Exploring the Role of Heritability and QTNs on Genomic Selection and Prediction

Sudha GC Upadhaya

Department of Plant Pathology, Washington State University, Pullman, WA

Summary:

Advancement in genomics, molecular biology, and computational methods have helped to make significant progress in modern crop breeding by accelerating the process in selection of desirable traits. Marker Assisted Selection (MAS), a very popular tool among plant breeders, has been successful for selection of Mendelian traits such as disease resistance, but is not that suitable for complex traits like yield and quality. For selection of polygenic traits and others for which phenotype data collection is expensive, genomic selection (GS) has been found to be effective, and thus has started to be a popular tool among plant breeders in recent years. The accuracy of GS, however, can vary depending on various factors such as heritability of the trait, number of quantitative trait nucleotides (QTNs), and population structure. This case study explores the effect of heritability and the number of associated QTNs on genomic selection and prediction in a panel of 400 worldwide barley accessions, which were genotyped using a 9K Illumina SNP chip. A total of 6,186 high-quality SNP markers were used to evaluate the performance of genomic best linear unbiased prediction (gBLUP) versus genome-wide association study (GWAS)-assisted marker-assisted selection (MAS) utilizing simulated phenotype data at three levels of heritability (0.25, 0.5, and 0.75) and five levels of QTNs number (10, 25, 50, 100, and 500). Four different methods were implemented and evaluated for their accuracy, including genomic selection (GS), GWAS-assisted gBLUP calculation by incorporating in the fixed effect, GWAS-assisted MAS, and a support vector machine (SVM) approach. The methods were replicated 30 times, and the correlation coefficient was used to determine the accuracy of each model. In all the tested methods, predication accuracy was highest at the heritability of 0.75 and lowest at the heritability of 0.25. With the increasing number of QTNs, prediction accuracies were found to decrease. All the tested methods had higher accuracies for highly heritable traits, controlled by few genes, and genomic selection using all marker associations had higher accuracy for low heritable traits compared to other methods. The accuracy of genomic prediction and selection in barley lines is affected by the heritability of the traits and the number of genes controlling them, given the low-resolution genomic data used.

Introduction:

The selection of desirable traits in agricultural crops is a fundamental aspect of plant breeding. For over 10,000 years, farmers have been using phenotype-based selection to identify crops with the desired traits. The success of agriculture and its contribution to human civilization can be attributed to the selection of these desirable traits that enhance plant performance. However, phenotype-based selection has limitations for polygenic and low heritability traits and can be resource-intensive and time-consuming (Spindel et al. 2015; Bhat et al. 2016).

In the past century, advances in genetics and molecular biology have provided us with a better understanding of the underlying genetic mechanisms of these traits. High-throughput genotyping and computational capacity have enabled us to obtain genetic information more efficiently and at a lower cost. One important approach integrated into plant breeding strategies is Marker Assisted Selection (MAS), which uses genotypic information to select desirable traits. MAS has been particularly successful in selecting for Mendelian traits, such as disease-resistant genes, especially those controlled by one or few genes (Poland and Rutkoski 2016). However, this approach has not been effective for selecting complex traits such as yield and quality (Bernando 2002). Recently, there has been growing interest in using genomic selection and prediction-based approaches to improve crop productivity, stress resilience, and quality through plant breeding tools. These techniques are particularly useful for polygenic traits and unbalanced data (Bernardo 2008). Furthermore, genomic selection can prove to be a cost-effective and appropriate tool when it comes to quality traits, where evaluating a large number of breeding lines can be expensive. Also, GS can help to increase genetic gain by reducing the length of the breeding cycle.

Since the first description of genomic selection (GS) by Meuwissen et al. in 2001, it has become a popular tool among breeders. This technique involves the use of genome-wide SNP markers to calculate the breeding value of parents and progenies, allowing for the breeding of complex traits influenced by both small and large effect genes. However, the accuracy of GS can vary depending on factors such as the number of quantitative trait nucleotides (QTNs), trait complexity, population structure and relatedness, marker density, and gene effects (de Moraes et al. 2018). In addition, GS can be noisy when information from a large number of genomewide markers is used. To address this issue, one potential solution is to incorporate genome-wide association study (GWAS) results into GS (McGowan et al. 2021). This involves using SNP information only from the significantly associated markers to build a kinship matrix, rather than including all markers (sBLUP). Additionally, incorporating associated markers in the fixed effect has been shown to improve the accuracy of genomic prediction (McGowan et al. 2021). However, to prevent bias towards the training set when predicting the breeding value or phenotype, it is crucial to employ a valid procedure during the marker trait association analysis, which includes the use of the training data to identify the associated SNPs and predict the value on the test set (McGowan et al. 2021).

The aim of this case study is to investigate the influence of heritability and the number of associated Quantitative Trait Loci (QTL) on genomic selection and prediction. To achieve this, three research questions were asked: (i) Does heritability affect the accuracy of genomic selection and prediction? (ii) Does the number of associated QTNs impact prediction accuracy? (iii) Are there differences in accuracy between genomic selection alone and GWAS-assisted genomic selection and prediction using machine learning and GLM models? In this study, I am specifically trying to explore using the low-resolution SNPs of the 400 barley lines.

Methods:

For this case study, the Barley 9k SNP markers dataset was utilized. This dataset was previously generated by the USDA National Small Grains Collection by genotyping a total of

2,465 barley lines from the world barley core collection (WBCC) using the iSELECT Illumina SNP platform. The dataset was downloaded from The Triticeae Toolbox(T3) website (https://triticeaetoolbox.org/barley/downloads/downloads.php), and lines with less than 10% missing data were included. Additionally, markers with a minimum minor allele frequency of < 5% were dropped, resulting in a dataset with 7,080 marker genotype data for 2,283 lines. Missing genotype data was imputed using the LD KNNi function in Tassel 5 with default settings, and duplicate genotypes were removed in R software. To reduce the sample size, 400 barley lines were randomly selected from 2283 barley lines in R software. After quality control and sample selection, the final dataset consisted of 400 barley lines and 6,186 markers.

The aim of this study was to evaluate the performance of genomic best linear unbiased prediction (gBLUP) versus genome-wide association study (GWAS)-assisted marker-assisted selection (MAS) using simulated phenotype data. To achieve this, phenotype data was simulated with 5, 25, 50, 100, or 500 quantitative trait loci to represent both Mendelian and polygenic traits. Three levels of heritability (0.25, 0.5, and 0.75) were used for the phenotype simulation. The "GAPIT.Phenotype.Simulation" function in GAPIT version 3 was utilized to simulate phenotype data (Wang and Zhang 2021). Briefly, the genetic effect was simulated using the QTNs with their genotype data. $genetic = SNP \cdot N(n, 0,1)$; $Var(g) = var(genetic)$ Then, residual variance was calculated using the genetic variance and the heritability (h^2), $(var(e) = \frac{var(g) - var(g)h^2}{h^2}$ and the residual effect was simulated using a normal distribution with mean of 0 and variance equal to the residual variance ($residual \sim N(0, Var(e))$). Finally, the simulated phenotype was obtained by adding the genetic and residual effects (phenotype = genetic + residual). This function helped to simulate genetic, residual, and phenotype data for downstream analysis.

To evaluate the performance of gBLUP and GWAS-assisted MAS at three levels of heritability and five levels of QTNs, four different methods were implemented and evaluated for their accuracy. First, BLINK was utilized for GWAS, as it has been shown to have higher power than other models, such as generalized linear model (GLM) and mixed linear model (MLM) (Zhang et al. 2010; Huang et al. 2019). To minimize Type I error, the *p*-values were corrected using a stringent Bonferroni correction. The alpha level (p=0.05) was divided by the total number of markers to derive the new cutoff, and only markers with a *p*-value less than this cutoff were utilized for the GWAS-assisted MAS and genomic selection.

The first method used was genomic selection (GS), which uses all markers to build a fixed effect matrix and the kinship matrix. The equation implemented for the mixed linear model is provided below.

$$Y = X\beta + Zu + \varepsilon$$

where *X* is the design matrix for the fixed effect with all maker information, *Z* is the design matrix for the individual effect (kinship matrix), β is the vector of parameters to be estimated for the fixed effect, u is used to calculate the gBLUP, and ε represents the residual error. The second method was GWAS-assisted gBLUP calculation, where the kinship matrix was built using all markers as above but the fixed effect terms use only the significant QTN from GWAS. The

equation is similar to the first method, except that only the significant markers were used to build the design matrix X including PCA. The third method was GWAS-assisted MAS, where a general linear model (GLM) was implemented using only the significant markers in the fixed effect term. The model obtained was then used to make predictions using the equation $Y = X\beta + \varepsilon$. In this equation, X represents the design matrix of the significant markers obtained from GWAS results and the principal component analysis (PCA). The fourth approach utilized a support vector machine (SVM), a popular supervised machine learning algorithm for predicting phenotypes by using the dependent variable. SVM is frequently utilized in predictive modeling because of its ability to manage high-dimensional data, such as genetic data with numerous markers. The SVM was implemented with a linear kernel function (Desta and Ortiz 2014). To account for the structure among the barley genotypes, three principal components were included as covariates in all four methods.

The genotype and phenotype data were divided into training and test sets to implement all four methods. The training set consisted of 80% of the data, while the remaining 20% were included in the test set to avoid overfitting. Because the study involved a small sample size (n) and a larger number of predictors (p), there were no degrees of freedom to evaluate the model accuracy. Thus, unseen data (test set) was utilized to assess the model's generalizability, and the correlation coefficient was used to determine the model's accuracy. Also, the GWAS was conducted only in the training set to identify the associated markers. To obtain a more comprehensive interpretation and account for variations across each simulation, each method was replicated 30 times. The first three methods, i.e., MAS and GS, were implemented using GAPIT V.3, whereas the support vector machine was implemented using the R package "e1071" (Meyer et al. 2019; Wang and Zhang 2021).

Results:

To evaluate the accuracy of breeding value prediction methods, a simulated phenotype was generated using three levels of heritability and five levels of the number of QTNs. The results showed that heritability effects on the accuracy of genetic selection, with the highest accuracy observed at a heritability of 0.75, followed by 0.5 and 0.25 (Figure 1). At the heritability level of 0.75, the test set had an average correlation coefficient ranging from 0.92 to 0.95, with a standard deviation of less than 0.05. On the other hand, the heritability of 0.25 had the lowest overall accuracy in predicting genetic value. Even for the same number of QTNs, the highest correlation coefficient obtained was 0.61 for the heritability of 0.25. Moreover, the average correlation and standard deviation across all methods and QTN number was 0.658 and 0.07, respectively for the heritability of 0.75 in the test set whereas, the average correlation and standard deviation was 0.37 and 0.12, respectively for the heritability of 0.25 in the test set (Figure 2). In Figure 3, the median, first quarter, and third quarter of the correlation values from 30 replications are displayed. The plot indicates a wider range of correlation values for lower heritability, but the range varied depending on the method used in prediction.

The simulation results show that the number of QTN has a significant impact on the accuracy of genetic selection. As the number of QTN increases from 10 to 500, the correlation between genetic selection and predicted value decreases. Specifically, the average correlation

coefficients for 10, 25, 50, 100, and 500 QTNs are 0.79, 0.64, 0.45, 0.37, and 0.29, respectively in the test dataset (Figure 1). The standard deviation was relatively similar across different numbers of QTN, ranging from 0.07 to 0.11 (Figure 2). The decrease in correlation with increase in the number of QTN suggests that the accuracy of genetic selection is highly dependent on the number of QTN used and that the polygenic trait has the lower prediction accuracy compared to that of the Mendelian trait. Figure 3 clearly shows that the correlation between predicted and observed values decreases as the number of QTNs increases, at all three levels of heritability. This pattern was consistently observed across all levels of heritability.

On average,s the correlations and standard deviations of the various methods used in genomic selection showed relatively little variation in the test dataset. The highest accuracy of 0.515 was achieved using all markers in the kinship, while the use of only associated markers in the fixed effect resulted in a slightly lower accuracy of 0.504 (Figure 1). The generalized linear model and support vector machine methods exhibited similar accuracies, with values of 0.506 and 0.505, respectively. The presented standard deviations of the correlations ranged from 0.08 to 0.107 (Figure 2). The results indicated that all methods had higher accuracies for highly heritable traits that were controlled by a small number of genes. The accuracy varied across different methods for different levels of heritability and number of QTN. Genomic selection using all marker associations had a comparatively higher accuracy for low heritable traits. In contrast, for highly heritable traits, the GLM method also exhibited high accuracy. Overall, heritability and the number of QTN underlying the traits tend to have effect in the overall correlation coefficients.

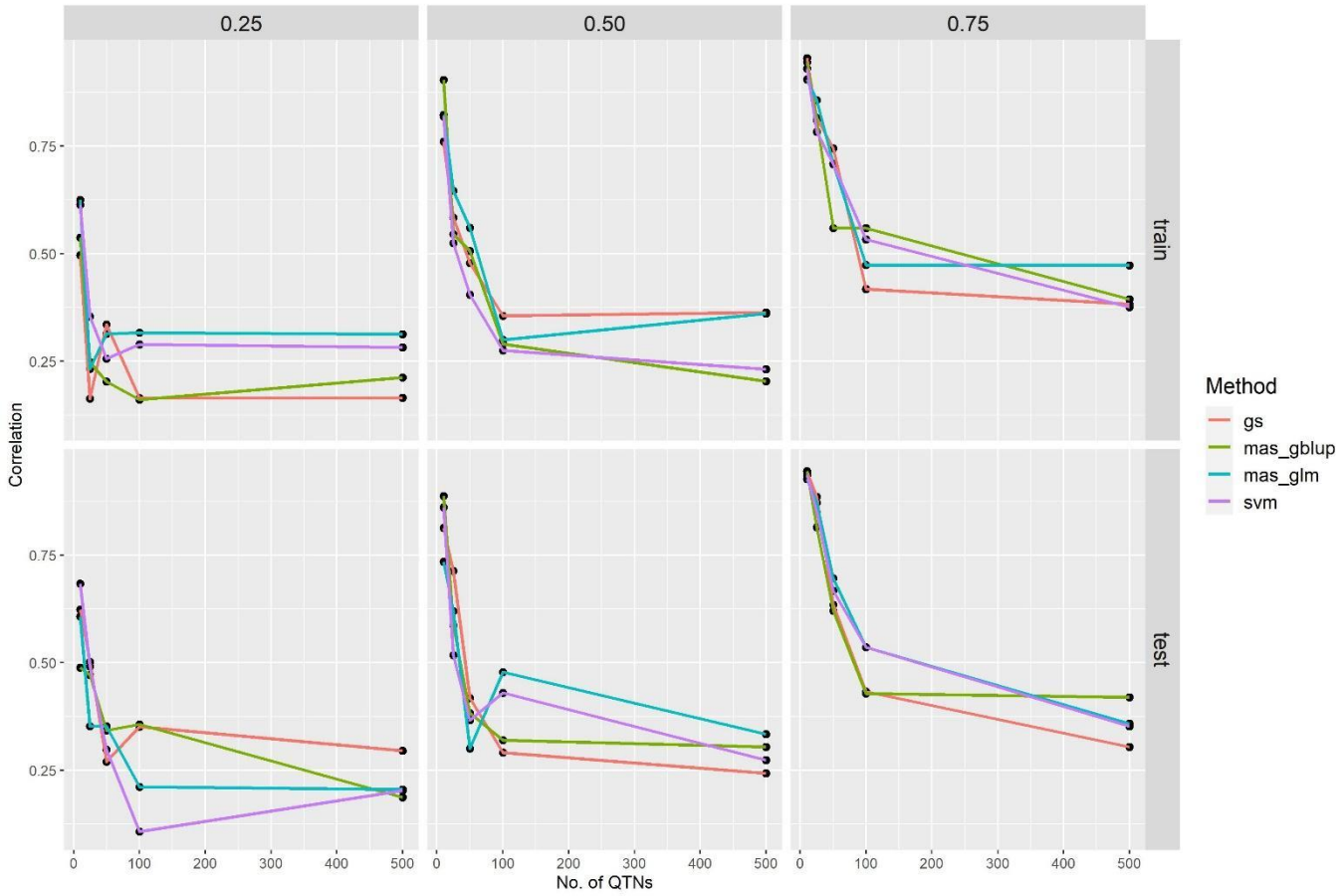

Figure 1: The line plots show the average correlation over 30 replications between the phenotype/genetic and predicted value for various combinations of heritability and number of associated QTN in both the training and test datasets. The x-axis represents the number of QTN with five levels (10, 25, 50, 100, and 500) and the y-axis represents the correlation. The color of the lines indicates the accuracy of four different methods: genomic selection, GWAS assisted genomic selection in fixed effect, phenotype prediction using a generalized linear model, and phenotype prediction using the support vector machine. Three levels of heritability are also displayed: 0.25, 0.5, and 0.75.

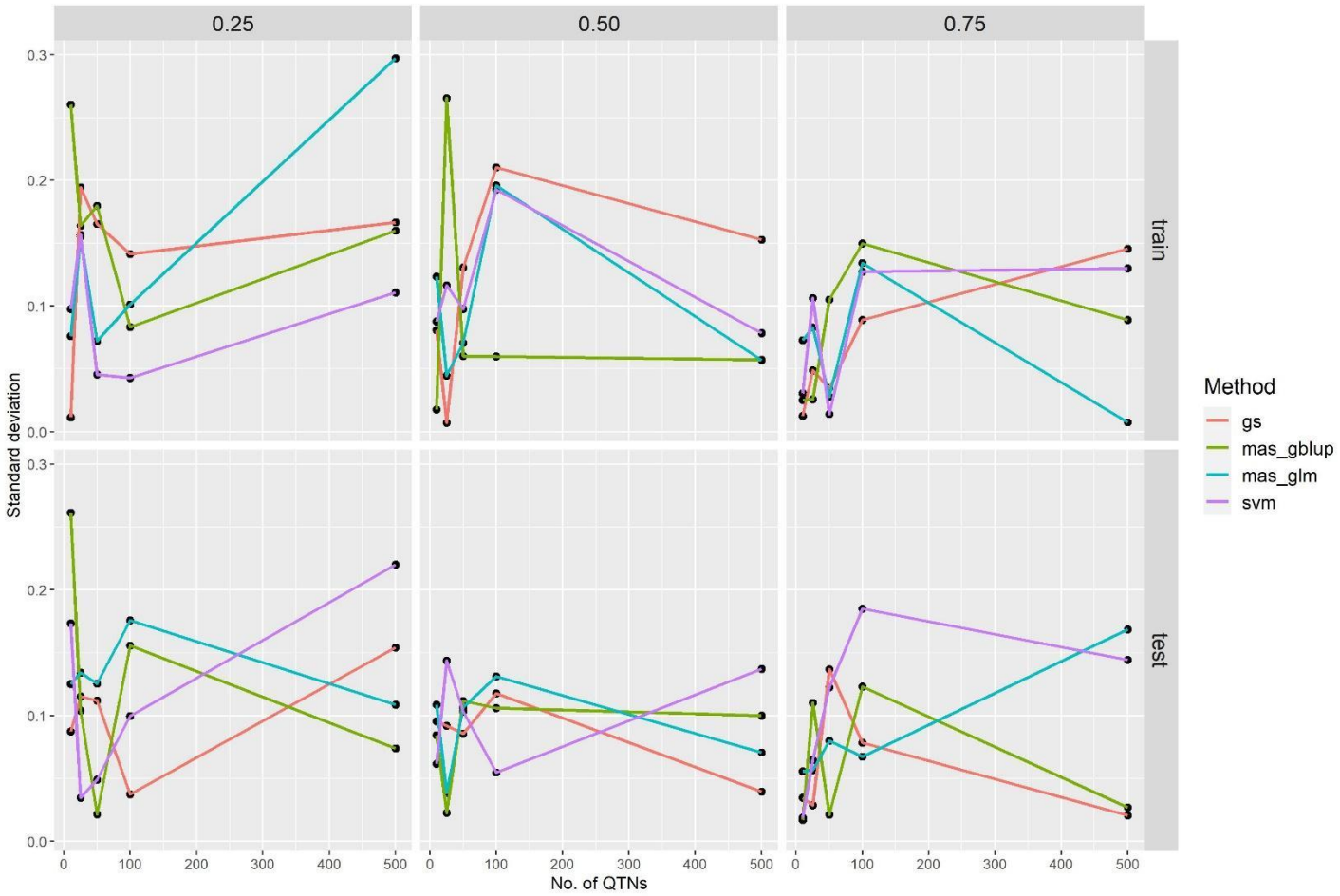

Figure 2: The line plots show the standard deviation of correlation over 30 replications between the phenotype/genetic and predicted value for various combinations of heritability and number of associated QTN in both the training and test datasets. The x-axis represents the number of QTN with five levels (10, 25, 50, 100, and 500) and the y-axis represents the standard deviation. The color of the lines indicates the accuracy of four different methods: genomic selection, GWAS assisted genomic selection (fixed effect), phenotype prediction using a generalized linear model, and phenotype prediction using the support vector machine. Three levels of heritability are also displayed: 0.25, 0.5, and 0.75.

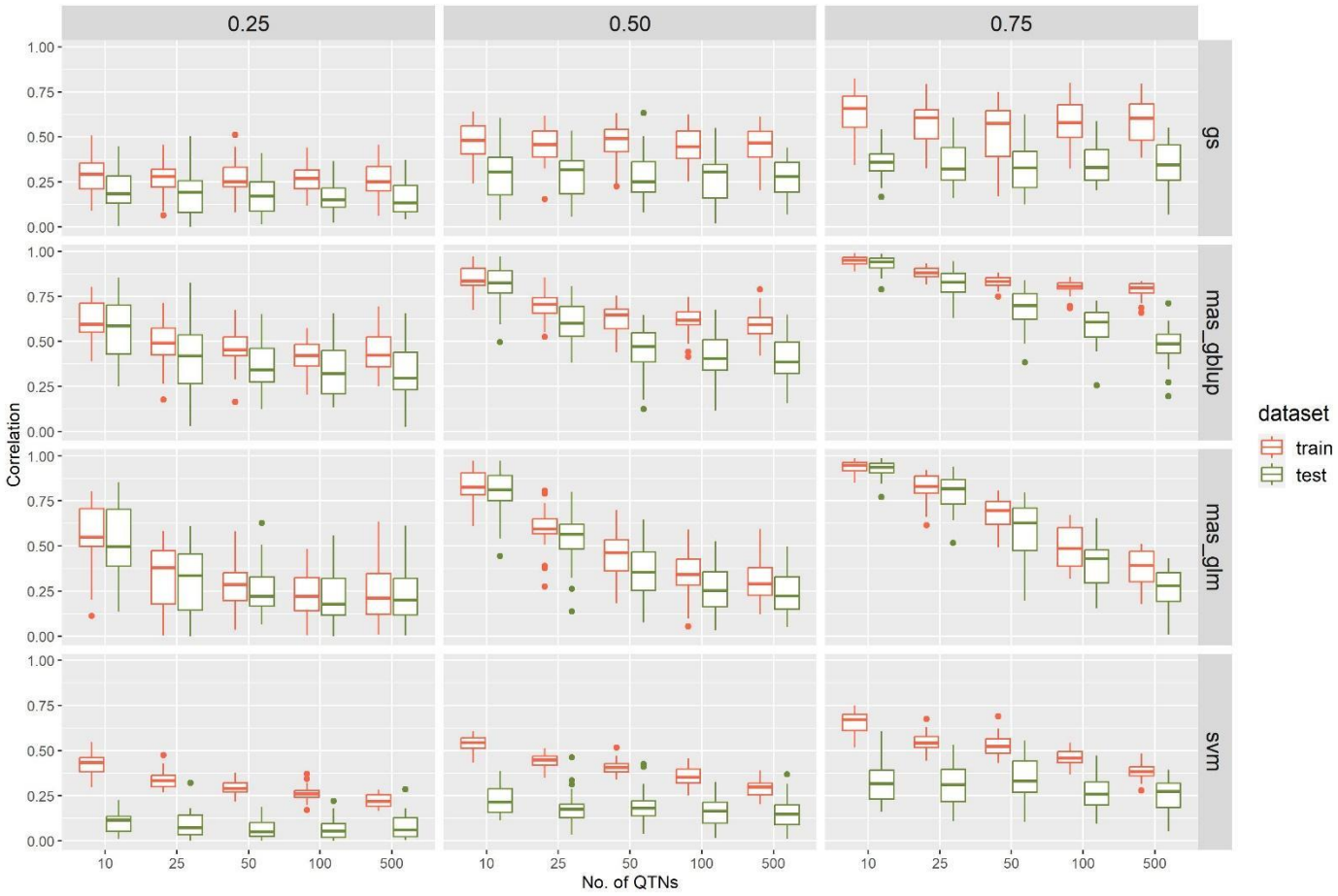

Figure 3: The box plots show the correlation between the phenotype/genetic value and predicted value for various combinations of heritability and number of associated QTN in both the training and test datasets over 30 replications. The x-axis represents the number of QTN, with five levels (10, 25, 50, 100, and 500) and the y-axis represents correlation values. The red and blue colors of the boxplots indicate the training and test sets, respectively, and each horizontal grid represents a genomic selection method, while the vertical grid represents the heritability.

Discussion and Conclusion:

Recent advances in genetics and molecular biology have enabled better understanding of desirable traits in crops. Genomic Selection (GS) is a promising tool for breeders, particularly for polygenic traits. This case study evaluated the impact of different levels of heritability and the number of associated QTN on genomic selection and prediction using low-resolution SNPs of 400 barley lines and tested the performance of four different methods for genomic prediction and selection in the test dataset. The results revealed that highly heritable traits with a low number of QTN had better accuracy, while low heritable and polygenic traits had higher accuracy using the genomic selection method with all markers used to build the kinship matrix and inclusion of all markers in the fixed effect.

The results showed that heritability has a significant impact on the accuracy of genetic selection and prediction. The highest accuracy was observed at a heritability of 0.75, followed by 0.5 and 0.25. Conversely, the lowest overall accuracy was obtained for a heritability of 0.25. Similarly, the number of QTN also has a significant impact on the accuracy of genetic selection. As the number of QTN increases from 10 to 500, the correlation between genetic selection and predicted value decreases. The decreasing correlation suggests that the accuracy of genetic selection is highly dependent on the number of QTN used, and polygenic traits have lower prediction accuracy compared to Mendelian traits.

All methods had higher accuracies for highly heritable traits that were controlled by a small number of genes. Genomic selection using all marker associations had higher accuracy for low heritable traits compared to other methods. For highly heritable traits, the GLM method also exhibited high accuracy. Overall, heritability and the number of QTNs underlying the traits tend to influence the overall correlation coefficients.

The possible limitation of the study includes the lower marker resolution, which may not fully represent the barley genome's size, which is estimated at 5 Gb. To improve the accuracy of prediction using machine learning method, generating high SNP density and exploring hyperparameter tuning, such as kernel functions like rbf and polynomial, could be considered. Other machine learning techniques like random forest and artificial neural networks could also be potentially explored. However, it is important to note that the interpretability and overfitting problems can limit the accuracy of machine learning-based methods.

In this study, the whole genotype information was used to run the machine learning model, but GWAS-assisted machine learning methods could be a potential area for future exploration to predict phenotypes. Furthermore, GWAS- assisted genetic value calculation by building the kinship with just the significant markers could provide better accuracy and is the potential area I could have explored. Plant breeders can exploit the advancements in genetics, statistics, and breeding to increase productivity and address issues such as climate change and depleting genetic resources. Future studies should also consider the potential impact of epigenetics and environmental factors on phenotype changes. Plant scientists should continue exploring and innovating to improve food production and sustainably feed the planet. In conclusion, this study highlights the impact of heritability and the number of QTN on the accuracy of genomic prediction and selection for low-resolution genomic data used for barley lines. Plant breeders should consider these factors when using genomic selection for breeding programs.

GitHub repo link:

https://github.com/gcsudha/StatisticalGenomics/tree/main/Script

References:

Bernardo, R., 2002. Breeding for quantitative traits in plants (Vol. 1, p. 369). Woodbury: Stemma press.

Bernardo, R., 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop science, 48(5), pp.1649-1664.

Bhat, J.A., Ali, S., Salgotra, R.K., Mir, Z.A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P.K. and Singh, G.P., 2016. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. Frontiers in genetics, 7, p.221.

de Moraes, B.F.X., dos Santos, R.F., de Lima, B.M., Aguiar, A.M., Missiaggia, A.A., da Costa Dias, D., Rezende, G.D.P.S., Gonçalves, F.M.A., Acosta, J.J., Kirst, M. and Resende, M.F., 2018. Genomic selection prediction models comparing sequence capture and SNP array genotyping methods. Molecular Breeding, 38, pp.1-14.

Desta, Z.A. and Ortiz, R., 2014. Genomic selection: genome-wide prediction in plant improvement. Trends in plant science, 19(9), pp.592-601.

Huang, M., Liu, X., Zhou, Y., Summers, R.M. and Zhang, Z., 2019. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience, 8(2), p.giy154.

McGowan, M., Wang, J., Dong, H., Liu, X., Jia, Y., Wang, X., Iwata, H., Li, Y., Lipka, A.E. and Zhang, Z., 2021. Ideas in genomic selection with the potential to transform plant molecular breeding: a review. Plant breeding reviews, 45, pp.273-319.

Meuwissen, T.H., Hayes, B.J. and Goddard, M., 2001. Prediction of total genetic value using genome-wide dense marker maps. genetics, 157(4), pp.1819-1829.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C., Lin, C.C. and Meyer, M.D., 2019. Package 'e1071'. The R Journal.

Poland, J. and Rutkoski, J., 2016. Advances and challenges in genomic selection for disease resistance. Annual review of phytopathology, 54, pp.79-98.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., Atlin, G., Jannink, J.L. and McCouch, S.R., 2015. Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS genetics, 11(2), p.e1004982.

Wang, J. and Zhang, Z., 2021. GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genomics, proteomics & bioinformatics, 19(4), pp.629-640.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M. and Buckler, E.S., 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, *42*(4), pp.355-360.