EFFICIENT MARKER SELECTION STRATEGY IN GENOMIC SELECTION

by

DEEPAK VITRAKOTI

(Under the Direction of Paul Schliekelman)

ABSTRACT

Genomic selection is a powerful tool for accelerating genetic gain in plant breeding by leveraging genome-wide markers. In this study, we evaluated genomic prediction accuracy across multiple scenarios using both simulated and real datasets. Our findings demonstrate that prediction accuracy was greater when phenotypes were strongly associated with genotypes and when markers were selected based on GWAS significance rather than randomly. Models containing a greater number of true QTLs consistently yielded higher prediction accuracies, especially at higher heritability levels, whereas the inclusion of non-causal (noise) markers reduced accuracy by diluting the true genetic signal. Moreover, we showed that q-value-based marker selection effectively optimized prediction models, with intermediate q-value thresholds (0.1, 0.2) capturing nearly all true QTLs while minimizing the inclusion of non-informative markers. Validation with real data mirrored the simulation trends, and a QTL recovery analysis confirmed the reliability of this strategy, highlighting the importance of precise and informed marker selection.

INDEX WORDS: Genomic selection, GWAS, Prediction Power, q-value, Marker pre-selection

EFFICIENT MARKER SELECTION STRATEGY IN GENOMIC SELECTION

by

DEEPAK VITRAKOTI

B.S., Tribhuvan University, 2015

M.S., University of Georgia, 2019

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2025

© 2025

Deepak Vitrakoti

All Rights Reserved

EFFICIENT MARKER SELECTION STRATEGY IN GENOMIC SELECTION

by

DEEPAK VITRAKOTI

Major Professor:	Paul Schliekelman
Committee:	Paul Severns
	XianYan Chen

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2025

DEDICATION

I would like to dedicate this thesis to my parents, my wife, and my lovely son, Aarav.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my major professor, Dr. Paul Schliekelman, for his steady guidance and constant encouragement throughout my M.S. in Statistics. His support was instrumental in helping me complete this important milestone. I am equally thankful to my committee members, Dr. Paul Severns and Dr. XianYan Chen, for their thoughtful feedback and invaluable advice throughout the course of this research.

I am profoundly grateful to my Ph.D. advisor, Dr. Andrew H. Paterson, for his unwavering mentorship, which has shaped me into an independent and confident researcher. His belief in my potential gave me the opportunity to immerse myself in every phase of the research process—from field experiments and laboratory work to advanced data analysis and interpretation. His encouragement also inspired me to pursue rigorous training in bioinformatics and statistics, allowing me to build a strong foundation in quantitative genetics and earn both a Graduate Certificate in Bioinformatics and an M.S. in Statistics alongside my Ph.D. in Plant Breeding and Genetics.

Finally, I extend my deepest gratitude to my family for their unconditional love, patience, and countless sacrifices. To my wife—your unwavering support, strength, and belief in me have been the foundation of this journey. You have been my anchor through every challenge, and I could not have accomplished this without you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Quantitative traits are typically controlled by multiple genes with small effects, thus complicating their selection and improvement [1] . Marker-assisted selection (MAS) is a breeding approach for trait improvement, which works well for simple, or qualitative traits controlled by a few genes with larger effects [2]. However, it struggles with complex traits that are controlled by multiple small effect loci [3]. To overcome this, Meuwissen, Hayes [4] introduced the genomic selection (GS) approach, which harnesses genome-wide marker data to estimate the genetic value of individuals without requiring phenotypic measurements for every generation, thereby accelerating selection cycles and improving genetic gain for complex quantitative traits. Heffner, Lorenz [5] reported that GS resulted in two-to three-times higher genetic gain per year in winter wheat and maize compared to MAS.

GS utilizes a training data consisting of individuals genotyped and phenotyped for the trait of interest. This training data is used to build a statistical model that captures the relationship between genome-wide marker information and the observed trait values. This model can then be used to predict phenotypic values for new individuals from genotype data only [4]. The predicted values for the phenotypes of a new set are called genomic estimated breeding values (GEBVs). GEBVs represent the expected genetic merit of individuals based solely on their marker profiles. The accuracy of a GS model can be evaluated using cross validation, where the dataset is split into training and testing subsets. The model is trained on the training data and then used to predict the phenotypes of the remaining individuals in the test data. Prediction accuracy is quantified as the

Pearson correlation coefficient between the predicted GEBVs and the observed phenotypic values in the test set. Several factors influence the accuracy of genomic selectio, including but not limited to trait heritability, underlying QTLs, training population size, marker density, and the statistical model used to estimate marker effects [6].

Genomic selection is based on assumption that quantitative trait loci (QTLs) or genes associated with traits are linked with at least one DNA marker. The linkage disequilibrium (LD) between QTLs and markers has a major effect on prediction accuracy because it determines how well markers can tag causal loci (QTL or genes) [4]. Previously, several studies have reported higher prediction accuracy when markers are in strong LD or even coinciding with causal mutations [7]. Hayes and Daetwyler [8] suggested that the inclusion of the markers tightly linked with a few large-effect QTLs, detected by GWAS could increase the prediction accuracy. Several studies have reported increases in prediction accuracy through large-effect SNP association in fixed-effects models [9-11].

A genome wide association study (GWAS) is a statistical method used to identify the markers significantly associated with a trait of interest. It is considered a powerful tool to dissect the genetic architecture of complex traits in plants and animals [12, 13]. Several studies have reported that marker preselection based on GWAS results can lead to slight or moderate improvement in prediction accuracy [14-17]. Ideally, genomic selection should capture the effects of all QTLs through LD between markers and causal variants. When at least one or two markers in strong LD with each QTL are included, the model can effectively account for majority of the QTL effects. With the advancement of high-throughput genotyping, generating hundreds of thousands of markers has become increasingly feasible. While this offers dense genome coverage, it also introduces many markers that are not in LD with any QTLs. These unlinked markers do not

contribute useful information for trait prediction. Instead, they can introduce statistical noise, inflate model complexity, and increase the risk of overfitting, especially in the models that do not shrink marker effects aggressively. This leads to a situation where the true signal from causal loci is diluted, thereby lowering the prediction accuracy. Therefore, marker preselection, such as that based on GWAS, can enhance the model efficiency and prediction accuracy by identifying markers in strong LD to put in the model while excluding redundant or uninformative ones.

In this study, we investigated how marker selection strategies influence genomic prediction accuracy in plant breeding, with the goal of identifying an optimal number of markers for inclusion in the genomic selection models. We began with a set of simulation-based scenarios, designed to evaluate genomic prediction under varying degrees of genotype-phenotype association. We applied the same analytical pipeline to the true genotype and phenotype data from an F_2 sorghum population to evaluate the model performance under actual breeding conditions. In each scenario, GWAS was conducted to compute p-values for each marker, which were then ranked based on their p-values and subsets of varying sizes were selected to perform genomic prediction. For comparison, we also tested randomly selected marker subsets of equivalent sizes. Additionally, we applied a q-value based marker selection approach to examine how different false discovery rate (FDR) thresholds affects prediction accuracy. We further evaluated the effect of trait architecture such as heritability and number of underlying QTLs on prediction performance. Together, this study provides insights into a how genetic architecture and marker selection strategies affect genomic prediction accuracy, offering practical guidance for optimizing prediction pipelines in plant breeding applications.

CHAPTER 2

MATERIALS AND METHODS

2.1 Theoretical Framework and Hypothesis

We hypothesized that the prediction accuracy in genomic selection can be improved by selecting statistically significant markers identified through genome-wide association studies (GWAS). To test this, we performed GWAS using single marker linear regression, where each marker was independently tested for its association with the phenotype. The p-values were used to rank markers based on statistical significance. The markers with N lowest p-values were used in genomic selection. We apply several strategies, detailed below, for determining N. For comparison, we also applied random marker selection, where N markers were selected at random, independent of GWAS results. This comparison allowed us to assess the effect of informed vs uninformed marker selection on genomic prediction accuracy.

2.2 Genomic Prediction Models

For genomic prediction, we implemented four commonly used Bayesian based models as follows [15]:

Genomic Best Linear Unbiased Prediction (GBLUP)

GBLUP is based on the standard linear mixed model framework, expressed as:

$$y_i = \mu + g_i + \varepsilon_i,$$

where y_i is the phenotypic value of individual I, μ is the overall population mean (a fixed effect), g_i is the genomic estimated breeding value (GEBV), and ε_i is the residual error. The model assumes $g \sim N(0, G\sigma_g^2)$, where $G = XX'/p$ is the genomic relationship matrix construction from

the standardized marker matrix X (coded as 0, 1, 2), and p is the number of markers. The model assumes all marker effects contribute equally to trait variation, making it well-suited for polygenic traits controlled by many small-effect QTLs. It is also effective in scenarios with strong population structure.

Bayesian Ridge Regression (BRR)

In BRR, phenotypes are fitted as a linear combination of all markers in the model as follows: $y_i = \mu + \sum_{j=1}^{p} x_{ij}\,\beta_j + \epsilon_i$, where x_{ij} is the genotype of the individual i at marker j, and $\beta_j \sim N\left(0, \sigma_\beta^2\right)$ is the marker effect, and ε_i is the residual error assumed to be normally distributed. The model assumes the normal distribution of effects which leads to shrinkage of estimates toward zero. It is particularly useful when the number of markers (p) exceeds the number of individuals (n), which is common in genomic datasets. This model is also suitable for polygenic traits where many markers contribute small effects.

Bayesian LASSO (BL)

BL assumes the marker effects follow a double-exponential (Laplace) prior $\beta_j \sim DE\ (\lambda^2, \sigma_\beta^2)$,which can be represented as a scale mixture of normal with an exponential prior in the variance:

$$\beta_j \sim \int N\left(\beta_j \middle| 0, \sigma_j^2\right) Exp\left(\sigma_j^2 \middle| \frac{\lambda^2}{2}\right) \partial\sigma_j^2$$

, where β_j is the effect of the jth marker, $N\left(\beta_j \middle| 0, \sigma_j^2\right)$ is a normal distribution centered at zero with variance σ_j^2. This model uses adaptive shrinkage, allowing each marker to have its own amount of shrinkage. In other words, it applies stronger shrinkage to small marker effects, effectively reducing noise, while retaining large-effect markers. As a result, it produces sparser solutions and is particularly well-suited for scenarios where only a subset of markers is truly associated with the trait.

BayesB (BB)

BayesB assumes a two-part mixture prior of each marker effect β_j. In particular, the model assumes only a subset of markers have non-zero effect, while the rest have exactly zero effect. The marker effects are assumed to be zero with probability π, and with probability $1-\pi$, they followed a scaled Students' t- distribution as follows:

$$\beta_j = \begin{cases} 0, & with\ probaiblity\ \pi \\ t(\beta_j | df_\beta, S_\beta), & with\ probaibility\ 1-\pi \end{cases}$$

The non-zero effects follow as scaled Student's t-distribution with degrees of freedom df_β and scale S_β. The t-distribution can be represented as follows:

$$\beta_j \sim \int N(\beta_j | 0, \sigma_j^2)\ \chi^{-2}(\sigma_j^2 | df_\beta, S_\beta) \partial \sigma_j^2)$$

The model is particularly well-suited for oligogenic traits, controlled by a few major QTLs. It effectively excludes non-informative markers by assigning them an effect size of exactly zero. For GBLUP, we computed genomic relationship matrix using the standardized genotype matrix coded as 0, 1, 2. For BRR, BL, and BayesB, the marker matrix was directly used as the design matrix.

We used a 5-fold cross-validation procedure, repeated over 100 replications, to evaluate genomic prediction accuracy at various marker subsets determined by GWAS-significance. The data were split into five folds; in each fold, four folds was used for training and one for testing. GWAS was performed only on the training set to select markers based on q-value thresholds. The selected markers were then used to train a model and predict phenotypes in test set. Prediction accuracy was calculated as the correlation between predicted and observed phenotype and averaged across replicates. All four models were implemented using the BGLR package in R [15].

2.3 Data

The genotype and phenotype data used in this study were derived from an F_2 sorghum population consisting of 189 individuals (Dr. Paterson's lab (Plant Genome Mapping L) at UGA). A total of 28,958 raw SNP data were generated based on genotyping by sequencing (GBS). We conducted filtering to retain only biallelic markers, reducing the number to 11,028. Markers with missing or heterozygous parental genotypes were removed, resulting in 8,465 SNPs. Of these, 4220 were polymorphic between parents. Finally, markers with more than 20% missing data were excluded, yielding a final set of 3,583 high-quality SNPs for all downstream GWAS and genomic prediction analyses.

2.4 Simulation Framework

2.4.1 Phenotype and Genotype Simulation

We performed several simulations for generating genotypes and phenotypes for conducting GWAS and GS. In some simulation scenarios the genotypes were sampled with replacement from true genotypes in the data. In others, frequencies were generated using the standard F_2 allele frequency (A:H: B = 1:2:1). The phenotype is simulated using the formula:

$$y_i = \sum_{j=1}^{n_{QTL}} g_{ij}\beta_j + \epsilon_i$$

Where y_i is the phenotype of individual *i*, n_{QTL} is the number of QTLs affecting the trait, g_{ij} is the genotype of individual *i* at $QTLj$ (coded as 0, 1, or 2), β_j is the effect of $QTLj$ on the phenotype and ϵ_i is the residual error for individual *i* (normally distributed: $N(0, \sigma_\epsilon^2)$}. To relate this to heritability h^2, we use $h^2 = \left(\sigma_g^2\right)/\left(\sigma_g^2 + \sigma_\epsilon^2\right)$, where σ_g^2 is genetic variance due to QTL effects, and σ_ϵ^2 residual variance. We tested the following scenarios with different combinations of

genotypes and phenotypes to investigate the effect of marker selection strategy in the genomic selection.

2.4.2 Simulation Scenarios

We evaluated four scenarios to investigate the impact of genotype-phenotype relationships on the prediction accuracies as detailed below. In each scenarios, we used 189 samples and 3,583 markers to match the dimensions of the true phenotype and genotype datasets.

Scenario 1. Simulated Phenotype and Simulated Genotype (Type I: True association between phenotype and genotype)

In this scenario, the genotype data was initially simulated based on an allele frequency distribution of 1:2:1, reflecting segregation patterns typical of an F_2 population and provide a realistic genetic framework for the study. This simulated genotype was then used to generate phenotype with a fixed narrow-sense heritability of 0.5, controlled by 50 QTLs with constant-effect sizes. GWAS and genomic selection were then conducted using simulated phenotypes and simulated genotypes, ensuring that the phenotype and genotype remained entirely dependent. This scenario served as a benchmark for evaluating the prediction accuracy when genetic effects are fully known and controlled.

Scenario 2. Simulated Phenotype and Simulated Genotype (Type II: No relationship between phenotype and genotype)

In this scenario, both genotype and phenotype were initially simulated as in scenario 1. Subsequently, a second simulation was performed using the real genotype dataset, where a new set of genotypes was generated by sampling individuals with replacement. Simulated phenotypes were then paired with these resampled genotypes in such a way that the phenotype and genotype

remained entirely independent and unassociated. GWAS and genomic selection were performed under the assumption that there was no intrinsic association between genotype and phenotype, allowing for a controlled evaluation of statistical models and prediction accuracy in the absence of true genetic effects. This scenario also served as a negative control to assess how prediction accuracy behaves when no true signal is present.

Scenario 3. Simulated Phenotype and True Genotype Data

In this scenario, marker genotypes were sampled with replacement from the sorghum data. Each individual in the simulation sample was assigned the full marker genotype vector of an individual in the real data. This procedure maintains a realistic correlation structure between markers. Markers were randomly chosen to be the QTLs and phenotypes were generated using the 'simplePhenotype' package in R [18]. The simulated phenotype was controlled by 50 QTLs with constant-effect sizes and had a narrow-sense heritability of 0.5. The model assumed no dominance or epistatic interactions and residuals followed a normal distribution. This scenario allowed for the evaluation of genomic prediction accuracy under a more realistic genetic background where true genotypic structure was preserved but phenotypes were still under controlled, simulated architecture.

Scenario 4. True phenotype and Genotype Data

In this scenario, we utilized the actual genotype and phenotype data from the F_2 sorghum population for both GWAS and genomic selection. This real-data analysis served as a validation step, enabling us to assess how the results from simulation-based scenarios compare to an actual breeding dataset.

2.5 Effect of Underlying QTLs and Heritability on Prediction Accuracy

To assess how prediction accuracy is influenced by underlying QTLs and heritability, we simulated phenotypes randomly selecting 10, 100, and 500 QTLs from the simulated genotype data (sampling with replacement of true genotypic data) at different heritability values of 0.2, 0.4, 0.6, and 0.8. Although we implemented four genomic prediction models, we focused on the GBLUP model for downstream analyses, as no substantial differences in prediction accuracy were observed among the models across Scenarios 1-4. For each simulation, GWAS was performed first to rank markers based on their p-values. Genomic selection was then performed using successive subsets of the top 500, 1000, 1500, 2000, 2500, 3000, and 3500 markers. Prediction accuracy was computed for different combinations of QTLs and heritability values across different subsets of top-ranked markers, which was measured as the Pearson correlation coefficient between the genomic estimated breeding values (GEBVs) and the true simulated phenotypes in the test set.

2.6 QTL vs Noise Marker Simulation

To evaluate the effect of errors or noise markers related to the prediction accuracy, we randomly selected a fixed number of QTLs from the true genotype for simulating the phenotype. The noise makers were generated from the remaining markers from a true genotype that does not contain any QTL genotype. However, there are still some chances that some markers may show poor to moderate association and cause biases. To mitigate this, we ran a GWAS with non-QTL genotypes and sorted them based on their p-values. We selected the top 1000 non-significant markers (1000 highest p-values) as noise markers. The prediction models were then fitted to predict genomic accuracies for several combinations of QTLs and noise marker combinations to assess the effect of non-causal marker inclusion. The scenarios included combinations such as (25 QTLs, 0 noise), (25QTLs, 25 noise), (25 QTLs, 50 noise), (25 QTLs, 75 noise), and (25 QTLs, 100 noise) markers.

The scenarios were tested with increased numbers of causal loci, including 50, 75, and 100 QTLs for the four different narrow-sense heritability levels (0.2, 0.4, 0.6, 0.8).

2.7 q-value Based Marker Selection

We used the q value approach to identify statistically significant markers while controlling the false discovery rate (FDR) in the genome-wide association study. The q-values were computed from the GWAS-derived p-values using the method proposed by Storey and Tibshirani (2003)[19], which estimates the minimum FDR at which a particular markers can be considered significant. Unlike p-values, which assess the probability of observing a test statistics as extreme as the one obtained under the null hypothesis, q-values account for the multiple testing burden by estimating the expected proportion of false positives among all the markers declared significant.

We used the 'qvalue' package in R[20] to calculate the q-values and generate marker subsets at multiple FDR thresholds, ranging from very stringent cutoffs (0.001 – 0.05) , through moderate thresholds (0.1 – 0.3), up to baseline thresholds of (0.7 – 1.0). These marker sets were then used in genome-wide association studies and genomic prediction models to evaluate how different stringency of marker selection thresholds affect the prediction accuracy. We used both simulated and real datasets to evaluate the impact of q-value-based marker selection approach on prediction accuracy. Additionally, we compared the effect on prediction accuracy when GWAS was performed on full datasets prior to prediction, compared to that when GWAS was performed on training data only. This comparison enabled us to assess the potential inflation of accuracy when marker selection is not independent of the test set.

2.8 Validation with True Dataset and QTL Recovery Analysis

To further evaluate the effectiveness of q-value-based marker selection approach for capturing the genetic signals, we performed a QTL recovery analysis using the true dataset. A total

of 152 significant SNPs were identified using GAPIT package in R [21], and these were considered as the true QTLs controlling the trait for the purpose of recovery analysis. At each q-value cutoffs (0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0) we counted the number and proportion of true QTLs that were recovered among different marker subsets. This analysis allowed use to understand how well different levels of statistical stringency performed to retrieve the true QTLs and relate QTL recovery rates to trends in prediction accuracy.

CHAPTER 3

RESULTS

3.1 Marker Selection Strategies: GWAS-Ranked vs Random

3.1.1 Scenario 1: Simulated phenotype and simulated genotype (Type I)

In this scenario, the simulated phenotype is entirely dependent on the simulated genotype, representing the ideal case where there is a true underlying genetic basis for the trait, as in the case of true phenotype and true genotype. When markers were selected based on GWAS-rankings, the highest prediction accuracy was observed with small subsets of highly significant markers (i.e., 500 to 1000 top-ranked markers). However, as the number of selected markers increased, prediction accuracy gradually declined, likely due to the inclusion of less informative or non-causal markers that may dilute the signal from true QTLs (Figure 1a).

In contrast, when markers were selected at random, prediction accuracy showed a slight increase as the size of the marker subset grew (Figure 1b). This could be due to increased probability of capturing true QTLs as more markers are included. Notably, the higher prediction accuracy observed under the BayesB model in the random marker selection strategies aligns well with its underlying assumption of retaining the effects of large-effect markers while shrinking the effects of small or non-informative markers towards zero.

3.1.2 Scenario 2. Simulated phenotype and simulated genotype (Type II)

In this scenario, we assumed that there was no correlation, only a very weak correlation by chance, between phenotypes and genotypes. As expected, genomic prediction yielded zero or near-zero accuracy, regardless of whether markers were selected randomly or based on GWAS rankings.

This outcome confirms the absence of any true association between genotype and phenotype and serves as a negative control, validating that our prediction models do not capture spurious signals when no real genetic effects are present.

3.1.3 Scenario 3: Simulated phenotype and true genotype.

In this scenario, where phenotypes were simulated based on the true genotype matrix, the prediction accuracy was generally high across all models, reflecting the strong underlying genetic control. As shown in Figure 3a, when top-ranked markers were used, the prediction accuracy gradually declined with increase in the size of marker subset. However, when markers were selected at random, accuracy remained stable over different marker subsets (Figure 3b).

3.1.4 Scenario 4. True phenotype and genotype data

When real genotype and phenotype data was used, the prediction accuracy declined as the marker subset size increased as observed in simulation performed in Scenario 1 and 3 (Figure 4a), however, the overall accuracy levels were lower than in the simulated scenario, reflecting the complexity of the trait architecture and environmental noise present in real data. For random marker subsets (Figure 4b), accuracy remained relatively stable across all subset sizes.

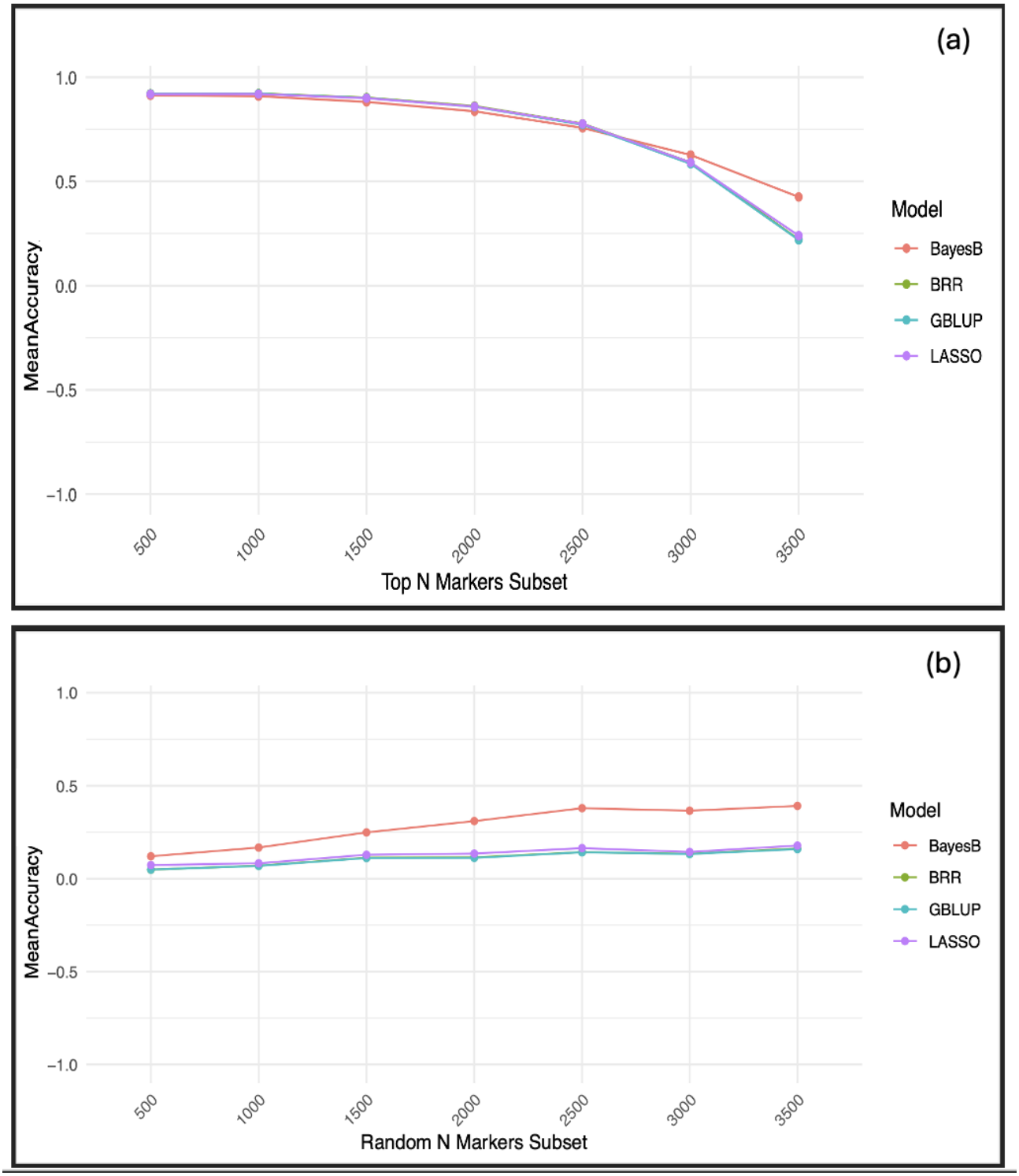

Figure 1. Prediction accuracy across different models using simulated phenotype and simulated genotype (Type I). Phenotype and genotype are entirely dependent. (a) when markers are sorted based on p-values, (b) when markers are selected randomly.

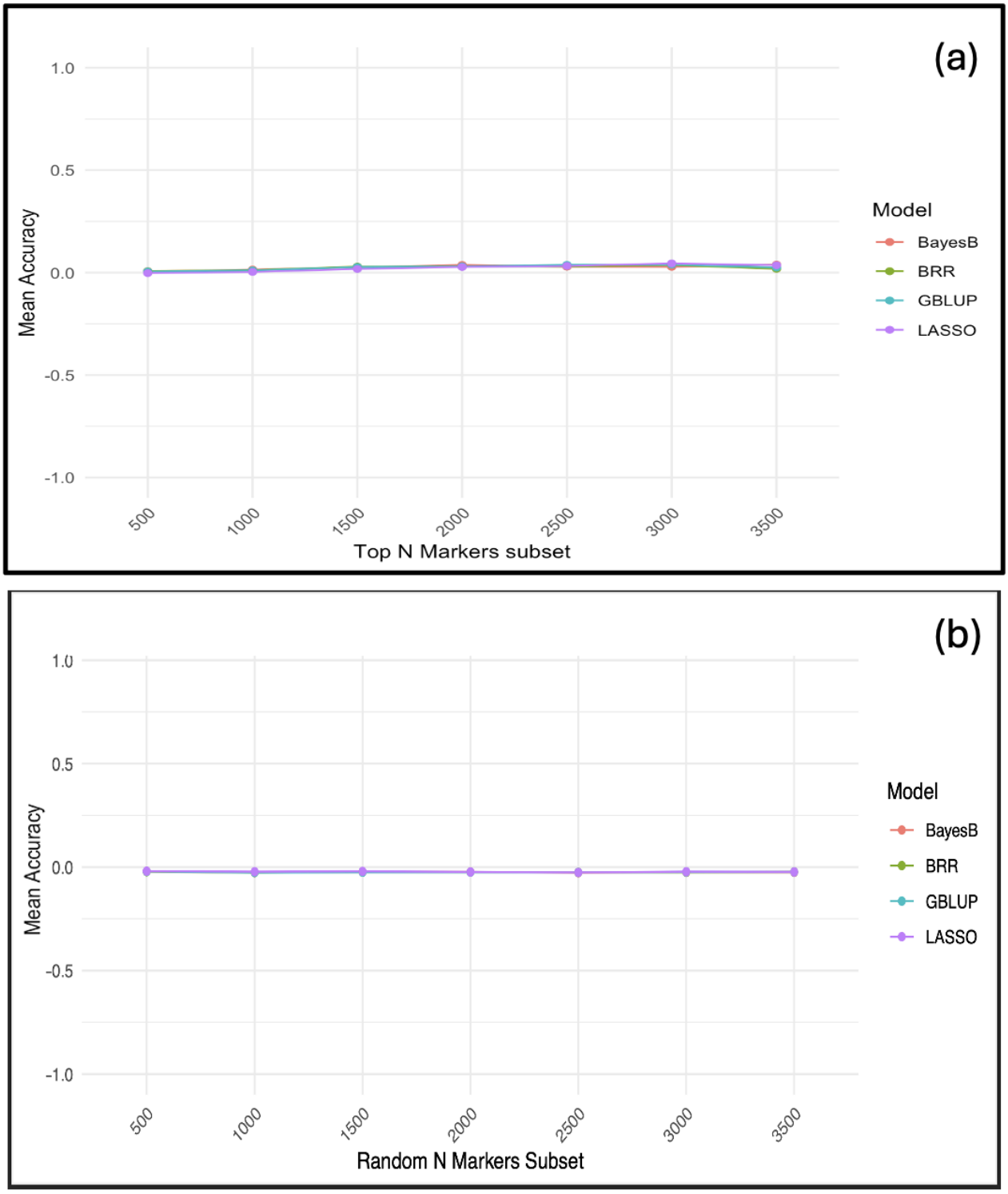

Figure 2. Prediction accuracy across different models using simulated phenotype and simulated genotype (Type II). Phenotype is independent of genotype. (a) when markers are sorted based on p-values, (b) when markers are selected randomly.

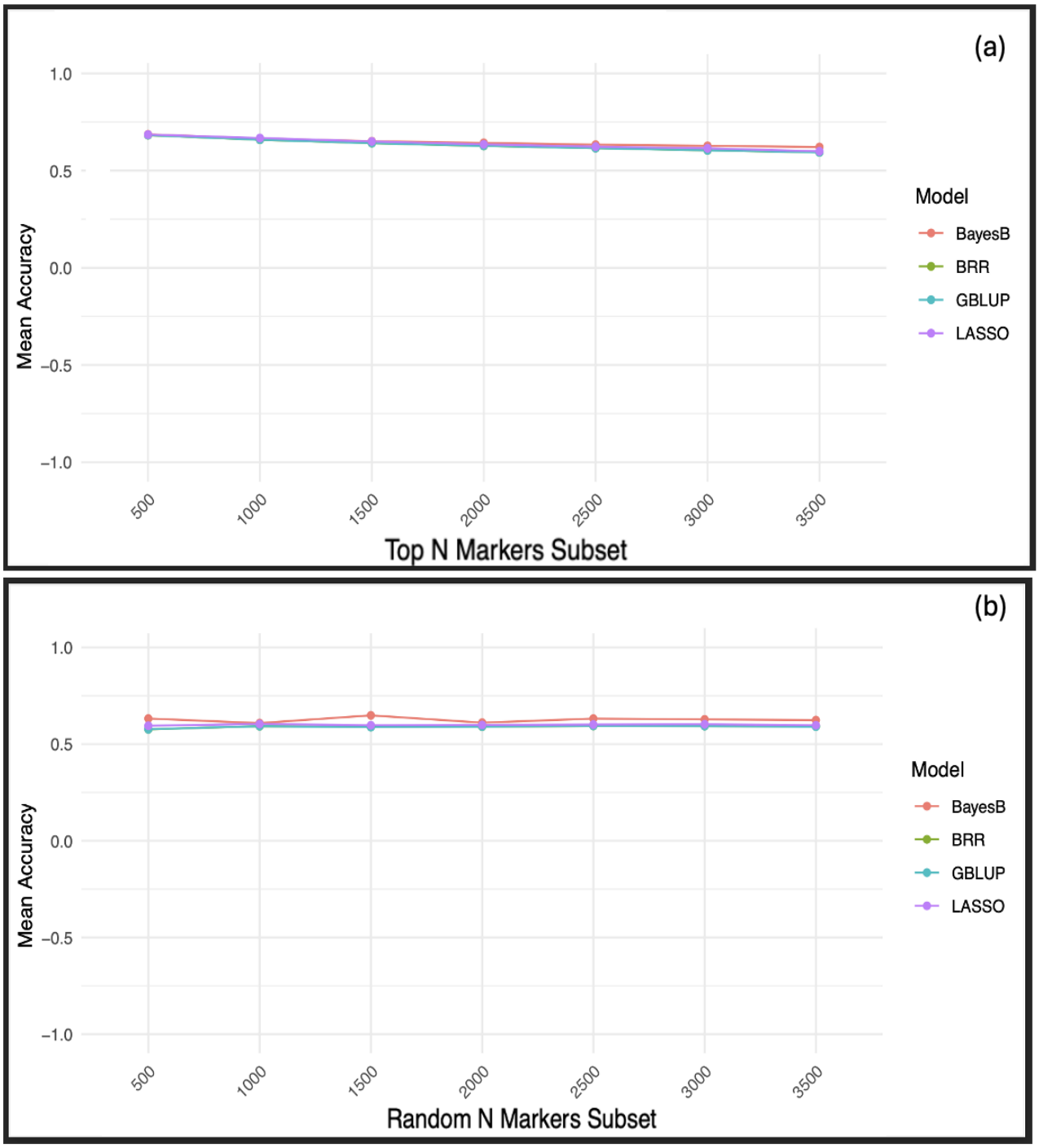

Figure 3. Prediction accuracy across different models using simulated phenotype and true genotype. (a) when markers are sorted based on p-values, (b) when markers are selected randomly.

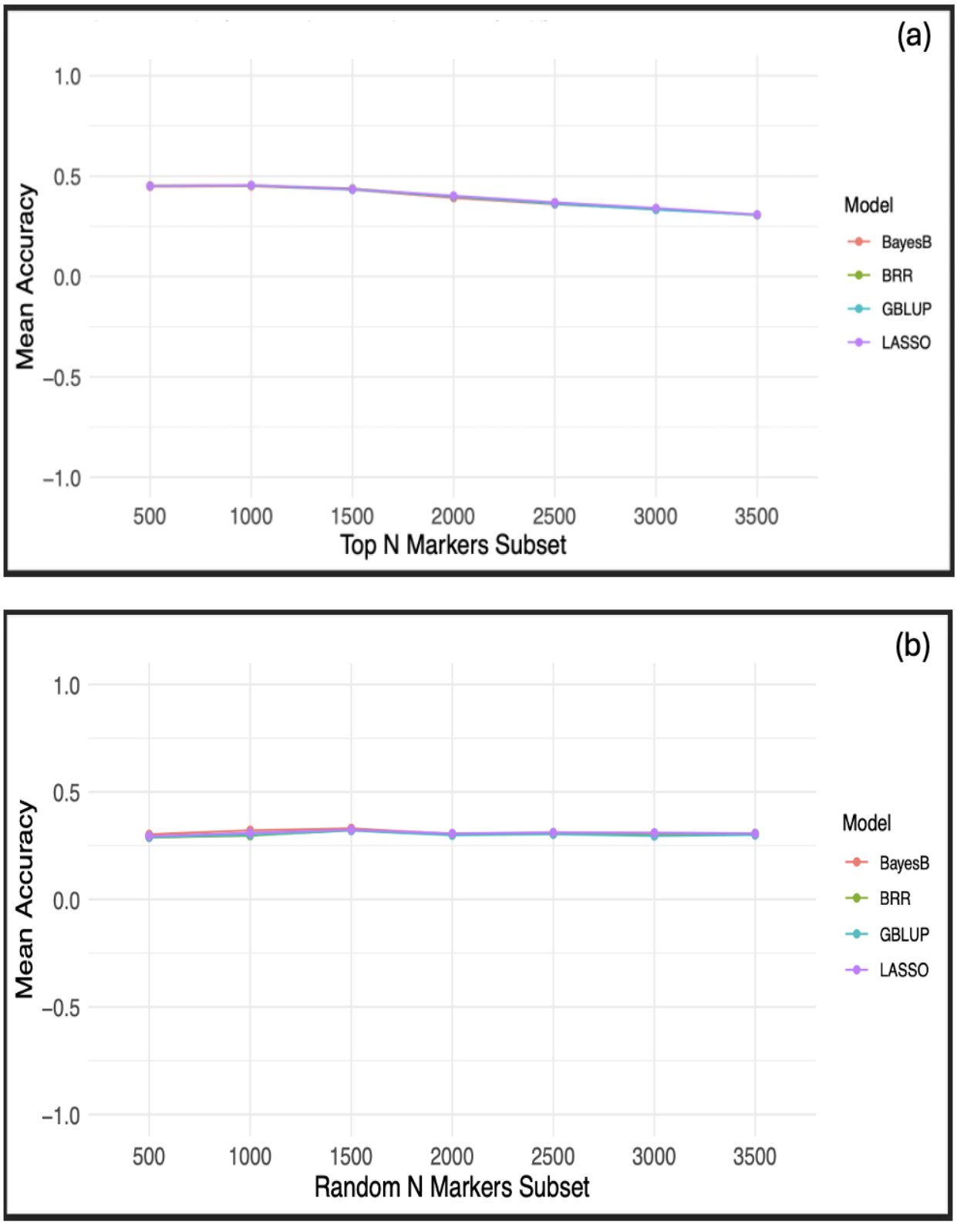

Figure 4. Prediction accuracy across different models using true datasets. (a) when markers are sorted based on p-values, (b) when markers are selected randomly.

3.2 Effect of Underlying QTLs and Heritability on Prediction Accuracy

Next, we explored how the relationship between prediction accuracy and number of markers varied with the heritability and the number of QTL controlling the trait. We observed little to no difference in prediction accuracy at low heritability of 0.2, regardless of the number of QTLs controlling the trait. However, as heritability increased, traits controlled by few QTLs such as 10 QTLs (large-effect QTLs) consistently exhibited higher prediction accuracy compared to those controlled by many small effect QTLs. This suggests that at higher heritability values, traits governed by fewer major QTLs are more predictable, highlighting the advantage of large-effect QTLs in genomic prediction models.

With 10 QTLs, prediction accuracy was always highest with 500 markers selected and decreased as the number of selected markers increased. In contrast, with 100 and 500 QTLs, prediction accuracy peaked at 1000-1500 selected markers for higher (0.6 and 0.8) heritabilities. This is likely because as the number of QTL increases it takes more markers being selected to get the same number of QTLs in the selected group. However, it is not clear why this same pattern doesn't hold for lower heritabilities.

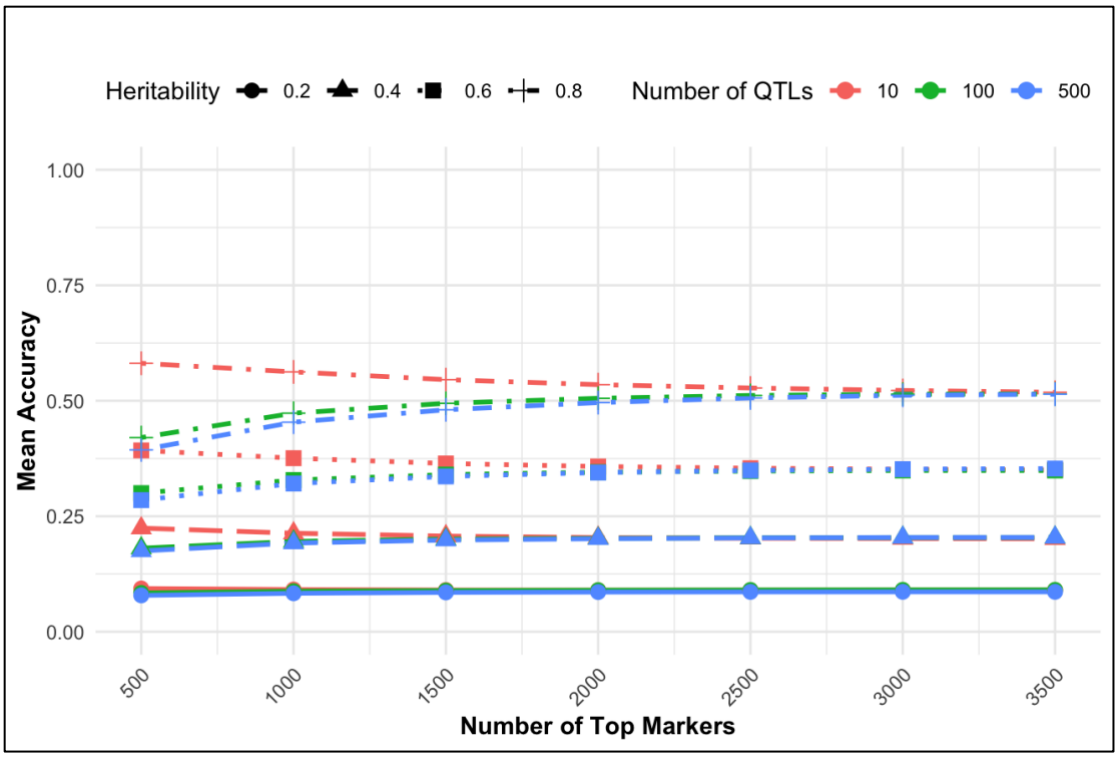

Figure 5. Prediction accuracy across different combination of heritability values and number of QTLs.

3.3 Effect of Noise markers on Prediction Accuracy

In order to better understand the how the balance of true QTLs versus noise markers in the prediction model impacts prediction accuracy, we conducted simulations in which the number of QTLs and noise markers was explicitly specified. In contrast, for the previous simulations the number of noise markers versus true QTLs was not specified and arose naturally from the marker ranking process.

We simulated scenarios in which 25, 50, 75, and all 100 QTLs were included in the model and cases with 0, 25, 50, 75, and 100 noise markers of equal variance. As expected, the prediction accuracy increases with increased number of true QTLs in the model and fewer noise markers. We see that getting a higher proportion of QTLs is generally more important than getting fewer noise

markers, up to a point. For example, the model with 75 QTL and 100 noise markers has equal or higher accuracy than all of the models with 25 or 50 QTLs, regardless of the number of noise markers.

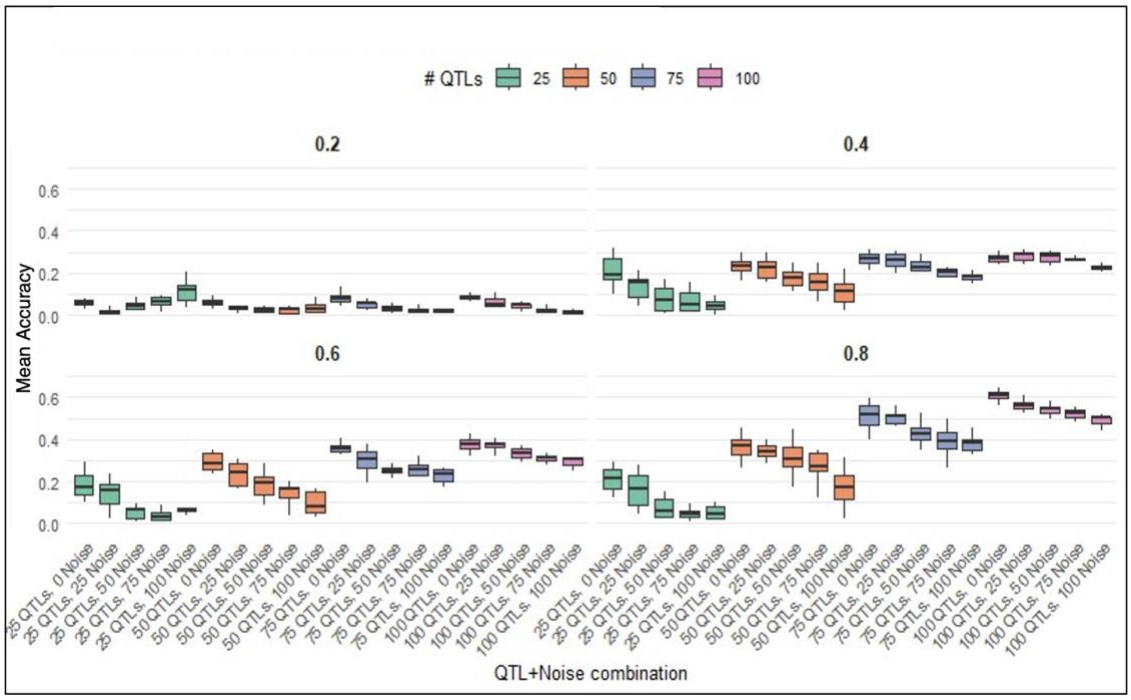

Figure 6. Prediction accuracy across different combination of heritability values and number of QTLs including noise markers in the model

3.4 q-value Cutoffs and Marker Selection Efficiency

When simulated phenotypes based on simulated genotypes were used to assess the prediction accuracy across different q-value cutoffs, we observed a sharp increase in accuracy as the q-value threshold increased from 0.001 to 0.2 (GWAS on training data, Figure 7). Beyond this threshold, prediction accuracy remained stable despite the inclusion of more markers (Figure 7, Table 1). In contrast, when phenotypes were simulated independently of genotypes (i.e., using noise markers), prediction accuracies remained consistently near zero across all q-value thresholds, as expected (Figure 8). This is because the GWAS generated p-values from unrelated phenotype and genotype dataset had uniform-distribution (Figure 9, left panel). The flat distribution is characteristics of the

null-scenario, where no true genotype-phenotype association exists. When these p-values were converted into q-values (Figure 9, right panel), nearly all q-values were clustered near 1.0, indicating a complete lack of statistically significant markers. This outcome is consistent with expectations under the null hypothesis and confirms that the q-value method effectively controls the false discovery rate and does not detect spurious associations when genotype and phenotype were unrelated.

An important observation was the difference in prediction accuracy depending on whether GWAS was conducted on the full dataset or only within the training data during cross-validation (Figure 7). When GWAS was conducted on the full dataset prior to prediction, accuracy was slightly inflated. In contrast, when GWAS was conducted solely within training folds before marker selection, the prediction accuracy remained lower but more realistic. These finding highlight the importance of maintaining independence between marker selection and predictions steps to avoid overestimating genomic prediction performance.

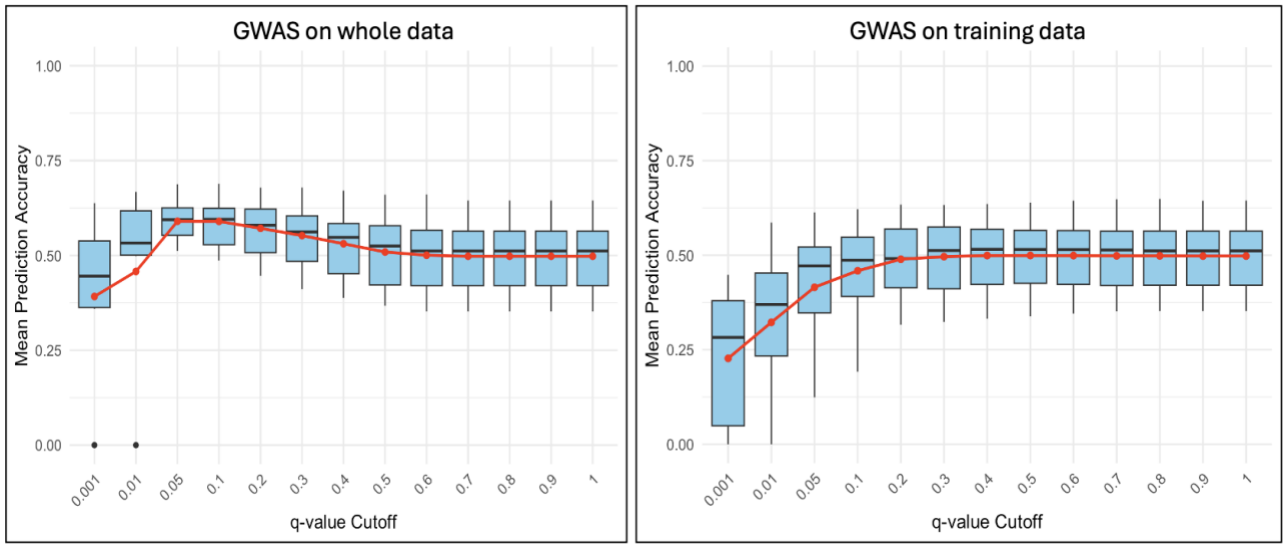

Figure 7. Prediction accuracy using markers selected based on q-values on simulated dataset

Table1: QTLs recovered at different q-value cutoffs using simulated dataset

q-value	Number of Marker	No of QTLs recovered	Proportion of QTLs recovered	Average Accuracy
0.001	33.3	4.1	0.08	0.29
0.01	133.6	8.1	0.16	0.41
0.05	408.5	16.2	0.32	0.49
0.1	641.4	20.8	0.42	0.52
0.2	1005.2	28.7	0.57	0.53
0.3	1396.1	36.8	0.74	0.53
0.4	1824	43.2	0.86	0.52
0.5	2324.5	48.2	0.96	0.53
0.6	2827.2	49.7	0.99	0.53
0.7	3240.6	50	1.00	0.53
0.8	3349.8	50	1.00	0.53
0.9	3526.5	50	1.00	0.53
1	3583	50	1.00	0.53

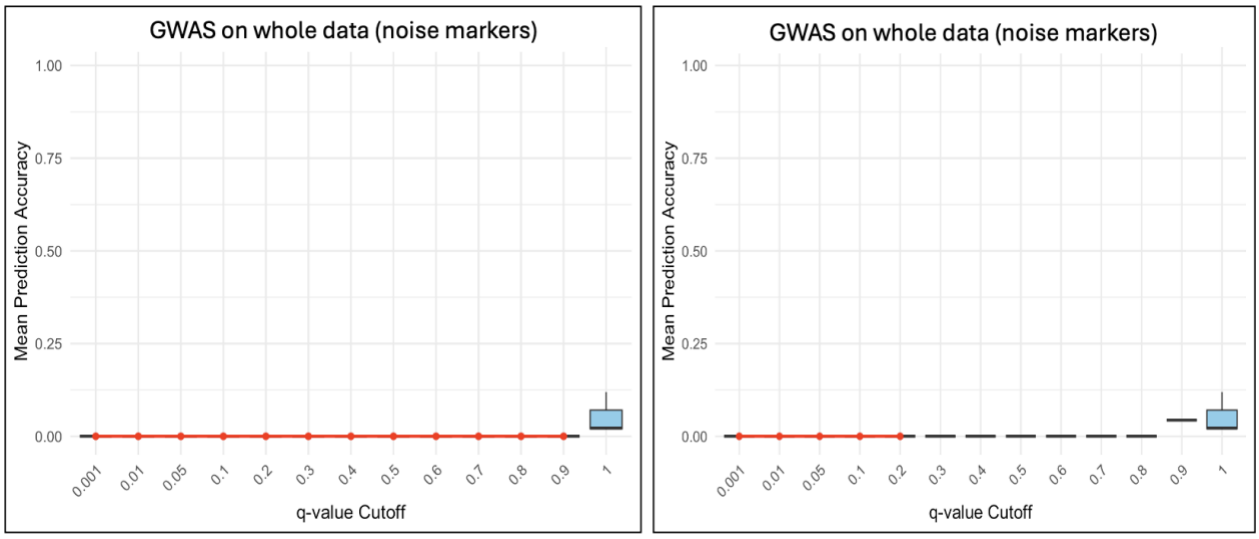

Figure 8. Prediction accuracy using markers selected based on q-values (model with noise markers)

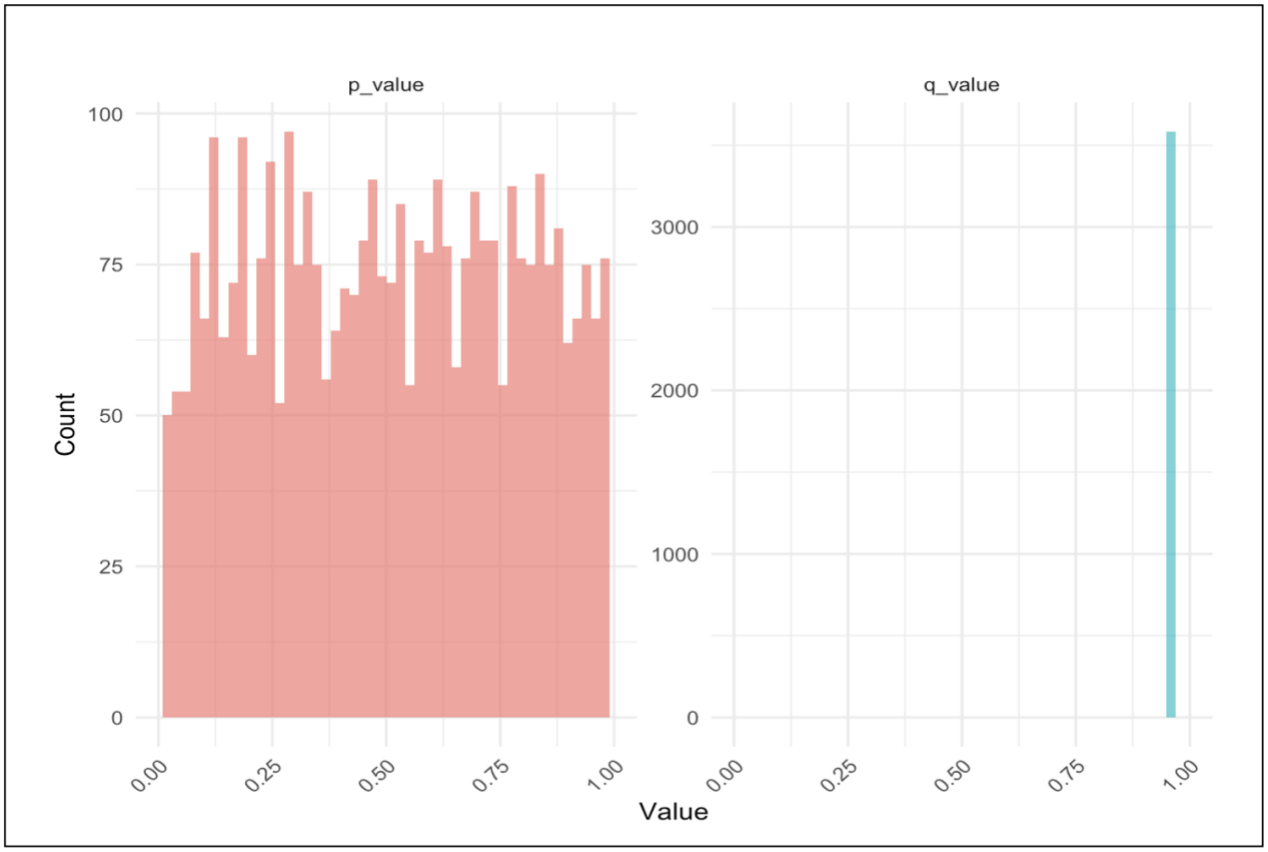

Figure 9. Distribution of p-values and q-values in model with noise markers.

3.5 **Validation with Real data: q-value thresholds and QTL Recovery**

When the true phenotype and genotype dataset was used, the prediction accuracy across q-value cutoffs followed the similar pattern as in the simulated scenario where the phenotype was fully dependent on genotype. As shown in the Figure 10, prediction accuracy sharply increased as the q-value threshold was relaxed from 0.001 to 0.2, beyond which accuracy plateaued. At very stringent thresholds such as 0.001, prediction power was reduced due to missing many true QTLs. However, at moderate q-value thresholds such as 0.1 or 0.2, the most informative markers for prediction were captured. Adding more markers beyond that point did not contribute additional predictive power, likely due to the inclusion of non-informative markers.

We validated the results by assessing the count and proportion of true QTLs recovered at each q-value threshold for the trait. We considered all 152 significant SNPs identified via GWAS using GAPIT as true QTLs for the trait. At very stringent threshold of 0.001, the marker subset included ~43% of true QTLs, indicating many true QTLs were missed, reducing the prediction power. However, at intermediate threshold such as 0.1 or 0.2, nearly 100% of true QTLs were recovered (Table 2). Once all the true QTLs were recovered, additional markers did not increase prediction accuracy, yet they may introduce noise into the prediction model (Figure 10, left panel). These findings confirm that q-value-based marker selection is effective in prioritizing informative markers and optimizing prediction accuracy, especially when the phenotype is strongly influenced by underlying genetic variation. The consistency between real and simulated data results also validates the robustness of this approach under realistic genetic architecture.

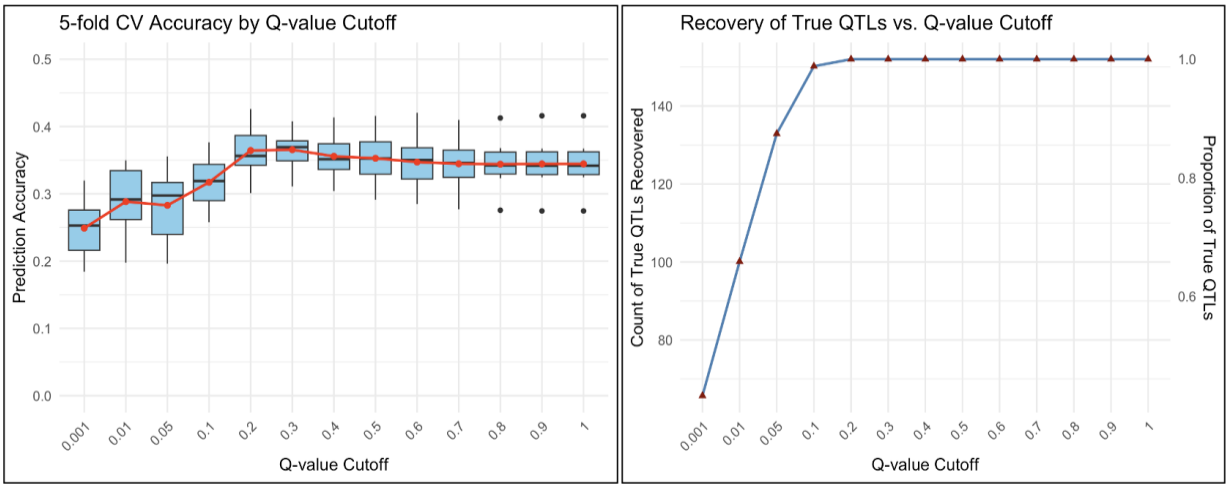

Figure 10. Prediction accuracy and QTL recovery using markers selected based on q-values on true data set

Table 2: QTLs recovered at different q-value cutoffs using true dataset

q-value	Number of Marker	No of QTLs recovered	Proportion of QTLs recovered	Average Accuracy
0.001	66.9	65.7	0.43	0.25
0.01	113.4	100.1	0.66	0.29
0.05	223	132.9	0.87	0.29
0.1	376.8	150.2	0.99	0.32
0.2	803.3	152	1	0.36
0.3	1502.8	152	1	0.36
0.4	2448.6	152	1	0.35
0.5	3262.7	152	1	0.35
0.6	3569.4	152	1	0.35
0.7	3583	152	1	0.34
0.8	3583	152	1	0.34
0.9	3583	152	1	0.34
1	3583	152	1	0.34

CHAPTER 4

DISCUSSION

In this study, we presented a comprehensive evaluation of genomic prediction accuracy using several simulated and real data scenarios, focusing on the impact of marker selection strategies and statistical thresholds. A key findings across all scenarios was that prediction accuracy was highest when phenotypes had a strong association with genotypes and when markers were selected based on GWAS significance rather than at random. This aligns with prior studies suggesting that incorporating informative or trait-associated markers improves prediction performance [11, 22]. Specifically, in Scenario 1, where phenotype was entirely dependent on genotype, accuracy peaked with smaller subsets of top-ranked markers and declined as more non-informative markers were added, likely due to the dilution of true genetic signal by non-causal variants. This observation supports the theory that, in highly polygenic traits, prediction performance can be weakened by overfitting or the introduction of irrelevant features [23]. However, when markers were selected randomly, accuracy increased slightly with larger subsets. This could be due to increased probability of including true QTLs as the marker subset size increased.

Scenario 2 provided a negative control, where phenotype was simulated independently of genotype. As expected, the prediction models showed near zero accuracy regardless of marker selection strategy. This outcome confirms the statistical validity of the models as well as support the notion that strong genetic signal as a pre-requisite for meaningful prediction.

To further validate the prediction pattern observed in Scenario 1, we incorporated two additional scenarios: Scenario 3, where phenotype was simulated using the true genotype matrix,

and Scenario 4, which utilized real observed phenotype and genotype data from a sorghum F_2 population. In both cases, we observed higher prediction accuracy with smaller marker subsets derived from GWAS-ranked p-values, which was consistent with the simulated scenario (Scenario 1) and highlights the potential of genomic prediction in real breeding contexts [24]. Random selection of markers, however, demonstrated stable, but generally lower prediction accuracy, reinforcing the advantage of informed marker selection.

We also evaluated how prediction accuracy is affected by the genetic architecture of traits, namely the number of QTLs controlling the traits and trait heritability. Prediction accuracy was substantially higher for traits governed by a small number of large-effect QTLs, particularly at higher heritability levels. These results aligns with theoretical expectations and empirical findings that traits with simpler, more additive architectures tend to yield higher prediction accuracies [25, 26]. On the other hand, the inclusion of non-informative (noise markers) in the model consistently reduced prediction performance regardless of the trait architecture, highlighting the importance of marker quality over quantity in building effective models.

Another major focus of our study was evaluating the utility of q-value-based marker selection. The q-value, as proposed by Storey and Tibshirani (2003) [19], controls the false discovery rate (FDR), and provides a way to balance true discovery and Type I error in multiple testing scenarios as in genome-wide association studies. We found that using q-values to filter markers at moderate thresholds (i.e., 0.1 to 0.2) led to optimal prediction accuracy, corresponding to the recovery of nearly all true QTLs as evident in the true datasets. Importantly, at more stringent thresholds such as 0.001, the model excluded many true positives, while at more relaxed threshold, the model added non-informative markers (noise), confirming the importance of an intermediate q-value cutoff for practical use. These results are consistent with prior work suggesting that

stringent threshold reduces power, while liberal thresholds increase the false positives [19, 27]. The findings from our study demonstrate that similar or even higher prediction accuracy can be achieved using fewer but more informative markers. This approach offers several advantages, including and an increased the signal to noise ratio, reduced computational burden, and lower genotyping costs–ultimately making the genomic prediction process more efficient and cost-effective.

Furthermore, our study showed that performing GWAS on the full dataset prior to marker selection, resulted in slightly inflated prediction accuracy, likely due to data leakage that violates the independence assumption in cross-validation. In contrast, GWAS solely on the training data, preserved the statistical integrity and yielded more realistic estimated of prediction accuracy, thus reinforces the best practices in genomic selection pipelines, where marker selection must remain independent of the test data to avoid biased results.

The effectiveness of q-value based marker selection was further validated by QTL recovery analysis using true datasets. We utilized 152 significant SNPs identified through GAPIT as a proxy for true QTLs. We found that only ~43% of true QTLs were recovered at a q-value threshold of 0.001, while 100% were recovered at 0.2. Beyond this point, adding more markers failed to improve prediction accuracy, suggesting that q-value based selection is not only statistically robust, but also biologically meaningful in prioritizing markers that influences trait variation.

In summary, our findings demonstrate that q-value–based marker selection approach is an effective and reliable method for optimizing genomic prediction. The consistency of results across simulated and real datasets supports the robustness of this framework for use in real-world practical breeding applications.

CHAPTER 5

CONCLUSION

In this study, we evaluated genomic prediction accuracy across multiple simulated scenarios and real datasets, focusing on the effects of marker selection strategies and statistical thresholds. Our results consistently showed that prediction accuracy was highest when phenotypes had a strong genetic basis and when markers were selected based on GWAS significance rather than at random. Marker selection using q-values at moderate thresholds (e.g., 0.1– 0.2) was especially effective since they recovered most true QTLs and avoided the inclusion of noise or irrelevant markers, thereby optimizing prediction accuracy.

Additionally, we found that the genetic architecture of traits strongly influenced the prediction accuracy. Traits with few, large-effect QTLs and high heritability were more predictable than highly polygenic traits or those with low heritability. Importantly, conducting GWAS on the training set rather than the full dataset avoided data leakage and provided unbiased estimates of model performance, reinforcing best practices for genomic selection pipelines. Real dataset from sorghum F_2 populations confirmed the patterns observed in simulations, highlighting the practical relevance of these findings. Overall, our study demonstrates that q-value–based marker selection is an effective strategy for enhancing genomic prediction in plant breeding.

References

1. Lynch, M. and B. Walsh, *Genetics and analysis of quantitative traits*. Vol. 1. 1998: Sinauer Sunderland, MA.

2. Collard, B.C. and D.J. Mackill, *Marker-assisted selection: an approach for precision plant breeding in the twenty-first century.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2008. **363**(1491): p. 557-572.

3. Podlich, D.W., C.R. Winkler, and M. Cooper, *Mapping as you go: An effective approach for marker-assisted selection of complex traits.* Crop Science, 2004. **44**(5): p. 1560-1571.

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

5. Heffner, E.L., et al., *Plant breeding with genomic selection: gain per unit time and cost.* Crop science, 2010. **50**(5): p. 1681-1690.

6. Combs, E. and R. Bernardo, *Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers.* The Plant Genome, 2013. **6**(1): p. plantgenome2012.11.0030.

7. Meuwissen, T., B. Hayes, and M. Goddard, *Accelerating improvement of livestock with genomic selection.* Annu. Rev. Anim. Biosci., 2013. **1**(1): p. 221-237.

8. Hayes, B.J. and H.D. Daetwyler, *1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes.* Annual review of animal biosciences, 2019. **7**(1): p. 89-102.

9. Bernardo, R., *Genomewide selection when major genes are known.* Crop Science, 2014. **54**(1): p. 68-75.

10. Bian, Y. and J. Holland, *Enhancing genomic prediction with genome-wide association studies in multiparental maize populations.* Heredity, 2017. **118**(6): p. 585-593.

11. Sehgal, D., et al., *Incorporating genome-wide association mapping results into genomic prediction models for grain yield and yield stability in CIMMYT spring bread wheat.* Frontiers in plant science, 2020. **11**: p. 197.

12. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proceedings of the National Academy of Sciences, 2009. **106**(23): p. 9362-9367.

13. Atwell, S., et al., *Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines.* Nature, 2010. **465**(7298): p. 627-631.

14. Chen, Z.-Q., et al., *Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce.* BMC genomics, 2018. **19**: p. 1-16.

15. Tan, B., et al., *Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F 1 hybrids.* BMC plant biology, 2017. **17**: p. 1-15.

16. Thumma, B.R., K.R. Joyce, and A. Jacobs, *Genomic studies with preselected markers reveal dominance effects influencing growth traits in Eucalyptus nitens.* G3, 2022. **12**(1): p. jkab363.

17. Resende Jr, M., et al., *Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.).* Genetics, 2012. **190**(4): p. 1503-1510.

18. Fernandes, S.B. and A.E. Lipka, *simplePHENOTYPES: SIMulation of pleiotropic, linked and epistatic phenotypes.* BMC bioinformatics, 2020. **21**: p. 1-10.

19. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proceedings of the National Academy of Sciences, 2003. **100**(16): p. 9440-9445.

20. Storey, J.D., et al., *qvalue: Q-value estimation for false discovery rate control.* R package version, 2015. **2**(0): p. 10-18129.

21. Lipka, A.E., et al., *GAPIT: genome association and prediction integrated tool.* Bioinformatics, 2012. **28**(18): p. 2397-2399.

22. Spindel, J. and H. Iwata, *Genomic selection in rice breeding.* Rice genomics, genetics and breeding, 2018: p. 473-496.

23. De Los Campos, G., et al., *Predicting quantitative traits with regression models for dense molecular markers and pedigree.* Genetics, 2009. **182**(1): p. 375-385.

24. e Sousa, M.B., et al., *Increasing accuracy and reducing costs of genomic prediction by marker selection*. Euphytica, 2019. **215**: p. 1-14.

25. Daetwyler, H.D., et al., *The impact of genetic architecture on genome-wide evaluation methods*. Genetics, 2010. **185**(3): p. 1021-1031.

26. Wientjes, Y.C., R.F. Veerkamp, and M.P. Calus, *The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction*. Genetics, 2013. **193**(2): p. 621-631.

27. Stephens, M., *False discovery rates: a new deal*. Biostatistics, 2017. **18**(2): p. 275-294.