	0.08642
GK h 1	0.5341	0.0655	0.4789	0.0603	0.5751	0.0636	0.5478	0.0533	0.5511	0.056	0.5374	0.05974
GK h 1 qtn1	0.517	0.0682	0.4721	0.064	0.5672	0.0758	0.5259	0.0579	0.5255	0.066	0.52154	0.06638
GK h 1 qtn3	0.502	0.0691	0.4546	0.0662	0.5567	0.0848	0.4884	0.0581	0.5017	0.0574	0.50068	0.06712
GK h 1 qtn5	0.4873	0.0705	0.4461	0.0655	0.5561	0.0973	0.4653	0.0638	0.4883	0.0584	0.48862	0.0711
GK h 2	0.5422	0.0623	0.4719	0.0618	0.5673	0.0686	0.555	0.0524	0.5342	0.0552	0.53412	0.06006
GK h 2 qtn1	0.518	0.0666	0.4665	0.0672	0.5592	0.0809	0.5273	0.0581	0.51	0.0673	0.5162	0.06802
GK h 2 qtn3	0.5031	0.0693	0.4505	0.0683	0.5542	0.0901	0.4882	0.0595	0.4926	0.0601	0.49772	0.06946
GK h 2 qtn5	0.4863	0.0712	0.4484	0.0652	0.5555	0.1006	0.466	0.0649	0.4829	0.061	0.48782	0.07258
cblup	0.1877	0.0751	-0.1103	0.0842	0.1072	0.2892	0.1571	0.093	-0.1563	0.0838	0.03708	0.12506
cblup qtn1	0.2151	0.0888	-0.1093	0.0843	0.2258	0.2597	0.1801	0.0949	-0.0162	0.1117	0.0991	0.12788
cblup qtn3	0.2826	0.0918	-0.1082	0.0839	0.2917	0.2499	0.2173	0.0879	0.0592	0.1071	0.14852	0.12412
cblup qtn5	0.3126	0.0913	-0.1077	0.0839	0.3498	0.2374	0.2507	0.0865	0.1159	0.1086	0.18426	0.12154
gBLUP	0.4912	0.0706	0.4523	0.0612	0.5537	0.0628	0.4965	0.0544	0.5365	0.0601	0.50604	0.06182
gBLUP qtn1	0.4768	0.0747	0.4452	0.0638	0.5511	0.0711	0.4775	0.057	0.5163	0.0651	0.49338	0.06634
gBLUP qtn3	0.4635	0.0732	0.4302	0.0655	0.5438	0.0806	0.4419	0.0603	0.4937	0.0584	0.47462	0.0676
gBLUP qtn5	0.4516	0.0742	0.4217	0.068	0.5423	0.0944	0.415	0.0642	0.477	0.0563	0.46152	0.07142
polynomial	0.5124	0.0666	0.4666	0.0606	0.5658	0.0625	0.5206	0.0547	0.5455	0.0574	0.52218	0.06036
polynomial qtn1	0.4963	0.0716	0.4555	0.063	0.5602	0.0731	0.4996	0.0577	0.5217	0.0657	0.50666	0.06622
polynomial qtn3	0.483	0.0696	0.4389	0.066	0.5488	0.0818	0.4627	0.0591	0.4967	0.0577	0.48602	0.06684
polynomial qtn5	0.4676	0.0713	0.4301	0.0669	0.5456	0.0965	0.4322	0.065	0.4787	0.0578	0.47084	0.0715
sigmoid	0.436	0.0817	0.3969	0.0648	0.5126	0.0656	0.4221	0.0585	0.4887	0.0643	0.45126	0.06698
sigmoid qtn1	0.4186	0.0806	0.3889	0.0703	0.5103	0.0796	0.3864	0.0618	0.4624	0.0665	0.43332	0.07176
sigmoid qtn3	0.4068	0.0791	0.3715	0.0727	0.5037	0.0949	0.3401	0.0615	0.4372	0.0633	0.41186	0.0743
sigmoid qtn5	0.3935	0.0801	0.3726	0.0751	0.5058	0.1105	0.3253	0.0651	0.4221	0.0611	0.40386	0.07838

Within Year Soft 2022 Prediction Accuracy average and standard deviation over 50 replicates.

Looking at pairwise comparison over the fifty replicates for yield with an RKHS GK 1 h model in prediction of yield, we can see that in the Soft 2022 data included GWAS results did not improve accuracy as clearly as it did in 2021.



Repeating this pairwise comparison for prediction of Test Weight, here the inclusion of fixed effects from GWAS as shown by the model on the X axis leads to loss of accuracy like with the 2021 soft data.



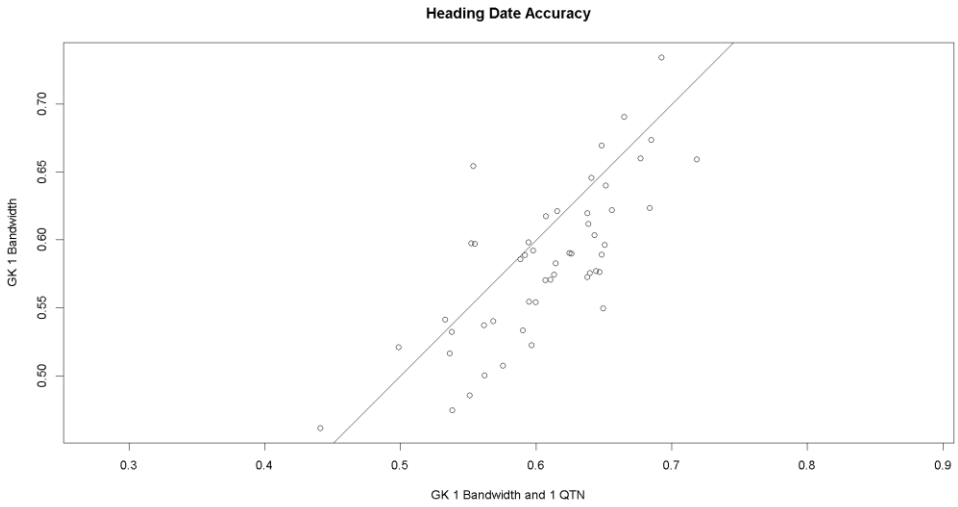
Hard 2021

The RKHS regression method with a Gaussian Kernal created with 1 bandwidth hyperparameter again performed the best on average for all traits in this scenario. Including fixed effects from GWAS results did not improve accuracy in general, it only improved accuracy of one trait. The only trait that benefitted from inclusion of GWAS results was heading date. GBLUP and RKHS methods had improved accuracy for heading date with the inclusion of one SNP as a fixed effect, including more SNPs as fixed effects reduced accuracy. The three fixed effect models performed the worst in terms of average accuracy and spread of accuracy.

Model	Height_Mean_Cor	Height_Mean_SD	Yield_Mean_Cor	Yield_Mean_SD	Heading_Mean_Cor	Heading_Mean_SD	Protein_Mean_Cor	Protein_Mean_SD	TW_Mean_Cor	TW_Mean_SD	Grand_Mean_Cor	Grand_Mean_SD
QTN1	0.3113	0.087	0.3355	0.1033	0.4097	0.1032	0.4411	0.0759	0.4314	0.0877	0.3858	0.09142
QTN3	0.3056	0.0838	0.3308	0.1144	0.4334	0.0867	0.4417	0.0737	0.4503	0.0818	0.39236	0.08808
QTN5	0.3118	0.0736	0.3271	0.1041	0.4427	0.0812	0.4346	0.0725	0.4652	0.0825	0.39628	0.08278
GK h .01	0.4284	0.0698	0.4334	0.079	0.5556	0.0577	0.5928	0.0696	0.6408	0.0551	0.5302	0.06624
GK h .01 qtn1	0.3379	0.09	0.3895	0.0988	0.5857	0.0569	0.5752	0.0744	0.6352	0.0591	0.5047	0.07584
GK h .01 qtn3	0.32	0.0779	0.3451	0.1079	0.556	0.0561	0.5373	0.0925	0.6217	0.0595	0.47602	0.07878
GK h .01 qtn5	0.3198	0.0754	0.3271	0.1056	0.5113	0.0696	0.5018	0.0862	0.6058	0.0629	0.45316	0.07994
GK h 1	0.4765	0.077	0.4969	0.0734	0.5842	0.0568	0.6185	0.0649	0.6591	0.0517	0.56704	0.06476
GK h 1 qtn1	0.4621	0.0762	0.4852	0.0798	0.6079	0.0536	0.5997	0.0673	0.651	0.0558	0.56118	0.06654
GK h 1 qtn3	0.4312	0.0779	0.4447	0.082	0.5875	0.0536	0.5821	0.0697	0.636	0.0561	0.5363	0.06786
GK h 1 qtn5	0.4189	0.0769	0.4173	0.0772	0.5671	0.0576	0.5579	0.0696	0.6231	0.057	0.51686	0.06766
GK h 2	0.4799	0.0787	0.5064	0.0747	0.5879	0.0561	0.62	0.064	0.652	0.0518	0.56924	0.06506
GK h 2 qtn1	0.4624	0.0814	0.492	0.0815	0.6067	0.0561	0.5959	0.0667	0.6411	0.0593	0.55962	0.069
GK h 2 qtn3	0.4341	0.0784	0.4515	0.0844	0.5862	0.0548	0.5742	0.0685	0.6224	0.0586	0.53368	0.06894
GK h 2 qtn5	0.4233	0.0791	0.4236	0.0781	0.5673	0.061	0.5496	0.0708	0.6077	0.0599	0.5143	0.06978
cblup	0.4023	0.0799	0.4406	0.0836	0.5552	0.0599	0.5785	0.0744	0.6261	0.0589	0.52054	0.07134
cblup qtn1	0.3889	0.0828	0.4436	0.0909	0.5871	0.0587	0.5701	0.0751	0.6254	0.0579	0.52302	0.07308
cblup qtn3	0.3574	0.0798	0.4162	0.0867	0.568	0.068	0.5513	0.0775	0.6077	0.0583	0.50012	0.07406
cblup qtn5	0.3571	0.074	0.3935	0.07	0.5485	0.0668	0.53	0.0778	0.5944	0.0582	0.4847	0.06936
gBLUP	0.4311	0.0688	0.4349	0.0783	0.555	0.0589	0.5914	0.0713	0.6383	0.0563	0.53014	0.06672
gBLUP qtn1	0.4173	0.0708	0.4297	0.0846	0.5867	0.0567	0.578	0.0731	0.6335	0.0589	0.52904	0.06882
gBLUP qtn3	0.38	0.0721	0.3872	0.0834	0.5661	0.0541	0.5641	0.0763	0.624	0.059	0.50428	0.06898
gBLUP qtn5	0.3641	0.0733	0.3591	0.08	0.5441	0.0583	0.5397	0.0745	0.6147	0.059	0.48434	0.06902
polynomial	0.4568	0.0724	0.4703	0.0755	0.5689	0.0577	0.6035	0.0687	0.6528	0.0532	0.55046	0.0655
polynomial qtn1	0.4383	0.071	0.4567	0.0842	0.5962	0.0555	0.5884	0.0705	0.6474	0.0565	0.5454	0.06754
polynomial qtn3	0.4005	0.0726	0.4132	0.084	0.5743	0.0536	0.5711	0.0738	0.6346	0.0572	0.51874	0.06824
polynomial qtn5	0.3879	0.0727	0.3856	0.0777	0.5505	0.0587	0.5444	0.0736	0.6225	0.0573	0.49818	0.068
sigmoid	0.3825	0.0714	0.3591	0.0877	0.5007	0.0625	0.5606	0.0722	0.5948	0.0612	0.47954	0.071
sigmoid qtn1	0.3476	0.0794	0.3647	0.1007	0.5476	0.0639	0.5453	0.0783	0.5968	0.0642	0.4804	0.0773
sigmoid qtn3	0.326	0.0778	0.3432	0.1047	0.52	0.0623	0.5195	0.085	0.5903	0.062	0.4598	0.07836
sigmoid qtn5	0.327	0.0732	0.3286	0.1014	0.499	0.0663	0.4913	0.0798	0.5793	0.0627	0.44504	0.07668

Within Year Hard 2021 Prediction Accuracy average and standard deviation over 50 replicates.

The pairwise accuracy comparison for heading date prediction with RKHS and GK with bandwidth of 1 shows that in most training/test fold splits, including the GWAS fixed effects shows an improvement in accuracy.



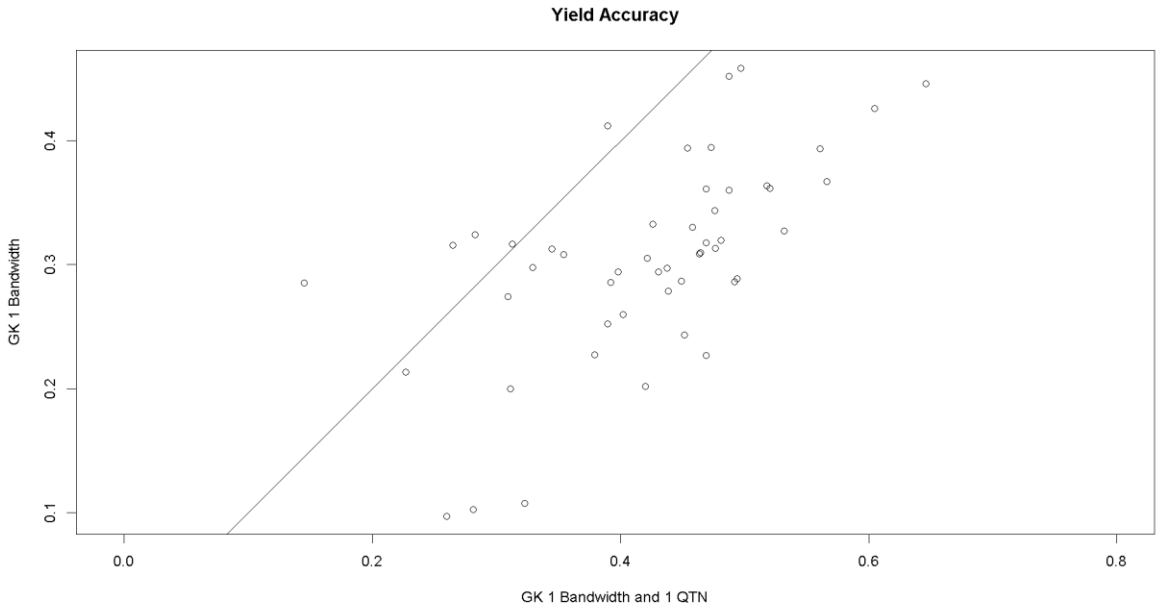
Hard 2022

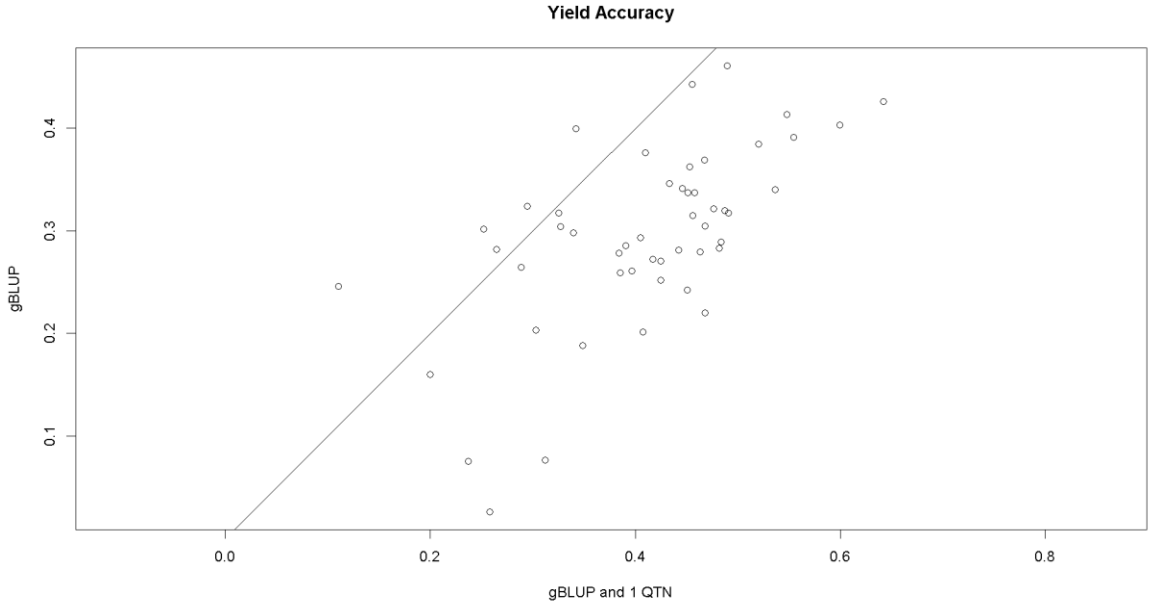
Highest average accuracy was achieved with an RKHS model using a GK with 2 bandwidth hyperparameter. For the four individual experimental validation sets, this is the only time including GWAS results in the GS models increased accuracy on average. However, only Yield benefited from including GWAS results, no other traits benefited in accuracy.

Within Year Hard 2022 Prediction Accuracy average and standard deviation over 50 replicates.

Model	Height_Mean_Cor	Height_Mean_SD	Yield_Mean_Cor	Yield_Mean_SD	Heading_Mean_Cor	Heading_Mean_SD	Protein_Mean_Cor	Protein_Mean_SD	TW_Mean_Cor	TW_Mean_SD	Grand_Mean_Cor	Grand_Mean_SD
QTN1	0.3418	0.1123	0.3951	0.1065	0.2121	0.11	0.2245	0.0914	0.2493	0.1308	0.28456	0.1102
QTN3	0.343	0.1085	0.3688	0.0971	0.2265	0.1191	0.231	0.1014	0.3606	0.1286	0.30598	0.11094
QTN5	0.348	0.1068	0.3615	0.095	0.2325	0.1236	0.2359	0.0946	0.3718	0.1231	0.30994	0.10862
GK h .01	0.4823	0.0935	0.2928	0.083	0.4109	0.0794	0.3219	0.1016	0.3742	0.0914	0.37642	0.08978
GK h .01 qtn1	0.435	0.1045	0.399	0.107	0.351	0.0986	0.2374	0.09	0.3366	0.0905	0.3518	0.09812
GK h .01 qtn3	0.3787	0.1053	0.3699	0.097	0.3088	0.1132	0.237	0.101	0.3735	0.1148	0.33358	0.10626
GK h .01 qtn5	0.3625	0.1122	0.3623	0.0955	0.2742	0.1187	0.2359	0.097	0.3821	0.1106	0.3234	0.1068
GK h 1	0.4904	0.095	0.3055	0.08	0.4194	0.0795	0.3484	0.1014	0.3783	0.0865	0.3884	0.08848
GK h 1 qtn1	0.466	0.1031	0.4226	0.0992	0.3869	0.0893	0.3194	0.0996	0.3789	0.0834	0.39476	0.09492
GK h 1 qtn3	0.4273	0.0989	0.386	0.095	0.3559	0.0987	0.295	0.101	0.4068	0.1048	0.3742	0.09968
GK h 1 qtn5	0.4083	0.1057	0.372	0.0915	0.3228	0.1048	0.2864	0.1011	0.4062	0.102	0.35914	0.10102
GK h 2	0.4906	0.095	0.3116	0.0752	0.4301	0.0702	0.3677	0.0978	0.3741	0.0844	0.39482	0.08452
GK h 2 qtn1	0.4589	0.1066	0.4293	0.0955	0.3858	0.0878	0.3367	0.0987	0.3779	0.082	0.39772	0.09412
GK h 2 qtn3	0.4231	0.1018	0.3917	0.0941	0.3506	0.0988	0.3111	0.103	0.4108	0.1028	0.37746	0.1001
GK h 2 qtn5	0.4042	0.1071	0.3773	0.091	0.3207	0.1075	0.2993	0.1023	0.41	0.1024	0.3623	0.10206
cblup	0.2799	0.1144	0.1041	0.0921	0.2811	0.1407	0.2057	0.1069	0.1757	0.0992	0.2093	0.11066
cblup qtn1	0.2934	0.1218	0.1186	0.0931	0.2734	0.1458	0.2297	0.1046	0.2241	0.0945	0.22784	0.11196
cblup qtn3	0.2993	0.1178	0.1221	0.093	0.278	0.1339	0.2432	0.1067	0.2972	0.0946	0.24796	0.1092
cblup qtn5	0.3112	0.1161	0.1249	0.0922	0.2648	0.1255	0.2508	0.1002	0.3189	0.0901	0.25412	0.10482
gBLUP	0.4807	0.0936	0.2948	0.0886	0.4061	0.083	0.3192	0.1029	0.3735	0.0934	0.37486	0.0923
gBLUP qtn1	0.4587	0.1029	0.4095	0.1042	0.3726	0.0934	0.2914	0.095	0.3728	0.086	0.381	0.0963
gBLUP qtn3	0.4195	0.0972	0.3712	0.0977	0.343	0.1017	0.2644	0.1005	0.3956	0.1048	0.35874	0.10038
gBLUP qtn5	0.3935	0.1073	0.3618	0.0936	0.3138	0.107	0.2567	0.0975	0.3943	0.1027	0.34402	0.10162
polynomial	0.4866	0.0936	0.2991	0.0852	0.4125	0.0817	0.3291	0.1032	0.3778	0.0911	0.38102	0.09096
polynomial qtn1	0.4602	0.1039	0.4156	0.1025	0.3759	0.0921	0.299	0.0972	0.3725	0.0855	0.38464	0.09624
polynomial qtn3	0.4231	0.099	0.381	0.0964	0.3426	0.0982	0.2768	0.1015	0.4012	0.1064	0.36494	0.1003
polynomial qtn5	0.4021	0.1063	0.3686	0.0919	0.3185	0.1075	0.2705	0.0988	0.402	0.1032	0.35234	0.10154
sigmoid	0.4727	0.0938	0.288	0.0899	0.4056	0.0772	0.3072	0.1015	0.3668	0.0922	0.36806	0.09092
sigmoid qtn1	0.4395	0.1058	0.4039	0.1063	0.3508	0.0932	0.2563	0.0893	0.3514	0.0895	0.36038	0.09682
sigmoid qtn3	0.3925	0.1011	0.3719	0.0975	0.3058	0.1121	0.2456	0.1004	0.3859	0.109	0.34034	0.10402
sigmoid qtn5	0.3768	0.1107	0.3628	0.0937	0.2827	0.1065	0.2447	0.0958	0.3892	0.1078	0.33124	0.1029

Including a single SNP from GWAS results as a fixed effect improved prediction accuracy of yield for GBLUP and RKHS. This improvement was dramatic, in the case of GLBUP an improvement from 0.2948 to 0.40915 average correlation when the most significant SNP from GWAS was included. The best model, the RKHS with GK and 2 bandwidth improved from 0.3116 with no GWAS results included to 0.4293 with the inclusion of GWAS results.





Both figures show pairwise comparison between a genomic selection model without GWAS results and with most significant SNP from GWAS included as a fixed effect. GBLUP and RKHS were improved in the prediction of yield when the most significant GWAS result were included.

Results for Scenario Two

Soft Wheat Prediction Across Years

In this scenario the phenotypic data from one field season was used to predict performance of the other year's preliminary yield trial. Results show that the RKHS regression

with Gaussian Kernels performed worse than GBLUP in this scenario. Including GWAS results improved prediction accuracy of heading date for several models. Predicting 2021 heading dates from 2022 data with GBLUP had a correlation of 0.3279, including three or five SNPs as fixed effects increased accuracy to 0.4054 and 0.4357 respectively. Predicting years in the opposite direction increased accuracy from 0.3304 without SNPs as fixed effects to 0.3846 and 0.4053 with 3 and 5 SNPs included as fixed effects. This general trend was also present for RKHS with the Polynomial Kernel and all the Gaussian Kernels. Yield prediction across years in each direction was improved by the inclusion of one or three SNPs as fixed effects. GBLUP accuracy increased from 0.1652 to 0.1968 and from 0.1548 to 0.1885 with the inclusion of 3 SNPs when predicting 2021 or 2022 population. The Kernel methods also had mostly increased predictive accuracy when GWAS results were incorporated to as fixed effects.

Soft Cross Year Prediction Accuracies

Average correlation for all traits, the year indicates the testing population. As an example the

Model	Height2021	Height2022	Protein2021	Protein2022	Yield2021	Yield2022	TW2021	TW2022	Heading2021	Heading2022	Average	SD	Median
QTN1	0.1145	0.1768	-0.0374	0.0142	0.2233	0.1907	0.2461	0.1507	0.2566	0.1738	0.1509	0.0915	0.1753
QTN3	0.1912	0.1910	0.0041	0.1206	0.1533	0.1760	0.1933	0.1681	0.2687	0.1173	0.1584	0.0655	0.1720
QTN5	0.2296	0.2128	0.0378	0.1355	0.1704	0.0792	0.1660	0.1682	0.3122	0.2645	0.1776	0.0780	0.1693
Gk h .01	0.2880	0.2419	0.1008	0.2602	0.1828	0.1515	0.2774	0.2252	0.3281	0.3142	0.2370	0.0693	0.2510
Gk h .01 qtn1	0.2803	0.2332	0.0531	0.2448	0.1750	0.1999	0.2622	0.2183	0.3512	0.3758	0.2394	0.0861	0.2390
Gk h .01 qtn3	0.3105	0.2481	0.0711	0.2280	0.2070	0.1932	0.2329	0.2180	0.4179	0.3761	0.2503	0.0931	0.2305
GK h .01 qtn5	0.2951	0.2542	0.0882	0.2221	0.1776	0.1224	0.2061	0.1857	0.4330	0.4059	0.2390	0.1064	0.2141
Gk h1	0.2635	0.2366	0.0937	0.2568	0.1265	0.1632	0.2846	0.2387	0.3084	0.3243	0.2296	0.0733	0.2477
Gk h 1 qtn1	0.2643	0.2405	0.0637	0.2763	0.1503	0.1884	0.2689	0.2148	0.3295	0.3756	0.2372	0.0846	0.2524
Gk h 1 qtn3	0.3202	0.2518	0.0742	0.2429	0.1661	0.1744	0.2269	0.2240	0.4005	0.3598	0.2441	0.0918	0.2349
Gk h 1 qtn5	0.2851	0.2603	0.0955	0.2359	0.1503	0.1343	0.1956	0.2310	0.4279	0.3940	0.2410	0.1018	0.2334
Gk h 2	0.2231	0.2124	0.1020	0.2208	0.1269	0.0880	0.2897	0.2103	0.2934	0.2914	0.2058	0.0731	0.2166
Gk h 2 qtn1	0.2012	0.2224	0.0558	0.2249	0.1641	0.1975	0.2622	0.2299	0.3242	0.3706	0.2253	0.0813	0.2236
Gk h 2 qtn3	0.2656	0.2321	0.0771	0.2136	0.1678	0.2094	0.2169	0.2194	0.3716	0.3469	0.2320	0.0795	0.2182
Gk h 2 qtn5	0.2651	0.2472	0.1000	0.2138	0.1491	0.0924	0.1939	0.2232	0.3970	0.3897	0.2271	0.0994	0.2185
cblup	0.2871	0.0711	0.0858	0.0614	0.1550	0.0429	0.2665	0.0839	0.3292	0.2425	0.1626	0.1027	0.1204
cblup qtn1	0.2758	0.1742	0.0450	0.0604	0.1646	0.1583	0.2603	0.0964	0.3555	0.2567	0.1847	0.0961	0.1694
cblup qtn3	0.3174	0.1885	0.0635	0.0649	0.1949	0.1612	0.2289	0.1253	0.4054	0.1629	0.1913	0.1008	0.1757
cblup qtn5	0.2618	0.2115	0.0817	0.0650	0.1732	0.0646	0.1969	0.1266	0.4357	0.3012	0.1918	0.1121	0.1850
gBLUP	0.2898	0.2496	0.0905	0.2541	0.1652	0.1548	0.2798	0.2312	0.3279	0.3304	0.2373	0.0745	0.2518
gBLUP qtn1	0.2847	0.2421	0.0518	0.2515	0.1714	0.1955	0.2650	0.2265	0.3494	0.3667	0.2405	0.0855	0.2468
gBLUP qtn3	0.3174	0.2509	0.0705	0.2391	0.1968	0.1885	0.2372	0.2248	0.4175	0.3846	0.2527	0.0951	0.2382
gBLUP qtn5	0.3044	0.2613	0.0907	0.2354	0.1753	0.1201	0.2117	0.2227	0.4381	0.4053	0.2465	0.1062	0.2290
polynomial	0.2812	0.2367	0.1002	0.2758	0.1662	0.1545	0.2777	0.2207	0.3193	0.3134	0.2346	0.0696	0.2562
polynomial qtn1	0.2707	0.2453	0.0606	0.2665	0.1536	0.1903	0.2685	0.2398	0.3443	0.3601	0.2400	0.0836	0.2559
polynomial qtn3	0.3207	0.2536	0.0707	0.2431	0.1911	0.1738	0.2382	0.2332	0.4141	0.3790	0.2518	0.0951	0.2407
polynomial qtn5	0.3073	0.2644	0.0949	0.2414	0.1709	0.1182	0.2088	0.2325	0.4337	0.4025	0.2475	0.1050	0.2369
sigmoid	0.2508	0.2030	0.1073	0.2148	0.2227	0.1339	0.2781	0.2000	0.3035	0.3273	0.2241	0.0659	0.2187
sigmoid qtn1	0.2362	0.2147	0.0565	0.2131	0.2150	0.1843	0.2664	0.2083	0.3354	0.3526	0.2283	0.0780	0.2148
sigmoid qtn3	0.2946	0.2257	0.0675	0.2009	0.2131	0.1836	0.2313	0.1972	0.4067	0.3732	0.2394	0.0926	0.2194
sigmoid qtn5	0.2691	0.2421	0.0880	0.2042	0.1784	0.1224	0.1978	0.1864	0.4306	0.3931	0.2312	0.1032	0.2010

Height2021 column shows correlation between predicted vs observed height values with the 2021 trial as the testing population and 2022 as the training.

Hard Wheat Prediction Accuracy Across Years

GBLUP with no SNPs from GWAS results included yielded the highest average correlation. There was a decrease in the standard deviation of correlation as more fixed effect SNPs were included in the model, but this decreased accuracy on average. Yield accuracy was higher with GBLUP and RKHS regression methods as GWAS results were incorporated as fixed

effects. GBLUP accuracy increased from 0.0864 to 0.1788 as the most significant SNP was incorporated for 2021 prediction; 2022 prediction accuracy increased from 0.1063 to 0.1427 with the inclusion of a single SNP. Unlike the soft wheat trials, prediction accuracy for height was not improved with the inclusion of GWAS results. Accuracy steadily goes down for all GS methods has more SNP are included as fixed effects in the models. The three fixed effect only models performed the worst for average accuracy and had relatively high spread in accuracy seen through standard deviation.

Hard Cross Year Prediction Accuracies

Model	Height2021	Height2022	Protein2021	Protein2022	Yield2021	Yield2022	TW2021	TW2022	Heading2021	Heading2022	Average	SD	Median
QTN1	0.2384	0.3052	0.2990	0.0810	0.1909	0.1070	-0.0521	-0.0520	0.1000	0.1141	0.1331	0.1204	0.1105
QTN3	0.2264	0.2790	0.3393	0.1321	0.1786	0.2263	0.0696	0.0083	-0.0251	0.0503	0.1485	0.1151	0.1553
QTN5	0.1878	0.2639	0.3302	0.1555	0.1863	0.1867	0.0359	0.0129	0.0468	0.0762	0.1482	0.0987	0.1709
Gk h .01	0.2985	0.3386	0.3658	0.2673	0.0986	0.0978	0.0683	0.0677	0.3237	0.2181	0.2144	0.1140	0.2427
Gk h .01 qtn1	0.2853	0.3094	0.3029	0.2227	0.1891	0.1095	0.0931	0.0856	0.2850	0.2288	0.2111	0.0837	0.2258
Gk h .01 qtn3	0.2303	0.2742	0.3404	0.2109	0.1797	0.2252	-0.0430	0.0925	0.1631	0.1416	0.1815	0.0996	0.1953
GK h .01 qtn5	0.1944	0.2632	0.3328	0.2035	0.1880	0.1877	-0.1023	0.0909	0.1592	0.1424	0.1660	0.1089	0.1879
Gk h1	0.2888	0.3131	0.3687	0.2888	0.1356	0.1054	0.0774	0.0544	0.3404	0.2220	0.2195	0.1108	0.2554
Gk h 1 qtn1	0.2862	0.2821	0.3561	0.2294	0.1977	0.1457	0.0804	0.0893	0.2891	0.2085	0.2164	0.0862	0.2189
Gk h 1 qtn3	0.2501	0.2710	0.3688	0.2263	0.1836	0.2195	-0.0354	0.1115	0.1957	0.1462	0.1937	0.1016	0.2076
Gk h 1 qtn5	0.2204	0.2687	0.3524	0.2232	0.1916	0.1748	-0.0939	0.0987	0.1867	0.1388	0.1761	0.1116	0.1892
Gk h 2	0.2887	0.3212	0.3680	0.2799	0.1729	0.0933	0.0634	0.0540	0.3431	0.1965	0.2181	0.1125	0.2382
Gk h 2 qtn1	0.2728	0.2632	0.3493	0.2223	0.2087	0.1287	0.0830	0.0855	0.2909	0.2040	0.2109	0.0843	0.2155
Gk h 2 qtn3	0.2429	0.2559	0.3615	0.2287	0.1921	0.2194	-0.0256	0.1073	0.1762	0.1259	0.1884	0.0983	0.2058
Gk h 2 qtn5	0.2177	0.2717	0.3512	0.2191	0.2073	0.1851	-0.0894	0.0974	0.1515	0.1310	0.1743	0.1112	0.1962
cblup	0.2989	0.3133	0.2492	0.2638	0.0850	0.0891	0.1795	0.1874	0.2160	0.1998	0.2082	0.0740	0.2079
cblup qtn1	0.2433	0.2938	0.3072	0.2151	0.1804	0.1373	0.0493	0.1906	0.1886	0.2160	0.2022	0.0706	0.2028
cblup qtn3	0.2285	0.2543	0.3447	0.2130	0.1722	0.2282	0.0827	0.1752	0.0109	0.1859	0.1895	0.0870	0.1994
cblup qtn5	0.1799	0.2591	0.3332	0.2108	0.1773	0.2105	0.0479	0.1807	0.0754	0.1311	0.1806	0.0790	0.1803
gBLUP	0.3069	0.3392	0.3497	0.2680	0.0864	0.1063	0.1543	0.1857	0.3225	0.2125	0.2331	0.0927	0.2403
gBLUP qtn1	0.2929	0.3099	0.3561	0.2188	0.1788	0.1427	-0.0010	0.1887	0.2804	0.2334	0.2201	0.0966	0.2261
gBLUP qtn3	0.2601	0.2755	0.3393	0.2128	0.1680	0.2159	0.0696	0.1736	0.1860	0.1515	0.2052	0.0708	0.1994
gBLUP qtn5	0.2222	0.2738	0.3302	0.2112	0.1756	0.1826	0.0359	0.1793	0.1733	0.1437	0.1928	0.0739	0.1809
polynomial	0.3062	0.3218	0.3817	0.2860	0.1007	0.1049	0.0678	0.0795	0.3322	0.2179	0.2199	0.1146	0.2520
polynomial qtn1	0.2879	0.3130	0.3616	0.2246	0.1808	0.1550	0.0751	0.0880	0.2891	0.2345	0.2210	0.0908	0.2296
polynomial qtn3	0.2646	0.2832	0.3618	0.2213	0.1729	0.2187	-0.0383	0.0946	0.1891	0.1445	0.1912	0.1045	0.2039
polynomial qtn5	0.2188	0.2765	0.3525	0.2201	0.1788	0.1814	-0.0929	0.0938	0.1759	0.1423	0.1747	0.1119	0.1801
sigmoid	0.2808	0.3169	0.3528	0.2504	0.0690	0.0816	0.0763	0.0584	0.2768	0.1953	0.1958	0.1088	0.2228
sigmoid qtn1	0.2751	0.2961	0.3214	0.1924	0.1845	0.1170	0.0847	0.0807	0.2400	0.2040	0.1996	0.0812	0.1982
sigmoid qtn3	0.2391	0.2632	0.3457	0.1907	0.1737	0.2250	-0.0420	0.0909	0.1303	0.1098	0.1726	0.1020	0.1822
sigmoid qtn5	0.1998	0.2520	0.3379	0.1868	0.1827	0.1869	-0.0980	0.0757	0.1311	0.1160	0.1571	0.1096	0.1848

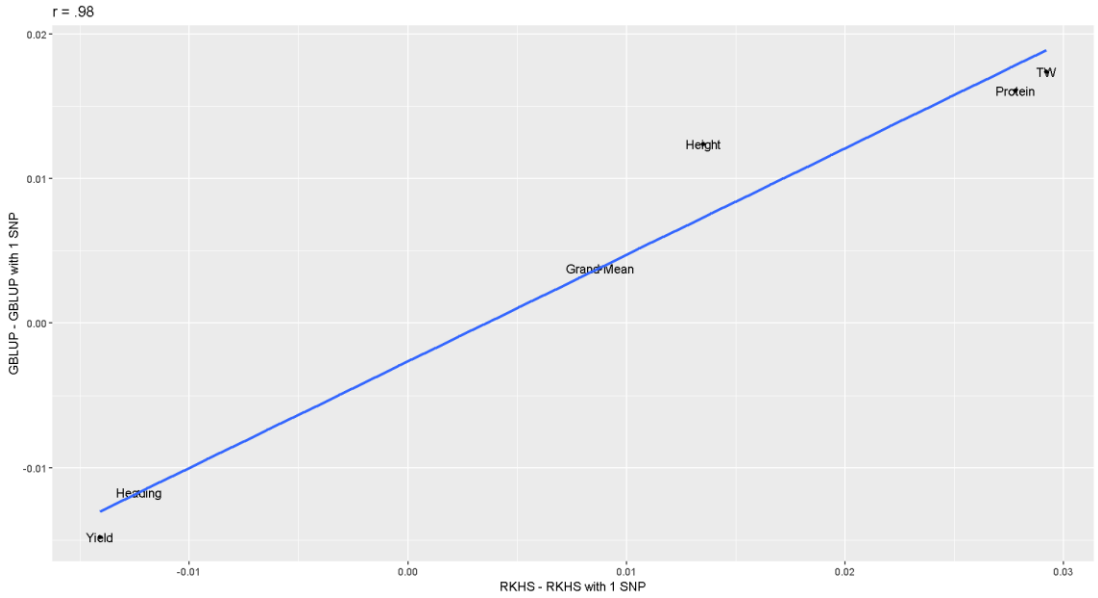
Discussion

The inclusion of SNPs with the lowest P values from GWAS analysis did not improve prediction accuracy for most traits in all analyzed scenarios. Test weight and protein content consistently had the best testing accuracy with GS models that did not involve GWAS results. Yield was the only trait that repeatedly benefitted from the inclusion of GWAS results. Including five SNPs as fixed effects did not improve yield prediction accuracy. Including one or three significant SNPs seemed to be the most likely to improve prediction accuracy for Yield with both GBLUP and RKHS regression. This improvement was most dramatic within the Hard 2022 wheat trials. GBLUP accuracy increased from an average accuracy over fifty replicates from 0.2948 to

0.4095 with the inclusion of the single most significant SNP from GWAS analysis. The hard wheat market class in the WSU spring wheat breeding program has generally much more diverse crosses than the soft market class. It is possible that distance or pest resistance genes that are fixed in the soft advanced germplasm are segregating in the hard germplasm. If that is the case, it would explain why the hard materials benefitted from the inclusion of GWAS results in the prediction of yield.

Plant height and maturity are two traits that have historically benefitted from MAS. The WSU Spring Wheat Program runs molecular markers on crossing block parents whose progeny eventually compose future preliminary yield trials. Two crossing block parents may each be homozygous for a different Rht gene. This results in an F2 and F3 generation with genotypes that are segregating for height, including lines that have two dwarfing genes present. These lines may be too short and suffer yield penalties due to having two semi-dwarfing genes. Visual selection on early generation materials can be used to cull and remove these lines. As a result, genomic selection models trained on the preliminary yield trials generally do not have all possible genotypes that would have been present without early generation visual selection. Genomic selection results for plant height presented in this analysis are contingent on future testing materials to be subject to visual selection. Heading date is also subject to visual selection in the breeding program. As families are advanced through generations of self-pollination, siblings that are relatively earlier in maturity are selected and advanced. As a result, genomic selection models trained on preliminary yield trials are tested on materials that on average have earlier heading dates than expected midparent performance. Prediction accuracy for height and heading date would be lower when tested on materials that were not subject to visual selection. Since most of the genotypes in the prelims are homozygous for one Rht gene or Ppd gene, the results from GWAS analysis and subsequent incorporation into GS explain small portions of variance compared to the effect if these genes were segregating.

Most research into the incorporation of GWAS results into GS methods focus on GBLUP or RRBLUP as the GS method to improve. This work represents one of the first investigations into the potential to improve RKHS regression with most significant GWAS results as fixed effects. When comparing the improvement in including GWAS results into GS, both GBLUP and the RKHS regression with Kernels had a correlated change in prediction accuracy.

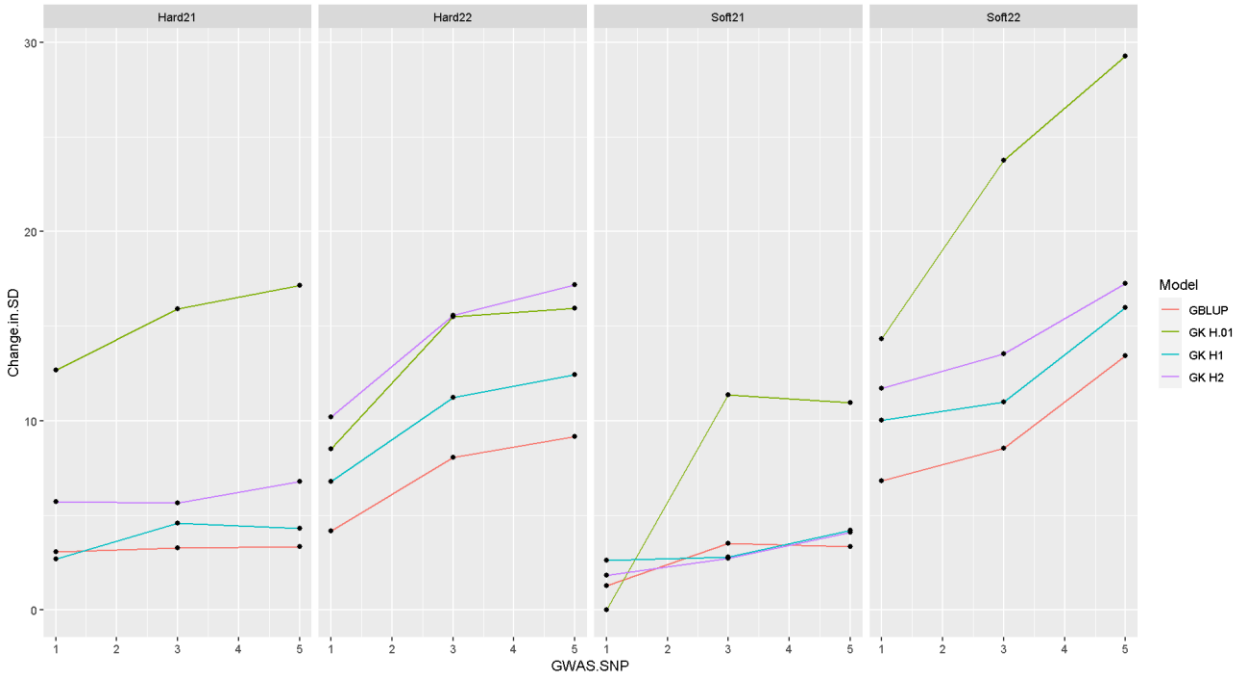


This figure shows the change in prediction accuracy by correlation for Soft 2021 testing when one SNP from GWAS results are included in GBLUP and RKHS with a .01 h GK. Every word shows the same trait and average difference for the two methods. Traits that are negative for a respective axis show an increase in prediction accuracy with incorporation of GWAS results while traits on a positive coordinate have decreased accuracy with GWAS incorporation. The correlation of change in accuracy between both methods is very high. When GBLUP accuracy is improved by including GWAS results, it is generally true that RKHS accuracy will also improve when GWAS results are incorporated. Yield and heading date are both improved with GWAS results included, almost no change in the average correlation of all traits when GWAS results are incorporated, and both methods have a decrease in accuracy for Protein and Test Weight when GWAS results are incorporated. These results may not hold true for different Kernel approaches or if a researcher attempts to tune the Gaussian Kernel while incorporating fixed effects in the tuning process.

For the first scenario of incorporating GWAS results within experiments, only the Hard 2022 scenario had the highest average correlation across all traits with a model that incorporated GWAS results. The other three scenarios of Soft 2021 and 2022 and Hard 2021 had greatest average accuracy with models that did not incorporate any GWAS results as fixed effects. Even in those three cases, the difference between inclusion of a single SNP or standard GS was very small. The Hard 2021 results were only due to the large increase in prediction accuracy for Yield when GWAS results were incorporated in predictions. However, the second scenario of predicting across years into new populations showed the two methods of predicting come close in accuracy. Highest average accuracy was achieved for all traits in the Soft market class with a GBLUP model that incorporated the three most significant SNPs as fixed effects. The Hard market class had the greatest average accuracy with GBLUP that did not incorporate fixed effects.

The advantage in prediction accuracy of the RKHS regression methods that was apparent for the within year prediction accuracy scenario was not present in predicting across years. The three fixed effect only models performed the worse generally in all scenarios. CBLUP performed the worst in both the 2022 Soft and Hard year scenarios. The poor performance of CBLUP could be due to the population structure present in the preliminary yield trials. Several families are evaluated from crosses of advanced materials, meaning a population of several sets of siblings. This rigid population structure may not be suitable for CBLUP in comparison to GBLUP. When testing predictions across years however, the large gap in accuracy between CBLUP and GBLUP or RKHS was much reduced.

The inclusion of GWAS results into GS increased the average standard deviation of testing correlation values for scenario one over the fifty replicates of training and testing splits. This indicates that in this scenario of validating models within populations, there is potentially more variability in the accuracy of GS results when SNPs are included as fixed effects. The average standard deviation for all traits of GBLUP and RKHS models with the Gaussian Kernel increased as SNPs were included. This was determined by finding the percent difference between the average standard deviation of a GS model without any fixed effects and as more fixed effects were include. Over all four Year and Market class within population scenarios, the variability of prediction accuracies increased on averaged once GWAS results were included.



The graph shows that variability in accuracy increased through inclusion of GWAS results. The increase of the Standard Deviation for correlation results was mostly moderate, around or

under 5% for GBLUP in Soft2021 and Hard 2021. However as seen in the Hard2022 and Soft 2022 panels, the increase in accuracy variability can be dramatic.

Conclusion

The inclusion of the most significant GWAS results by P value into GS models as fixed effects did not improve prediction accuracy for Test Weight, Protein Content, or Plant Height. Heading date prediction accuracy was improved in two scenarios, cross year validation in the Soft market class and within year validation for the Hard 2022 trial. Yield prediction accuracy had the greatest improvement in accuracy with the inclusion of GWAS results in both GBLUP and RKHS regression models. However, the variability of prediction accuracies achieved with cross-validation testing within years increased as GWAS results were included in GS models. Fixed effect models that only incorporated Principal Components and GWAS results as predictors performed the worst overall, however this gap in accuracy was smaller when predictions accuracy was measured across years. The three fixed effect only models that mimic a form of marker assisted selection are not effective in comparison to GS methods. Breeding programs looking to implement GS into should evaluate the impact of including significant SNPs as fixed effects, as the genetic architecture of traits present in advanced breeding materials may not benefit from their inclusion.

Literature Cited

Bernardo, R. (2014). Genomewide selection when major genes are known. *Crop Science*, *54*(1), 68–75. https://doi.org/10.2135/cropsci2013.05.0315

Bernardo, R. N. (2020). *Breeding for quantitative traits in plants*. Stemma Press.

De Los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., & Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research, 92(4), 295–308. https://doi.org/10.1017/s0016672310000285

Ellis, M., Spielmeyer, W., Gale, K., Rebetzke, G., & Richards, R. (2002). "perfect" markers for the RHT-B1B and RHT-D1B dwarfing genes in wheat. *Theoretical and Applied Genetics*, *105*(6), 1038–1042. https://doi.org/10.1007/s00122-002-1048-4

Gianola, D., & van Kaam, J. B. (2008). Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of Quantitative Traits. *Genetics*, *178*(4), 2289–2303. https://doi.org/10.1534/genetics.107.084285

Geranios, N. K. (2021, September 21). 2021 a record-breaking drought year in parts of Washington. The Seattle Times. Retrieved April 28, 2023, from https://www.seattletimes.com/seattle-news/environment/2021-a-record-breaking-drought-year-in-much-of-eastern-washington/

Huang, M., Liu, X., Zhou, Y., Summers, R. M., & Zhang, Z. (2017). Blink: A package for next level of genome wide association studies with both individuals and markers in millions. *GigaScience*. https://doi.org/10.1101/227249

McGowan, M., Wang, J., Dong, H., Liu, X., Jia, Y., Wang, X., Iwata, H., Li, Y., Lipka, A. E., & Zhang, Z. (2021). Ideas in genomic selection with the potential to transform plant molecular breeding. *Plant Breeding Reviews*, 273–319. https://doi.org/10.1002/9781119828235.ch7

Merrick, L. F., Burke, A. B., Chen, X., & Carter, A. H. (2021). Breeding with major and minor genes: Genomic selection for quantitative disease resistance. *Frontiers in Plant Science*, *12*. https://doi.org/10.1101/2021.05.20.444894

Miedaner, T., & Korzun, V. (2012). Marker-assisted selection for disease resistance in wheat and barley breeding. *Phytopathology®*, *102*(6), 560–566. https://doi.org/10.1094/phyto-05-11-0157

Montesinos-López, O. A (2022). Chapter 8 Reproducing Kernel Hilbert Spaces Regression and Classification Methods. In *Multivariate Statistical Machine Learning Methods for genomic prediction*. essay, SPRINGER.

Montesinos-López, A., Montesinos-López, O. A., Montesinos-López, J. C., Flores-Cortes, C. A., de la Rosa, R., & Crossa, J. (2021). A guide for kernel generalized regression methods for genomic-enabled prediction. *Heredity*, *126*(4), 577–596. https://doi.org/10.1038/s41437-021-00412-1

Odilbekov, F., Armoniené, R., Koc, A., Svensson, J., & Chawade, A. (2019). GWAS-assisted genomic prediction to predict resistance to septoria tritici blotch in Nordic winter wheat at Seedling Stage. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01224

Prather, S., Schneider, T., Gaham Godoy, J., Odubiyi, S., Bosque-Perez, N. A., Rashed, A., Rynearson, S., & Pumphrey, M. O. (2022). Reliable DNA markers for a previously unidentified, yet broadly deployed Hessian fly resistance gene on chromosome 6b in Pacific Northwest spring wheat varieties. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.779096

Pérez, P., & de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/genetics.114.164442

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Snape, J., Butterworth, K., Whitechurch, E., & Worland, A. J. (2001). Waiting for fine times: Genetics of flowering time in wheat. *Wheat in a Global Environment*, 67–74. https://doi.org/10.1007/978-94-017-3674-9_7

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.-L., & McCouch, S. (2016). Genome-wide prediction models that incorporate de Novo Gwas are a powerful new tool for Tropical Rice Improvement. *Heredity*, *116*(4), 395–408. https://doi.org/10.1038/hdy.2015.113

Wang, J., Zhou, Z., Zhang, Z., Li, H., Liu, D., Zhang, Q., Bradbury, P. J., Buckler, E. S., & Zhang, Z. (2018). Expanding the blup alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity*, *121*(6), 648–662. https://doi.org/10.1038/s41437-018-0075-0

Yan, H., Guo, H., Xu, W., Dai, C., Kimani, W., Xie, J., Zhang, H., Li, T., Wang, F., Yu, Y., Ma, M., Hao, Z., & He, Z. (2023). GWAS-assisted genomic prediction of cadmium accumulation in maize kernel with machine learning and Linear Statistical Methods. *Journal of Hazardous Materials*, *441*, 129929. https://doi.org/10.1016/j.jhazmat.2022.129929

Wang, J. and Zhang, Z. (2021, September). *Gapit version 3: Boosting Power and accuracy for Genomic Association and Prediction*. Genomics, proteomics & bioinformatics. Retrieved April 28, 2023, from https://pubmed.ncbi.nlm.nih.gov/34492338/

Zegeye, H., Rasheed, A., Makdis, F., Badebo, A., & Ogbonnaya, F. C. (2014). Genome-wide association mapping for seedling and adult plant resistance to stripe rust in synthetic hexaploid wheat. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0105593