

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Rethinking Prediction and Evaluation in Plant Breeding: Integrating Statistical Methods and Phenomic Data for Genetic Gain

Permalink

<https://escholarship.org/uc/item/6d66w7qf>

ISBN

9798189233128

Author

Wang, Fangyi

Publication Date

2026-06-25

Peer reviewed|Thesis/dissertation

Rethinking Prediction and Evaluation in Plant Breeding: Integrating Statistical Methods and
Phenomic Data for Genetic Gain

By

FANGYI WANG
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Daniel Runcie, Chair

Mitchell Feldmann

C. Titus Brown

Committee in Charge

2026

To my grandfather, 周敬飞.

Contents

Abstract	v
Acknowledgments	vi
Introduction	1
Chapter 1. Why Usefulness Is Rarely Useful	4
1.1. Abstract	4
1.2. Introduction	4
1.3. Methods and Materials	7
1.4. Results and Discussion	11
1.5. Conclusions	21
1.6. Data Availability	22
1.7. Acknowledgement	22
1.8. Funding	22
Chapter 2. Do Not Benchmark Phenomic Prediction against Genomic Prediction Accuracy	23
2.1. Abstract	23
2.2. Introduction	23
2.3. Materials and Methods	26
2.4. Results and Theory	30
2.5. Discussion	36
2.6. Conclusions	40
2.7. Data Availability	40
2.8. Acknowledgements	40
Chapter 3. Evaluating the Performance of a Two-step Prediction Model by Incorporating Phenomic Data into Genomic Prediction for Strawberry Fruit Quality Traits	41
3.1. abstract	41

3.2. Introduction	42
3.3. Materials and Methods	44
3.4. Results and Discussion	49
3.5. acknowledgements	55
3.6. conflict of interest	55
Appendix A. Appendix of Chapter 2	56
A.1. Derivation of correction factor for ρ_y	56
A.2. Derivation of ρ_e when PP measures y directly	57
A.3. Derivation of estimating accuracy with bivariate linear mixed model	58
Appendix B. Supplementary Materials of Chapter 1	60
Appendix C. Supplementary Materials of Chapter 3	61
Bibliography	63

Abstract

New statistical methods and data resources have been introduced to improve plant breeding. However, evaluating their utility remains challenging. This dissertation examines how statistical models and emerging data types should be assessed and integrated within quantitative genetics. In the first chapter, I evaluate the benefit of the Usefulness Criterion for parental selection and show that estimating within-family genetic variance provides meaningful benefit only under limited conditions, as variation in predicted progeny mean typically dominates Usefulness Criterion. In the second chapter, I demonstrate that commonly used cross-validation estimates of phenomic prediction accuracy are biased due to environmental correlation between traits of interest and phenomic data, and are not the accuracy parameter in the breeder's equation. In the third chapter, I propose a two-step prediction method that first uses phenomic data to predict phenotypes, which are then incorporated into genomic prediction models, allowing phenomic data to augment rather than replace genetic data. Together, these results show that the benefits of new methods and data sources are highly context-dependent and that their evaluation must be aligned with breeding objectives. This work provides a framework for more principled integration of statistical models and high-dimensional data in plant breeding.

Acknowledgments

The success of graduate students relies heavily on the support of their mentors, family, friends, and community, as has mine.

First and foremost I would like to thank Dr. Dan Runcie and Dr. Mitchell Feldmann. Dan and Mitchell have made my PhD journey a fun and enjoyable experience. I greatly enjoyed our inspiring and stimulating conversations over the years. It is through their guidance that I became a scientist. I would like to thank Dr. Haixiao Hu. He gave me valuable advice and suggestions during the early years of my PhD on my projects and studies. He was always kind and encouraging.

I also thank Forrest Li. He was the one who introduced me to Dan's lab, and he was incredibly supportive and encouraging during my hardest times.

Most importantly, I would like to thank my family for their unwavering support. They have always believed in me and stood by me. It must have been difficult for my parents that their only child studies and lives abroad, but they support my decisions regardless. I can never fully repay their sacrifices.

I also want to thank Dr. Louise Ferguson. She introduced me to the scientific world. Together with Dr. Barbara Blanco-Ulate and Dr. Shuai Chen, she guided me as I began developing as a scientist.

I would also like to thank my friends, cohortmates, and members of the Runcie and the Feldmann labs, including Felina Loya, Kelly Chau, Olivia Ho, Panyue Wang, Paul Skillin, Tiezheng Fan, Yi-Ching Lee. There are many more that I cannot write a comprehensive list. It was reassuring to know that we were all in this together. It was through their support and encouragement that I was able to endure the most difficult times.

Lastly I would like to thank Zackary Tyler. He brings me peace.

Introduction

By the 2080s, the world population is projected to approach 10 billion people (67), placing unprecedented pressure on agriculture to produce more food of higher quality on constrained resources. Feeding this population sustainably will require crop varieties that are high-yielding, nutrient-rich, and resilience to environmental stress (5). To meet this challenge, plant breeders must accelerate genetic improvement, developing superior cultivars at a faster pace and a larger scale, which requires applications of new breeding methods and new forms of data. This dissertation examines how new sources of data and statistical methods should be evaluated and integrated within quantitative genetics, with a focus on their impact on genetic gain.

Quantitative genetics is the study of how continuous traits are influenced by multiple genes and environmental factors in a population. It is a study widely used in the field of plant breeding for plant enhancement and population improvement, since most economically important traits in crops are continuous instead of categorical, and quantitative genetics provides a foundation for artificial selection for crops (7). The core of breeding relies on iterations of choices: which breeding methods to use, which parents to cross, and which offspring to select either for commercial cultivar development or as parents for the next generation of the breeding population.

One of the fundamental theorems of quantitative genetics explains the sources of genetic gain Δg with the Breeder's Equation (53):

$$\Delta g = \frac{i\rho\sigma_A}{L},$$

where i is the selection intensity, ρ is the selection accuracy, σ_A is the additive genetic standard deviation, and L is generation interval. While many modern breeding methods aim to improve one component of this equation, most commonly selection accuracy, these improvements are often evaluated in isolation, without considering their impact on other components or on overall genetic gain. My dissertation focuses on implementing breeding strategies and statistical methods and how they affect the Breeder's Equation as a whole.

(71) proposed the Usefulness Criterion (UC), which includes both within-family mean (μ) and standard deviation (σ), for parental selection. The theory of UC is that the best crosses are the ones that can potentially produce the best possible offspring, instead of the ones that produce the best average offspring. In other words, a cross that can generate a higher tail of a distribution

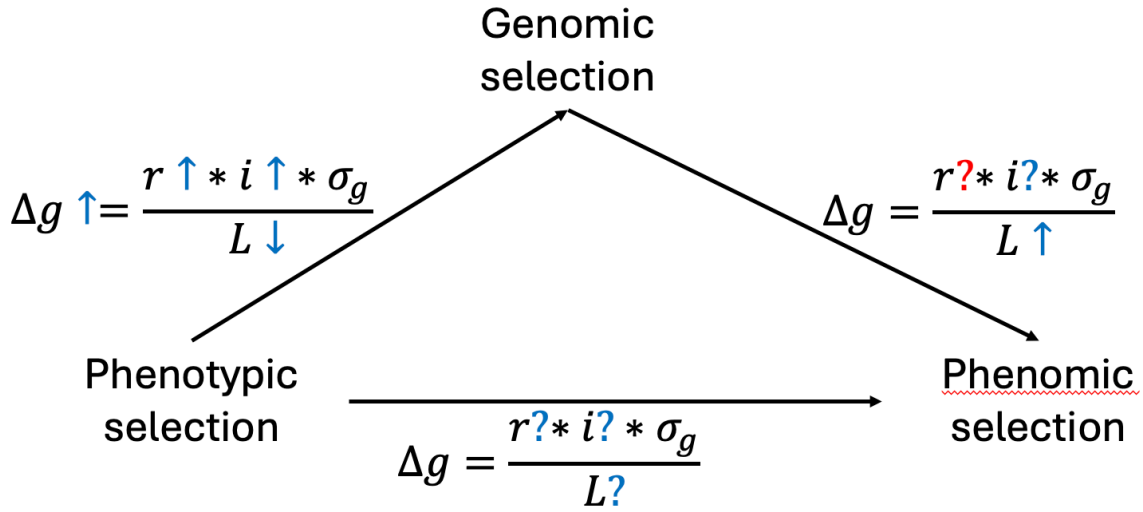


FIGURE 0.1. **Changes of components in Breeder's Equations with different selection methods**

of offspring can produce offspring with higher breeding values, compared to a cross with a similar center but a smaller tail. The UC is extremely attractive in plant breeding, because plants are able to produce large families, which makes producing offspring in the tails more likely. However, I show that the expected family mean often dominates UC, limiting the practical benefit of estimating within-family standard deviation in many realistic breeding scenarios.

Genomic selection has been introduced as a tool to accelerate genetic gain in breeding (55). It models the relationship between genetic markers and phenotypic values. Once a model is built, it can be used to predict future generations' phenotypic values, using their genetic information. Compared to the traditional phenotypic selection, genomic selection increases selection accuracy in Breeder's Equation by using the relationship information among genotypes, increases selection intensity and decreases generation interval. These make genomic selection an extremely attractive method in breeding. Phenomic prediction has garnered much interest (65; 47; 85) as an alternative tool for selection (Figure 0.1). Instead of genetic markers, it utilizes high-throughput phenotyping technologies, for example, hyperspectral image data, near-infrared spectroscopy (NIRS), etc. The advantage of phenomic prediction is that it can be less expensive and less logistically challenging compared to genomic prediction. It also utilizes similar statistical or machine learning methods as

genomic prediction. Many studies have shown that phenomic prediction achieved higher prediction accuracy, making it more desirable than genomic prediction. However, I show that commonly used cross-validation accuracy is not equivalent to the accuracy parameter in the Breeder's Equation and can be biased by environmental structure. Also, I discuss that the change in selection intensity component in the Breeder's Equation depends on the crop and system. If phenomic data is collected to replace phenotypic data, because it is costly and/or time-consuming to phenotype, it can potentially increase selection intensity compared to phenotypic selection. If the cost of phenomic data collection is less than the cost of genotyping, the selection intensity can be greater in phenomic selection compared to genomic selection. The generation interval component likely increases compared to genomic prediction (Figure 0.1) and stays the same compared to phenotypic selection. There are many approaches to utilize phenomic data. Earlier papers did not evaluate these approaches fairly or consistently (65; 47; 85). This is due to the environmental confounding between phenomic data and target traits. After establishing appropriate evaluation metrics, I try to develop effective phenomic selection methods. I propose a two-step framework in which phenomic data are first used to predict phenotypes, and these predicted values are then incorporated into genomic prediction models. This approach treats phenomic data as a form of indirect phenotyping, allowing it to augment rather than replace genomic information while mitigating environmental confounders. Together, this work demonstrates that the value of new methods and data types in breeding is inherently context-dependent, and that meaningful progress requires aligning statistical models, evaluation metrics, and biological objectives. By emphasizing this alignment, this dissertation contributes to a more principled framework for integrating new data sources and statistical methods into plant breeding.

CHAPTER 1

Why Usefulness Is Rarely Useful

1.1. Abstract

Mate selection plays an important role in breeding programs. The Usefulness Criterion was proposed to improve mate selection, combining information on both the mean and standard deviation of the potential offspring of a cross, particularly in clonally propagated species where large family sizes are possible. Predicting the mean value of a cross is generally easier than predicting the standard deviation, especially in outbred species when the linkage of alleles is unknown and phasing is required. In this study, we developed a method for estimating phasing accuracy from unphased genotype data on possible parental lines and evaluated whether the accuracy was sufficient to predict family standard deviations of possible crosses. We used simulations spanning a wide range of genetic architectures and used genotypes from a real strawberry breeding population to evaluate the conditions when usefulness could be accurately predicted. We found that with highly accurate computational phasing, predicting family standard deviations and usefulness criteria for potential crosses yields benefit over simply selecting crosses based on predicted family means only at high selection intensity and high heritability and with small numbers of QTL. However, even then the gain from using the family usefulness is small.

1.2. Introduction

Plant breeding programs aim to create genotypes with improved performance in one or more target traits. New genotypes are created by making crosses between existing genotypes, using recombination and Mendelian assortment to assemble sets of alleles that did not exist. These new genotypes can then be evaluated and released as commercial products or recycled to create new crosses. The success of a breeding program can be improved by evaluating genotypes more efficiently and making crosses that are more likely to generate better genotypes (6; 55; 39; 28). In recent years, genomic data has been leveraged in many breeding programs to make both steps more effective. Genomic

data can help make the evaluation of genotypes based on phenotypic data more precise by modeling the correlations between phenotypes and genotypes. Genomic data can also be used to select parents to be crossed by predicting their breeding value (BV), or the expected average performance of their offspring (55; 39; 42). Both uses of genomic data are now commonly implemented in plant breeding programs.

Predicting the breeding value of potential parents is widely used to choose crossing partners. When the number of offspring per cross is large, the best crosses are the ones that will create the best possible offspring, not the crosses that will make the best average offspring. This situation is common in many plant systems because large numbers of seeds can be generated from the same cross. This result was observed nearly 50 years ago by (71), who proposed the usefulness criterion (UC), or usefulness, to design optimal crosses (89). Usefulness is defined as

$$UC_j = \mu_j + i\sigma_j,$$

where μ_j is the average genetic value of offspring of cross j , σ_j is the standard deviation (SD) of genetic values among these offspring, and i is within family selection intensity, the difference in mean phenotypic value of selected offspring from each family relative to the family's mean. The consequence of the usefulness is that crosses between parents that individually have good breeding values are not necessarily the best to make. For example, the upper tail of the distribution of genetic values of offspring from a cross with high variance (high σ_j^2) may have higher values than the upper tail of the distribution of offspring from a family with higher mean but lower variance. Selection decisions based on usefulness have received little attention in animal breeding, possibly because the number of offspring per cross is very small so the chance of creating offspring in the tails is low. However, several recent plant breeding studies have suggested that usefulness-based breeding might be successful, particularly in clonally propagated crops (86).

A simple method to predict the mean breeding value of a family is to use the average of the breeding value of the two parents. This will be accurate as long as most QTL operate additively. However, even under additive gene actions, predicting the variance within a family is more challenging as patterns of co-segregation of causal loci among genotypes also contribute to variance. For example, a cross between two genotypes with two coupling heterozygous loci would produce a greater variance in the next generation, compared to one with two repulsion heterozygous loci (53; 89; 86). Predicting

co-segregation of causal loci from a cross of two outbred parents requires information on the phasing of alleles between all pairs of heterozygous loci in each parent and phased haplotypes are not directly measured by most genotyping platforms when individuals are heterozygous. Computational phasing using software like Beagle (15; 14) and FILLIN (74) have been used in plant populations (62; 72). However, whether these methods are sufficiently accurate to predict family variance is unknown. Once phased haplotypes are known, QTL effect sizes need to be estimated in a training population, and then either family population variance (89; 57) or sample variance can be predicted (89; 57; 50; 8; 86).

The benefit of choosing crosses based on their usefulness will likely vary among crops and breeding programs due to differences in genetic architecture, breeding system, and training population size. (89) evaluated the benefit of the usefulness in inbred populations by analytical studies and simulations. They found a generally limited utility of usefulness in inbred populations. A recent study by (86) evaluated the usefulness in a cassava breeding program with pre-phased parental individuals and empirical observations of cross-variance in 4 traits. They argued that both family mean and variance were important metrics for mate selection. In their results, the family mean and usefulness criterion were correlated and similar selections were made using either criterion. This was one of the first applications of usefulness in an outbred, clonally propagated species using whole-genome marker data and genomic prediction models of marker effects. Our work focuses on strawberry, another outbred, clonally propagated species, and on the University of California Davis (UCD) strawberry breeding program. This breeding program is currently investing in infrastructure to implement genomics data in breeding decisions (60; 44; 37; 26; 46) and has detailed records of the pedigrees of current varieties (61; 45; 25) which help evaluate the success of the statistical models necessary for predicting usefulness as detailed below.

Our study builds on the work of (89) and (86) in three key ways. First, we develop and apply a method to measure the accuracy of computationally phased parental haplotypes which is necessary for accurately predicting usefulness. Second, our study uses a different reference population with a different pedigree topology and a different marker density. (89) studied the benefit of usefulness in inbred species, while our study focuses on an outcross species. Therefore, our results test whether their conclusions are robust in other systems. Finally, we use numerical simulations of genetic architectures so that we can compare computationally predicted usefulness values to the actual

(simulated) values without measurement error. We repeat these simulations across a range of additive genetic architectures from oligogenic to polygenic and from low to high heritability so that our conclusions can generalize to many possible target traits. Overall, our results show that despite highly accurate computational phasing, the conditions under which selection decisions based on usefulness could be useful are relatively rare. This is consistent with earlier results from (8) who studied predicting genetic variance within biparental populations created from inbred parents and concluded that there was little benefit to modeling within family genetic variance.

1.3. Methods and Materials

1.3.1. Genetic Data and Genotyping Error Estimation. We used genetic and pedigree data from a population of 1007 strawberry genotypes with 37,441 SNPs. We identified 593 family trios based on the pedigree to assess genotyping error rates and pedigree accuracy. Each trio is composed of two parents and one of their progeny. These 593 progenies were not parents of any other genotypes in the pedigree. The parents of the 593 progenies were included in the rest of the 414 genotypes. Within each family trio, we examined SNP markers for which both parents were homozygous. We compared the parental and progeny genotypes at those markers and counted the number of mismatches between parents and progeny. For example, parental genotype AA and aa with a progeny genotype AA were considered a mismatch. The genotyping error was calculated as the number of mismatched markers divided by the total number of homozygous markers compared per trio. All of the family trios had a genotyping error rate of $< 5\%$.

1.3.2. Phasing. To predict the family breeding value variance, the linkage between heterozygous alleles within each parent must be known. Assigning heterozygous genotype calls to haplotypes is phasing. We used Beagle 5.0 (15; 14), version *05May22.33a* and seed 0, to phase the SNP data for all 1007 genotypes and to impute any missing genotype calls. We used a genetic map for the markers with 28 pairs of chromosomes averaging 223 centiMorgans (cM), measured as the average distance between the furthest markers on the chromosomes. We excluded 2,530 of the 37,411 markers due to their unknown map positions or low allele frequencies.

1.3.3. Simulation of offspring. We calculated family mean and standard deviation empirically by simulating crosses. We randomly selected 10,000 unique pairs of strawberry genotypes as parents for the crossing simulation. For each pair of parents, we used the *Hypred* R package (76) to

simulate crossover events and produce 400 gametes using the same genetic map as above, and the gametes were randomly paired to produce 200 offspring. Crossover events were simulated based on a count-location model. The number of crossover events (on a single chromosome) followed a Poisson distribution $Pois(L)$, where L was the length of the chromosome in Morgans. The locations of crossover events were sampled from a uniform distribution $U(0, L)$.

1.3.4. Estimation of imputation and phasing error. To measure the accuracy of phasing and imputation, we re-ran Beagle using the 414 genotypes that were not the progenies of the 593 family trios, and left the 593 progenies unphased. We estimated the imputation error using the same method of estimating the genotyping error as above. We compared genotypes of progeny to their computationally phased parents at pairs of markers for which at least one parent was heterozygous at both markers (e.g. $AaBb$ and $AABB$ parents). We measured the percentage of progeny genotypes that were inconsistent with the predicted parental haplotypes for markers, binned by the cM distance between markers.

We compared the calculated rates of phasing inconsistencies to the expected rates of inconsistencies given our genetic map, by introducing phasing errors in the haplotypes of each parent and then comparing the "incorrect" parent haplotypes to the simulated offspring genotypes using the "correct" parent haplotypes. We estimated the rate of offspring genotype inconsistency with the "incorrect" parental haplotypes using the method above. We defined 6 rates of phasing error per cM distance $p \in (0.005, 0.01, 0.025, 0.05, 0.1, 0.15)$ as the probability of a phasing error between two SNPs 1 cM apart. Phasing errors were introduced using the same method of crossover events as described above. The number of phasing errors per chromosome was sampled from a Poisson distribution $Pois(p * L)$, where L is the chromosome length in Morgans. The locations of phasing errors were sampled from a uniform distribution $U(0, L)$.

1.3.5. Simulation of breeding value and phenotypes. We simulated breeding values assuming a sparse additive genetic architecture. We created 15 scenarios spanning all combinations of three levels of heritability h^2 (0.2, 0.5, 0.8) and five levels of QTL number (4, 16, 64, 256, 1024) randomly selected from the SNP markers with effect sizes sampled from a standard normal distribution $N(0, 1)$. We defined the breeding value (BV) of genotype k as $BV_k = \sum_j Z_{k,j} \alpha_j$, where $Z_{k,j}$ was the number of copies of alternative allele for genotype k at QTL j and α_j was the effect

size of alternative allele at QTL j . To calibrate these architectures, we calculated the number of QTL segregating per pair of haplotypes, which is the average number of heterozygous QTL per individual. For the traits with 4 QTL, on average there were ≈ 1.5 segregating QTL per genotype, while for the traits with 1024 QTL, there were ≈ 400 segregating QTL (Figure: S3). Compared to a biparental population, this is equivalent to having 1.5 or 400 segregating loci. Focusing only on large-effect QTL explaining at least 20% of the variance in breeding values, haplotype pairs differed by ≈ 1.13 large-effect QTL when there were 4 QTL in the whole population, ≈ 3.43 when there were 16 QTL, and ≈ 7.58 , 5.57, and 0.06 with 64 - 1024 QTL (Figure S4).

We also simulated phenotypic values (Y) by adding random noise ϵ_k sampled from a normal distribution $N(0, \sigma_\epsilon^2)$, i.e. $Y_k = BV_k + \epsilon_k$, where σ_ϵ^2 was scaled based on different levels of heritability $h^2 = Var(\mathbf{BV}) / (Var(\mathbf{BV}) + \sigma_\epsilon^2)$. For each combination of heritability and QTL number, we ran 20 independent simulations, sampling different SNPs as QTL and with different effect sizes and environmental errors in each simulation, and we used 500 out of the 10,000 families for each simulation to ensure independence among the 20 simulations. We calculated breeding values for both the real 1007 parental genotypes and simulated offspring, but the phenotypes were only simulated for the real 1007 genotypes.

1.3.6. Prediction of family means and standard deviation of breeding value, and estimation of prediction accuracy. RR-BLUP and BayesC models were fitted to the simulated phenotypes of the 1007 parental genotypes and their SNP marker data, excluding the markers that were selected as QTL (20; 63; 75), i.e. $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\mathbf{Y}$ and $\mathbf{X}$ were the phenotype vector and marker matrix of 1007 genotypes and $\hat{\boldsymbol{\beta}}$ was the trained models. Predicted breeding values $\hat{\mathbf{BV}}_{offspring}$ were calculated for each offspring by plugging offspring genotypes $\mathbf{X}_{offspring}$ into the fitted models, i.e. $\hat{\mathbf{BV}}_{offspring} = \mathbf{X}_{offspring}\hat{\boldsymbol{\beta}}$. We calculated the within family breeding value mean and standard deviation based on both the true and predicted breeding value of offspring. For example, the true and predicted breeding value mean (μ and $\hat{\mu}$) of a family was calculated as $\mu = (\sum_k \mathbf{BV}_k)/200$ and $\hat{\mu} = (\sum_k \hat{\mathbf{BV}}_k)/200$, where k represented the k th offspring in the family. We calculated the prediction accuracy r as the Pearson correlation between the predicted values and the true values of each statistic (e.g. $r = cor(\hat{\mu}, \mu)$).

1.3.7. Prediction of family mean and standard deviations of all possible crosses using simulation of gametes. We used the randomly simulated 10,000 families to assess the prediction accuracies for family mean, standard deviation, and usefulness. For mate selection, however, 10,000 crosses were only a small fraction of all 506,521 possible unique crosses that can be made among the 1007 individuals, and the cross that yielded the highest mean or usefulness was likely not among them. We needed a method to exhaustively predict family means and standard deviations of all possible crosses without the simulation of offspring. For this method, we simulated 200 gametes from each of the 1007 individuals using *Hypred* with the same count-location model. We calculated the predicted breeding values of the gametes of each potential parent as $\hat{B}\hat{V}_{gametes} = \mathbf{X}_{gametes}\hat{\beta}$, where $\hat{\beta}$ was the trained BayesC model. We also calculated the predicted mean and variance of gametes from each parent, $\hat{\mu}$ and $\hat{\sigma}$. Since we only considered additive genetic architectures, we calculated predicted family means and variances of a cross as the summation of predicted means and variances of gametes from two individuals. For example, if a cross was made between individual j and individual l , the predicted family mean and variance would be calculated as $\hat{\mu}_{j,l} = \hat{\mu}_j + \hat{\mu}_l$ and $\hat{\sigma}_{j,l}^2 = \hat{\sigma}_j^2 + \hat{\sigma}_l^2$. Predicted family standard deviation was calculated as the square root of predicted family variance. We used this method to calculate the true and predicted family mean and standard deviations for all 506,521 crosses, using the simulated gametes. The estimates of family mean breeding values were highly correlated using both methods (greater than 99%, Figure S7a). The correlations of standard deviations of breeding values within families were also high, but decreased as the number of QTL increased (from $\sim 99\%$ to $\sim 75\%$, Figure S7b).

1.3.8. Calculation of selection intensity. The usefulness criterion is a function of the standardized selection differential or selection intensity i defined in (17). To compare true and predicted usefulness values we selected two values of i representing moderate and strong selection: 1.40 and 2.67, corresponding to selecting 20 or 1 percent of offspring within each family, using the equation:

$$i = f\left(\frac{\bar{Y}_{selected} - \bar{Y}_{population}}{sd_{population}}\right)/p.$$

Here, $f(\cdot)$ is the probability density function of the phenotypic distribution of the population, and we assume a normal distribution, and $(\bar{Y}_{selected} - \bar{Y}_{population})/sd_{population}$ represents how many population standard deviations the selected group mean is away from the population mean. p is the proportion of genotypes selected.

1.4. Results and Discussion

1.4.1. Computational phasing is highly accurate in this strawberry population. We compared genotypes of pairs of markers at different genetic distances between unphased offspring and phased parents. To account for recombination, we binned marker pairs by genetic distances in cM. Rates of inconsistencies increased with recombination distance (Figure 1.1), as expected. We then compared the observed rates of genotype inconsistency with expected rates if parents were correctly phased, or if phasing errors were introduced randomly within each parent with different frequencies. For marker pairs greater than 10 cM apart, our observed rate of genotype inconsistencies was close to the expected rate if parental phasing was perfect (Figure 1.1, black and red). However, for closer marker pairs, simulated inconsistency rates converged towards zero even when parental phasing was imperfect, but our observed inconsistency rate never dropped below $\approx 2.8\%$, perhaps due to a low amount of genotyping error or incorrect parent-offspring trios. These results are consistent with the computationally inferred phasing of markers being highly accurate in this population.

1.4.2. Prediction accuracy of family breeding value mean and standard deviation varied as heritability and number of QTL changed. Given the highly accurate computational phasing of parental haplotypes in this population (Figure 1.1), we used the haplotypes to simulate families of 200 offspring for 10,000 crosses between randomly paired genotypes and projected genotypes of causal alleles onto each simulated offspring. We then trained genomic prediction models (BayesC and RR-BLUP) on simulated parental phenotypes with varying levels of heritability (h^2) and used the trained models to predict the breeding values of the offspring (20; 63). We calculated the mean and standard deviation of these predicted breeding values per family and measured the accuracies (r) of these predictions using Pearson’s correlation with the true values ($r = cor(\hat{\mu}, \mu)$, and $r = cor(\hat{\sigma}, \sigma)$).

Using BayesC as a genomic prediction model, the average prediction accuracy of family means ranged between $r = 0.7$ and $r = 0.9$ (Figure 1.2a). We achieved moderate accuracy in predicting family standard deviations, averaging between $r = 0.3$ and $r = 0.8$ across scenarios (Figure 1.2b, blue). Results using RR-BLUP as a prediction model were similar (Figure S1). The prediction accuracies of family breeding values mean and standard deviations both increased as the heritability

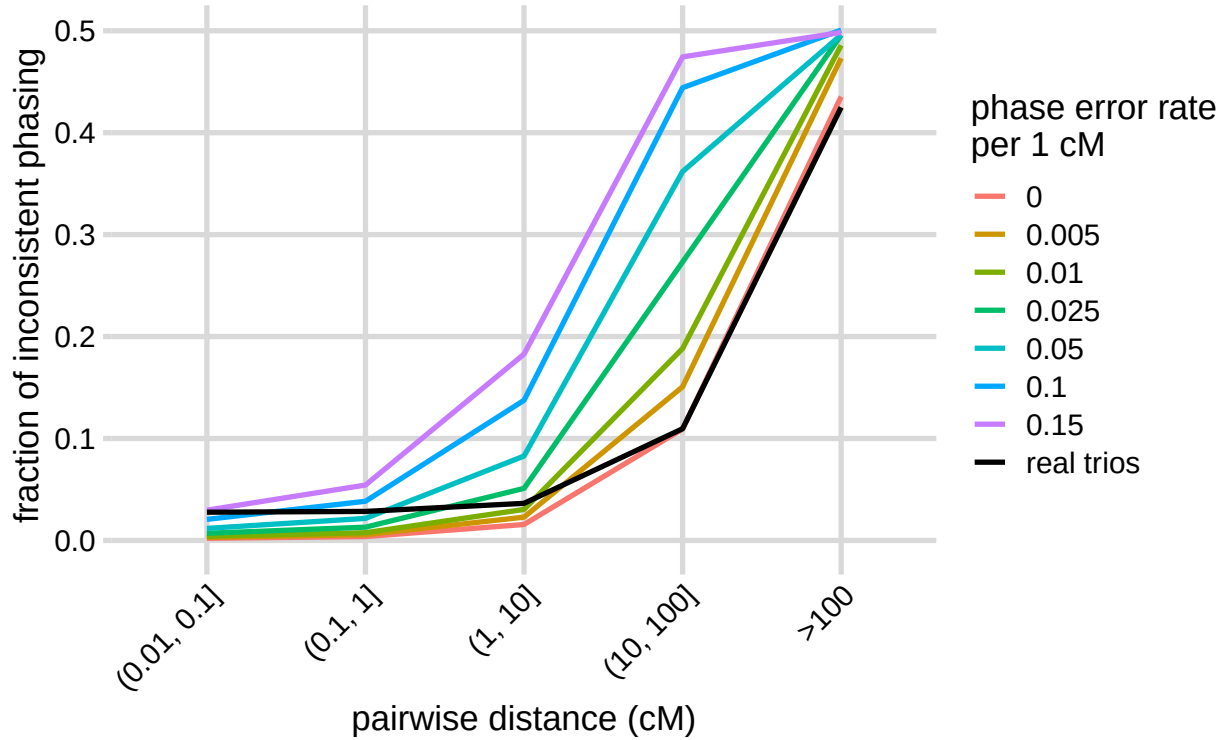


FIGURE 1.1. Rates of inconsistent parent-offspring genotypes at pairs of markers as a function of genetic distance. Pairs of marker genotypes, which both were heterozygous in at least one parent, were compared between parents and offspring across the 593 family trios. The black line shows the fraction of marker pairs in each genetic distance bin that are inconsistent between offspring and parents. Colored lines show simulated inconsistency rates for the same marker pairs after introducing haplotype switches (phasing errors) at the specified rates per cM in each parent.

increased, as expected. When the number of QTL increased, the prediction accuracy of the mean remained the same (Figure 1.2a), but that of standard deviation decreased (Figure 1.2b, blue). This might be due to less accurate estimates of individual QTL effects when there are more QTL. Even with a low heritability and a high number of causal loci, prediction accuracies of within family standard deviations did not decline to zero. This is likely because parents heterozygous for more loci also tend to be heterozygous for more causal loci. Only heterozygous causal loci contribute to the variance of breeding values among offspring in our simulations because the genetic architecture is purely additive. We used the number of heterozygous markers as a measure of an individual's heterozygosity, and calculated the Pearson's correlation between the average parental heterozygosity and the true family standard deviation. The average parental heterozygosity was more correlated

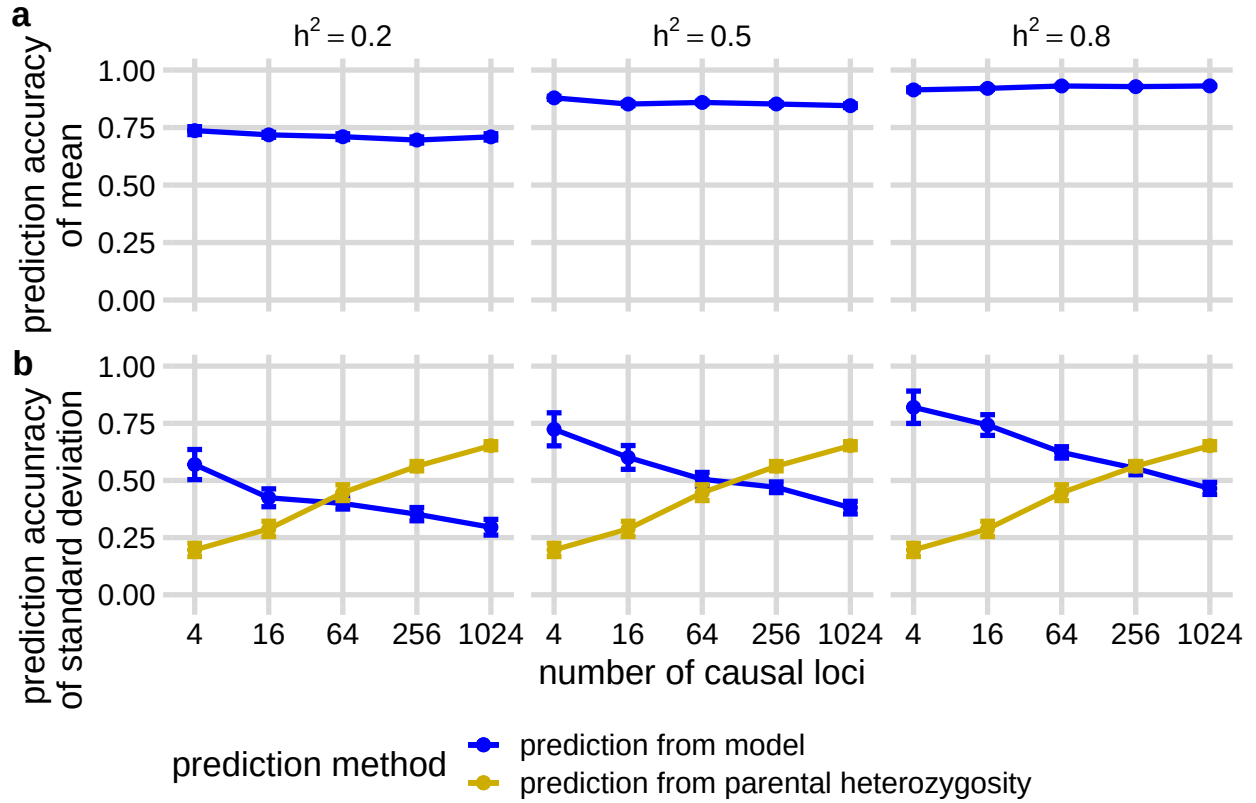


FIGURE 1.2. Prediction accuracies of the means (a) and standard deviations (b) of breeding values across simulated families as a function of heritability (h^2) and the number of causal loci. Panels show the means and standard error bars of Pearson's correlations (prediction accuracies, y-axis) between the true family means of breeding values (a) or true family standard deviations of breeding values (b, blue) based on a BayesC genomic prediction model trained on parental phenotypes simulated with different heritabilities (h^2) and different number of causal loci (x-axis). 20 simulations were run at each combination of h^2 and number of causal loci. Each simulation utilized 500 out of 10,000 simulated families to ensure independence among simulations. As a comparison, we also calculated the correlation between the true standard deviation of breeding values in each family and the average number of heterozygous markers of the two parents of that family calculated across all non-causal SNP markers in each simulation (b, gold). Points and bars show the means and standard errors of accuracies across 20 simulations. Curves are created by connecting the points of means and show the trend of prediction accuracies across simulations.

with the standard deviation of offspring breeding values than either the BayesC-based or RR-BLUP-based predicted standard deviations in some of our simulations (Figure 1.2b, gold, Figure S1b, gold), particularly when the number of causal loci was high.

Mean heterozygosity information may have been indirectly captured by the genomic prediction models to accurately predict standard deviation. With a small number of QTL, the predicted family standard deviations from our genomic prediction models were more accurate than predictions of standard deviation from parental heterozygosity. In contrast, with a large number of QTL, the prediction of standard deviation from parental heterozygosity was more accurate. Unfortunately, despite its high correlation with family standard deviations in some instances, it is not straightforward to use parental heterozygosity to directly predict family usefulness, because heterozygosity and family standard deviations have different units and scaling factors. In practice, real families with large enough numbers of genotypes would be needed to estimate the regressions between family standard deviations and parental heterozygosity before this statistic could be used to estimate usefulness.

1.4.3. Family usefulness can be accurately predicted with family mean. With the predicted family mean and standard deviation using the BayesC model, we were able to predict the usefulness of each family for a defined selection intensity i using the equation: $\hat{UC} = \hat{\mu} + i\hat{\sigma}$. We calculated the accuracy of the estimated usefulness as Pearson's correlation with the true usefulness ($cor(\hat{UC}, UC)$, Figure 1.3, gold) and compared this with the accuracy of predicting the usefulness using only the predicted family mean ($cor(\hat{\mu}, UC)$, Figure 1.3, blue) of each family. As heritability increased, the accuracy of both predictions increased. At low heritability, the prediction accuracies were relatively insensitive to the number of QTL. However, at high heritability, the prediction accuracies increased as the number of QTL increased. Predicting usefulness with both predicted mean and predicted standard deviation (i.e. $\hat{UC} = \hat{\mu} + i\hat{\sigma}$) only achieved meaningfully higher accuracies than predicting usefulness with the predicted mean itself (i.e. $\hat{\mu}$), when heritability was high and when there were small numbers of QTL (Figure 1.3, panels $h^2 = 0.5$ and $h^2 = 0.8$, at 4 and 16 QTL), the same conditions under which we could most accurately predict the family standard deviation (Figure 1.2b, blue). The results using an RR-BLUP model were similar (Figure S2).

Given the moderate accuracy at predicting both the standard deviation and mean of each family, we were surprised that direct predictions of usefulness were not much better than predictions based on predicted family means alone. One explanation for this observation is that the true values of family mean and family usefulness are highly correlated ($cor(\mu, UC)$), especially when the number of causal loci is large (Figure 1.4). The reason for this high correlation is that the variance among families in

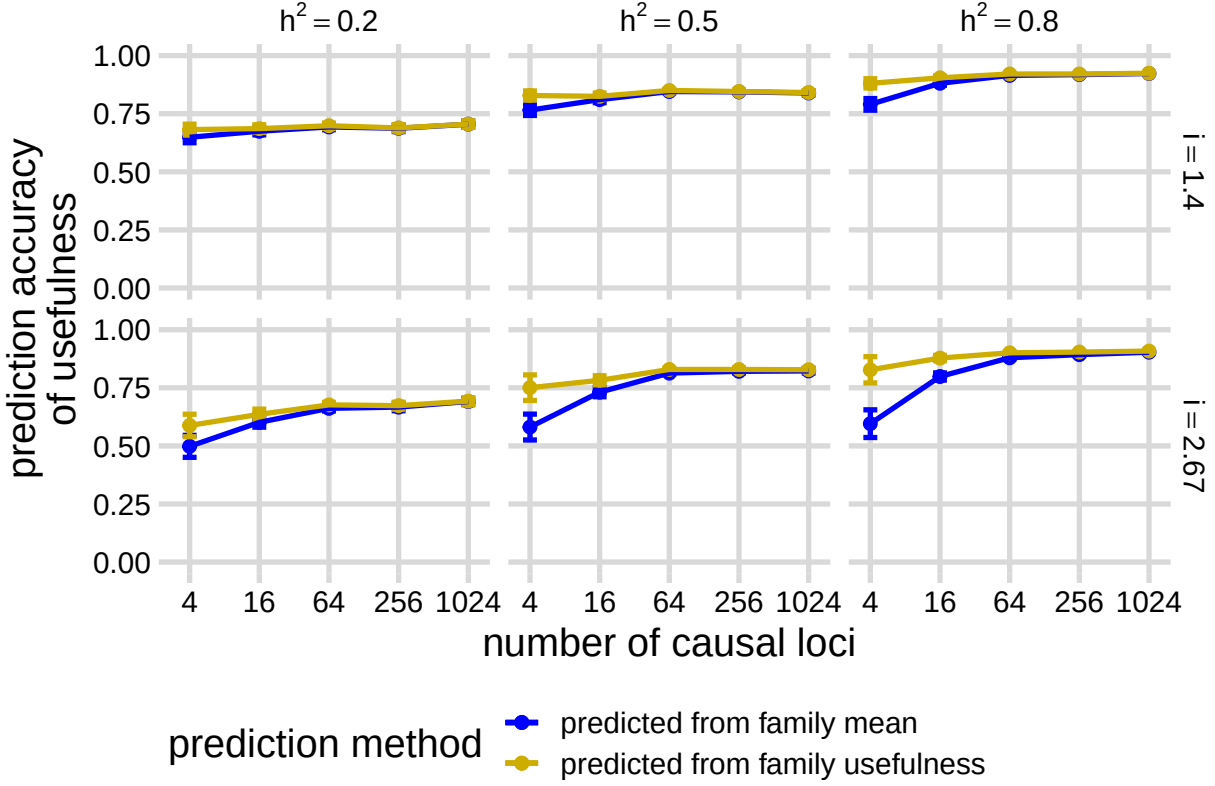


FIGURE 1.3. **Prediction accuracies of the usefulness across the simulated families using a BayesC genomic prediction model.** Panels show the means and standard error bars of Pearson's correlations (prediction accuracies, y-axis) between the true family usefulness and predicted family mean (blue) or predicted family usefulness (gold) based on a BayesC genomic prediction model trained on parental phenotypes simulated with different heritabilities (h^2) and different numbers of causal loci (x-axis), across different selection intensity (i). 20 simulations were run at each combination of h^2 and number of causal loci. Points and bars show the means and standard errors of accuracies across 20 simulations across each i . Curves are created by connecting the points of means and show the trend of prediction accuracies across simulations.

the quantity $i\sigma$ (selection intensity multiplied by standard deviation) is very small compared to the variance among families in mean (μ), so the first term in the usefulness dominates. This mirrors observations from (89) in biparental recombinant inbred line populations that the variance of family means increases at a greater rate than the variance of family standard deviations as the number of QTL increases. They defined the parameter $t = \text{var}(\sigma)/\text{var}(\mu)$, the ratio between the variance of true family standard deviations and variance of means and showed that in biparental populations t accurately predicts the value of usefulness. We estimated t for each simulation (Figure S5, blue) and found that it also decreased dramatically as the number of causal loci increased. However, t

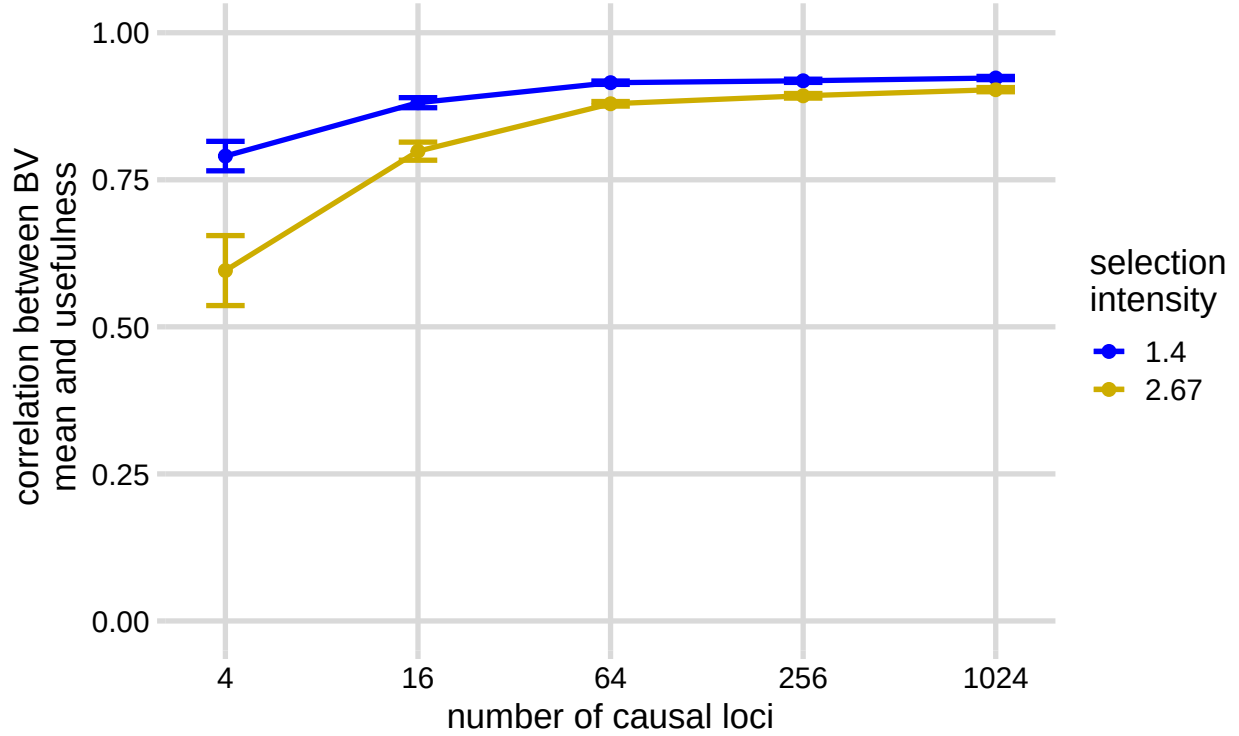


FIGURE 1.4. **Correlations between true family breeding usefulness and mean at two selection intensity levels.** Pearson’s correlation (y-axis) between the true family means of breeding values and true family usefulness across different selection intensities (i) and the number of causal loci (x-axis). 20 simulations were run at each number of causal loci. Points and bars show the means and standard errors of accuracies across 20 simulations at $i = 1.4$, blue, and at $i = 2.67$, gold. Curves are created by connecting the points of means and show the trend of correlations across simulations.

was generally larger in our populations than expected based on Equation 8 of (89): $t = (4L)^{-1}$, where L is the effective number of causal loci (Figure S5, gold), which assumes that all loci are independent, have equal allele frequencies ($q = 0.5$), and equal effect sizes. The increase of t in our simulations is expected because there was LD among loci, unequal effect sizes, and unequal allele frequencies. This reduces the “effective” number of loci, as in Equation 9 of (89). However, that equation does not account for unequal allele frequencies, which increases t even more, as they predicted. Nevertheless, t values are still small in our simulations when there were 4 causal loci, which explains the negligible value of usefulness.

1.4.4. Predicting usefulness improves cross-selection accuracy in a two-stage approach, but results in little additional genetic gain. (89) observed that by restricting possible crosses to only those among lines first selected for breeding value, (*i.e.*, elites), the variance in

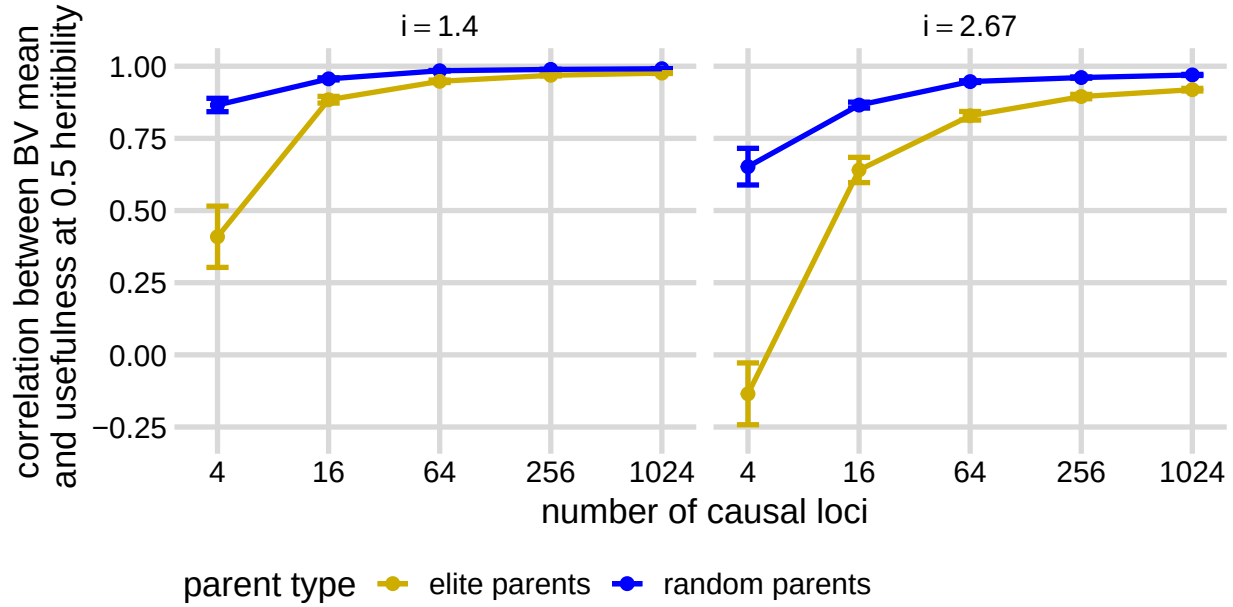


FIGURE 1.5. **True family breeding value means and true family usefulness are less correlated in the elite families.** Pearson's correlation (y-axis) between the true family means of breeding values and true family usefulness across different selection intensities (i) and the number of QTL (x-axis), from the initial 10,000 families (blue) and from the families with the highest predicted breeding value individuals as parents (gold) at 0.5 heritability. Points and bars show the means and standard errors of accuracies across 20 simulations at $i = 1.4$ and at $i = 2.67$. Curves are created by connecting the points of means and show the trend of correlations across simulations.

family means will be reduced, and thus the relative importance of standard deviation for usefulness will increase. We tested whether our prediction accuracy of usefulness in this context could improve crossing decisions. Across the range of genetic architectures, the true correlation between the family mean breeding value and usefulness was lower among crosses from the elite parents than among random crosses across all numbers of causal loci and selection intensities for these new families, reaching values as low as 0 when selection intensities were strong and there were few QTL, but remained moderately high when the number of QTL was greater than 60 (Figure 1.5, S6).

Using our previously trained BayesC models, prediction accuracies of usefulness were moderately higher using predicted usefulness for each candidate family than using predicted family means in the high heritability/few QTL/low selection intensity scenarios, so the benefit increased when the selection intensity was high and the number of QTL was small (Figure 1.6).

These results suggest that a two-stage crossing design strategy, in which parents are first selected by predicted breeding value and then crosses among these selected parents are designed in a second

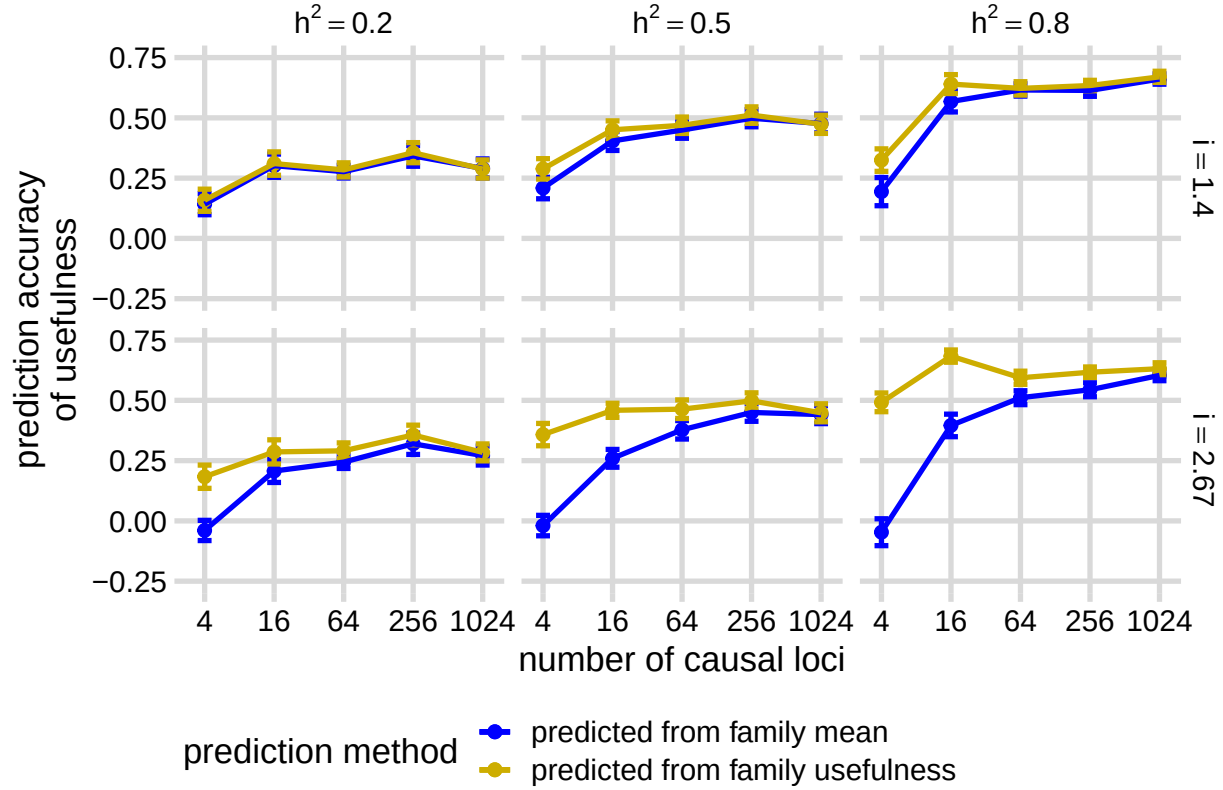


FIGURE 1.6. **Prediction accuracies of the usefulness across the elite families using a BayesC genomic prediction model.** Panels show means and standard error bars of Pearson's correlations (prediction accuracy, y-axis) between the true family usefulness and predicted family mean (blue) or predicted family usefulness (gold) based on a BayesC genomic prediction model trained on parental phenotypes simulated with different heritabilities (h^2) and different numbers of causal loci (x-axis), across different selection intensity (i), from the families with the highest predicted breeding value individuals as parents. 20 simulations were run at each combination of h^2 and number of causal loci. Points and bars show the means and standard errors of accuracies across each i . Curves are created by connecting the points of means and show the trend of prediction accuracies across simulations.

state, might be an effective method for leveraging usefulness to increase genetic gain, at least when heritability and/or selection intensity is high and the genetic architecture has a small number of QTL. However, the prediction accuracies and selection intensities shown in Figure 1.6 are not direct parameters of the traditional breeder's equation, usually written as $\Delta G = ir\sigma_A$. The parameter i from the breeder's equation represents the selection intensity relative to all possible offspring, while the i in the usefulness equation is the selection intensity within each family. The parameter r from the breeder's equation represents the accuracy of predicted genetic values of each offspring, while

the r from Figure 1.6 represents the accuracy of predicted average genetic values of the selected offspring in each cross. Additionally, Figure 1.6 does not show the population additive genetic variance σ_A , and the second stage of the two-stage crossing design will have less genetic variance because only crosses among the previously selected parents are considered. Therefore, Figure 1.6 provides an incomplete picture of the potential benefit of predicting usefulness for increasing genetic gain.

To assess whether the increases in accuracy from predicted usefulness shown in Figure 1.6 actually result in meaningful increases in genetic gain, we estimated the genetic gain expected for this two-stage approach and compared it to the expected genetic gain from the one-stage approach that selected crosses from all possible crosses using either predicted usefulness or predicted mean breeding values. The expected genetic gain is the mean genetic value of the top offspring from each of the selected crosses, using a selection intensity of i **within each cross**. For each approach, we created 10 crosses using 20 unique parents. For the two-stage approach, we selected the 40 parents by predicted breeding values, and predicted usefulness values for all 780 candidate crosses, and then chose 10 crosses using a greedy search algorithm, first choosing the cross with the highest predicted usefulness, then adding the next-best cross that did not use any previously selected parents, and so-on until 10 crosses were selected. For the one-stage approach, we predicted family means and family usefulness for all 506,521 possible crosses, and then applied the same algorithm to select 10 crosses using 20 unique parents. Note that a two-stage approach using predicted family means as the cross-section criteria in the second stage would choose identical crosses as the one-stage approach using predicted family means. Since simulating crosses for 506,521 crosses would be prohibitively computationally expensive, we instead simulated populations of gametes from each parental genotype, and predicted family means as the sum of the mean predicted gamete genetic values of each parent, and predicted family variances as the sum of the variances of predicted gamete genetic values.

Figure 1.7 shows the expected genetic gains in units of σ_A for each selection strategy. Consistent with the results of Figure 1.3, predicting usefulness can increase genetic gain, and never results in lower rates of gain on average. However, except in cases of extreme within family selection intensity, high heritability, and few QTL, the gain is rarely large. In contrast to the impression given by Figure 1.6 and to the hypothesis by (89), there was limited benefit of selecting crosses

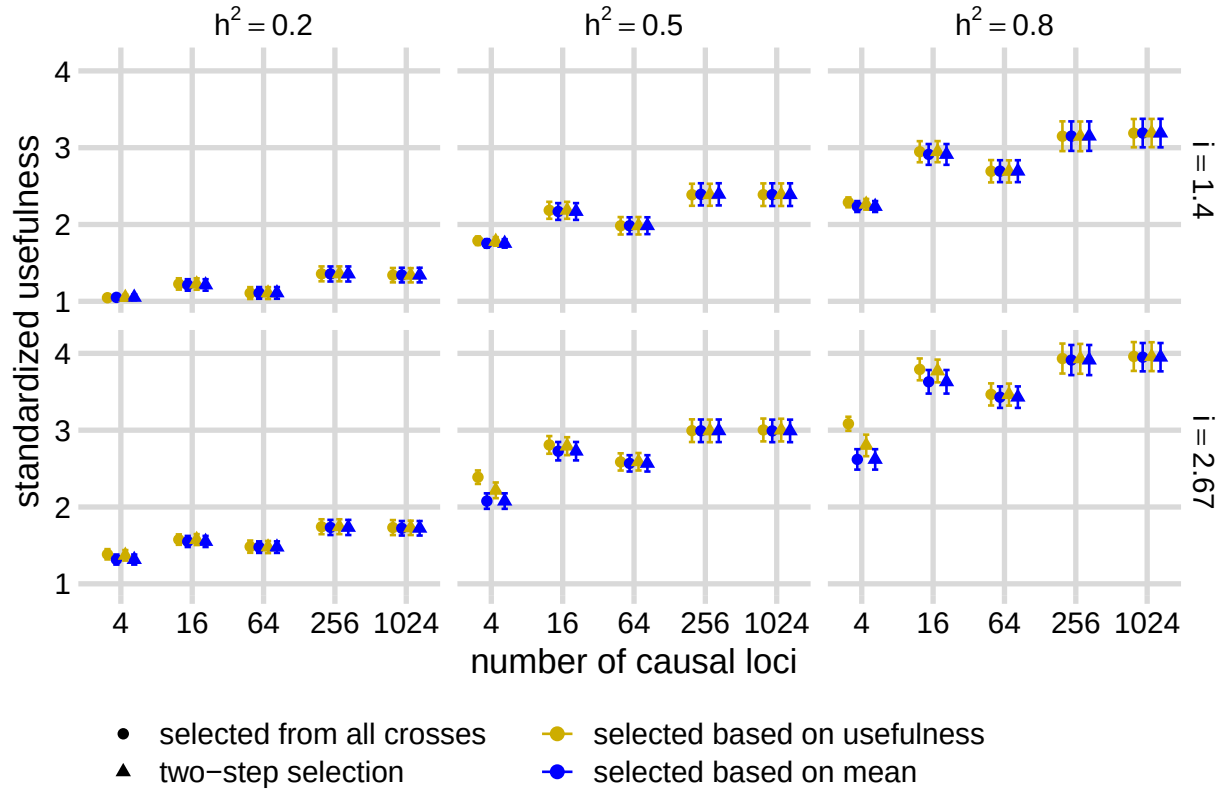


FIGURE 1.7. **Selection on predicted family usefulness from all possible crosses gave crosses with the highest breeding value usefulness.** Panels show means and standard error bars of standardized family breeding value usefulness (y-axis) against the number of QTL (x-axis) from crosses based on four selection methods across different heritability (h^2) and selection intensity (i). The selection methods included selecting the top 10 crosses among all possible crosses with the highest predicted family usefulness (gold, circle) without replacement, selecting the top 10 crosses among all possible crosses with the highest predicted family mean (blue, circle) without replacement, selecting top 40 ranked parents based on BayesC model predicted breeding values and then selecting the top 10 crosses among them with the highest predicted family usefulness (gold, triangle) without replacement, and selecting top 40 ranked parents based on BayesC model predicted breeding values and then selecting the top 10 crosses among them with the highest predicted family mean (blue, triangle) without replacement.

by predicted usefulness after selecting parents, even when the accuracy of predicted usefulness was high. The two-stage approach never had higher expected genetic gains than the one-stage approach using predicted usefulness.

1.4.5. Mate selection based on family usefulness provides little benefit in most cases and minimal benefit in some cases. Together, our results show that the true family mean is highly correlated with the true family usefulness under most scenarios in an outcross species like

strawberries. Our results validated conclusions made by (89), in which they stated limited benefit of usefulness in inbred species. We show that the accuracy of computational phasing is unlikely to be the limiting factor in the utility of usefulness predictions for improving genetic gain. Even when phasing is successful, as it appears to be in our strawberry population, and when predicted usefulness values are relatively accurate, predicting usefulness values for potential crosses only has meaningful benefit compared to only predicting family means when the genetic architecture is dominated by few QTL, when the heritability is moderate or high, and when family sizes are so large that very high within family selection intensities can be achieved (Figure 1.7, panels $h^2 = 0.5$, $h^2 = 0.8$, and $i = 2.67$ at 4 QTL). One possible explanation for our high phasing accuracy is that our reference strawberry breeding program includes many parent-offspring pairs which are highly informative for inferring phasing among markers. In other programs without such a strong pedigree structure, lower phasing accuracy may further limit the benefit of usefulness.

One limitation of our study is that we only simulated additive genetic architectures which is justified if we are only interested in population improvement. Under purely additive architectures, the usefulness of a cross is simply a function of the average usefulness of each parent. In this case, usefulness can be used equally as a line or mate selection metric. However, for variety release in a clonally propagated crop like strawberry, the usefulness criterion can be extended to total genetic values where dominance and epistasis effects will additionally be important (86). In this case, if the within family variance in total genetic values becomes large enough so that the variance of $i\sigma_j$ is of similar magnitudes to μ_j , the importance of usefulness would increase. However, predicting dominance and epistatic effects from genome-wide marker data remains very challenging and generally requires much larger sample sizes than are currently possible in most breeding programs. (86) found little benefit from predicting dominance variance. In future work, we could extend our simulations to include dominance and epistatic variance to confirm these intuitions.

1.5. Conclusions

In this study, we discuss whether selecting crosses based on predicted values of the usefulness of each cross could improve the rate of gain in breeding programs. Predicting the usefulness requires predicting the standard deviations of breeding values among the offspring produced by each cross. Our results show that predicting usefulness improves outcomes relative to simply predicting the

mean breeding value of each family only in very specific scenarios with high heritability, high selection intensity, and a small number of QTL (Figures 1.3, 1.4a). However, even in these scenarios, the relative gain in accuracy was small. We have also shown that the true usefulness of the top crosses selected either with predicted family mean or with usefulness only meaningfully differ in those same scenarios (Figure 1.7). Due to the rare instances where prediction of usefulness is beneficial, usefulness is rarely useful.

1.6. Data Availability

The data used in this study are available. The supplementary figures are provided. The codes are available at https://github.com/Faye-Wong-stat/family_variance.git.

1.7. Acknowledgement

We thank Lily Bhattacharjee for the initial data formatting.

1.8. Funding

This study was supported by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Agriculture Genomes 2 Phenomes Initiative (AG2PI), grant number 026418H in addition to funding from the University of California, Davis.

CHAPTER 2

Do Not Benchmark Phenomic Prediction against Genomic Prediction Accuracy

2.1. Abstract

Phenomic selection is a new paradigm in plant breeding that uses high-throughput phenotyping technologies and machine learning models to predict traits of new individuals and make selections. This can allow breeders to evaluate more plants in higher throughput more accurately, resulting in faster rates of gain and reduced labor costs. However, phenomic prediction models are frequently benchmarked against genomic prediction models using cross-validation to demonstrate their usefulness to breeders. We argue that this is inappropriate for two reasons: 1) differences in the accuracy statistic measured by cross-validation do not reliably indicate differences in the accuracy parameter of the breeder's equation, which we show analytically and through re-analysis of data from three representative phenomic prediction studies, and 2) phenomic and genomic selection tools influence other parameters of the breeder's equation, so comparing accuracy, even if done properly, is insufficient to advocate for one approach over the other. We conclude that phenomic selection may be useful, but comparisons of accuracy between genomic prediction and phenomic prediction models are not.

2.2. Introduction

Breeders create new varieties by iterating through cycles of population creation, evaluation, and selection to create the next generation. The expected outcome of each cycle is approximated by the breeder's equation (53):

$$(2.1) \quad \Delta g = \frac{i\rho\sigma_A}{L}$$

where Δg is the improvement in the average additive genetic value per year; i is the standardized selection intensity, the difference in the mean trait value of the selected individuals relative to

the whole population, standardized by the standard deviation of trait values; ρ is the accuracy of selection, the Pearson correlation coefficient between the estimated and true additive genetic values among the selection candidates; σ_A is the standard deviation of additive genetic values; and L is the cycle length, the difference in years between when the population is created and when selections are made. While the Breeder’s equation is only an approximation of the effect of selection in real breeding programs (59; 11), studying how changes in breeding procedures jointly impact these four parameters can give insight into the total impact on genetic gain.

In a traditional breeding program, selection decisions are based on phenotypic measurements of each candidate individual. In many species, collecting this data requires generating lines to grow in field trials and running replicated trials across multiple locations. Therefore, each selection cycle can take several years.

(55) introduced genomic selection as a tool to increase the rate of gain in breeding programs. Genomic selection means selecting candidate individuals based on estimates of genetic values from a genomic prediction (GP) model using genetic marker data. Genomic selection can increase the rate of gain (Δg) in a breeding program by increasing selection intensity (i), increasing selection accuracy (ρ), or reducing selection lengths (L). If genetic marker data are inexpensive, more candidates can be evaluated using a GP model than can be evaluated directly using field trials. Larger numbers of candidates allow breeders to impose a higher selection intensity while maintaining a constant number of selected individuals to conserve genetic variance (9; 18). GPs of additive genetic values can be more accurate than direct estimates from field trials when candidates are genetically related at causal loci because the sample size per allele can be higher than the sample size per candidate (88). Genomic predictions of additive genetic values can be made as soon as genotype data are collected, even on seeds that have yet to germinate in some species (78). This can enable multiple cycles per year (10). However, multiple factors continue to limit the use of genomic selection in plant breeding programs, including the cost of large-scale genotyping, the need to train and continuously update GP models, and the logistics of reorganizing existing breeding schemes.

(65) introduced phenomic selection as an alternative tool for breeders that can be less expensive and less logistically challenging than genomic selection. Like genomic selection, phenomic selection also relies on a statistical or machine learning-based model to make selections, called a "phenomic prediction" (PP) model, because instead of genetic markers, the features of the model are measured

by high-throughput phenotyping technologies (HTP). For example, (65) used multispectral image data from wheat leaves and grains and poplar wood tissue, (85) used spectral image data from a camera mounted on an aerial vehicle, and (47) used hyperspectral image data from a camera mounted on an aircraft. Phenomic selection has garnered interest in the breeding community for several reasons: i) HTP data can often be collected at lower cost and higher throughput than genotype data, enabling breeders to evaluate more candidates at once, or evaluate a common set of candidates more times to increase the heritability. ii) Logistically, phenomic selection may require fewer changes to breeding schemes than genomic selection, and many of the same software packages used for implementing GP models also work for PP models. iii) PP models often achieve relative high accuracy when evaluated using cross-validation.

Both genomic and phenomic selection schemes have the potential to accelerate plant breeding. But, since implementing either requires significant investment in equipment and changes to existing breeding schemes, there is considerable interest in determining which approach is best. Multiple recent papers have demonstrated the potential of phenomic selection by showing estimates of accuracies of PP models that are as high or higher than estimates of accuracies of GP models made using the same datasets. However, our objective is to show that these comparisons *are not useful* for evaluating the relative benefits of implementing phenomic and genomic selection schemes in breeding programs.

One of the reasons is that the accuracy statistic reported by these studies is not an appropriate estimator of the selection accuracy parameter of the breeder’s equation (ρ). For simplicity, we assume that we have evaluated a set of n candidates by some method (*i.e.* GP or PP), and that the candidates are unrelated and so independent and identically distributed. While this is not exactly true in any real breeding program, these assumptions are useful for illustrative purposes. The *accuracy* of selecting among the candidates, represented by the parameter ρ in the breeder’s equation, is the Pearson correlation coefficient between the estimated ($\hat{a}$) and true (a) breeding values (*i.e.* additive genetic values) of the candidates:

$$(2.2) \quad \rho = \text{cor}(\hat{a}, a) = \frac{\text{cov}(\hat{a}, a)}{\sqrt{\text{var}(\hat{a})}\sqrt{\text{var}(a)}}.$$

Unfortunately, it is not easy to create a consistent estimator of this parameter because additive genetic values (a) cannot be directly observed. Instead, we do observe phenotypes of candidates

(y). The Pearson correlation coefficient between estimated breeding values and phenotypes (ρ_y) is:

$$(2.3) \quad \rho_y = \text{cor}(\hat{a}, y) = \frac{\text{cov}(\hat{a}, y)}{\sqrt{\text{var}(\hat{a})}\sqrt{\text{var}(y)}}.$$

We can easily form a consistent estimator of ρ_y as:

$$(2.4) \quad r_y = \widehat{\text{cor}}(\hat{a}, y)$$

using cross-validation (58; 49), where $\widehat{\text{cor}}(\hat{a}, y)$ is the sample correlation. (49) coined the term *predictive ability* (ρ_y) to distinguish it from *accuracy* (ρ), but these terms are not used consistently in the plant breeding literature (85; 47). While r_y is a consistent estimator of ρ_y , it is not a consistent estimator of ρ because $\hat{a}$ and y share non-additive genetic and environmental effects, even for individuals not used for model training, as we elaborate on below.

Thus, using r_y to predict responses to selection will usually not be accurate.

Another reason is that genomic and phenomic selection schemes impact other parameters of the breeder’s equation differently, so comparisons need to consider the full effects on genetic gain. We will further discuss this in Discussion.

In this study, first, we used a more appropriate technique to estimate PP *accuracy* (24; 68; 52), which uses a bivariate linear mixed model to directly estimate the genetic correlation between predicted and observed trait values,

and re-analyzed published data from (65; 47; 85), to show that conclusions about the relative merits of GP and PP change when compared using this method. Then, we demonstrate these observations analytically. We conclude that phenomic selection may be useful, but comparisons of *accuracies* between GP and PP models are not.

2.3. Materials and Methods

We replicated the comparisons between GP and PP *accuracy* reported in three published studies to test how each set of conclusions would change if *accuracies* were estimated using the method in (68) instead of as *predictive ability*. In each case, we attempted to reproduce the GP and PP models exactly as reported in each original paper.

2.3.1. Case Study 1. (65) used two datasets, one from wheat (*Triticum aestivum* L.) and one from poplar (*Populus nigra* L.), to compare the *predictive abilities* of GP and PP models.

The wheat data consisted of 223 genotypes measured for heading date and grain yield in two treatments (dry and irrigated). We considered the 4 trait-treatment combinations as 4 traits (heading date in the dry treatment, heading date in the irrigated treatment, grain yield in the dry treatment, and grain weight in the irrigated treatment). Genomic data consisted of 84,529 SNPs, and we used it to train GP models for the 4 traits. Phenomic data consisted of near-infrared spectroscopy scans from both leaf and grain tissues in each treatment. Phenotypic data were reported as best linear unbiased estimators (BLUES) of genotype means correcting for random block effects. We trained PP model using phenotypes from the 4 traits and their corresponding phenomic data, following the methods in the source paper.

The poplar data consisted of 562 genotypes measured for height, circumference, bud flush, bud set, and rust resistance in two locations: Orleans, France and Savigliano, Italy. All 5 traits were measured at Orleans, but only circumference, bud flush and bud set were measured at Savigliano. Genomic data consisted of 7,808 SNPs. Phenomic data were collected from the wood samples in both locations. Phenotypic data were reported as BLUES of genotype means correcting for random block effects. We considered each of the 8 trait-location combination as independent traits and used the corresponding phenomic data along with genomic data to train PP and GP models following the methods in the source paper.

For each trait in each species, we fit the following GP model following the methods of (65):

$$(2.5) \quad y_i = \mathbf{x}_i^T \boldsymbol{\alpha} + \epsilon_i$$

where y_i is the BLUE of the phenotypic value of genotype i , $\mathbf{x}_i$ is the SNP markers for genotype i , $\boldsymbol{\alpha} = \{\alpha_k\}$, $\alpha_k \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma_a^2)$ are the random marker effects, and $\epsilon_i \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma_e^2)$ are residuals.

Also following (65), we fit the following PP model for each trait in each species:

$$(2.6) \quad y_i = \mathbf{p}_i^T \mathbf{h} + e_i$$

where $\mathbf{p}_i$ is the phenomic features for genotype i , $\mathbf{h} = \{h_k\}$, $h_k \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma_h^2)$ are the random phenomic feature effects, and $e_i \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma_e^2)$ are residuals.

We estimated the *predictive ability* of PP and GP models as the sample Pearson’s Correlation Coefficient between predictions and observed BLUEs using eight-fold cross-validation for the wheat data, and using five-fold cross-validation for the poplar data, following (65). We estimated PP *accuracy* using the method by (68) by fitting a bivariate linear mixed model (see Appendix Part 3) between the observed phenotypic traits and the predicted values from PP models, using the marker-based kinship matrix as the estimate of additive genetic relationships among individuals. We estimated GP accuracy as $r_y/\sqrt{h^2}$ (see Results and Theory Part 3) (49).

2.3.2. Case Study 2. (85) compared GP and PP *predictive abilities* for grain yield in wheat. The dataset consisted of grain yield data for 3,442 genotypes measured across multiple replicated trials across multiple years and locations. Genotype data consisted of 23,130 SNPs. We removed SNPs with a minor allele frequency of or less than 5%. Phenomic data were derived from images taken by unoccupied aerial vehicles from the same genotypes across different years and different locations. We followed Scenario 1 in (85) and calculated BLUEs of grain yield and of each spectral band at each genotype, correcting for the random effect of unique combinations of year and location. For GP, we used equation 2.5 as above, matching the GP model reported by (85). For PP, we followed the same methods of (85) and used the following linear model which treats the phenomic features as fixed effects:

$$(2.7) \quad y_i = \mathbf{p}_i^T \mathbf{l} + \epsilon_i$$

where $\mathbf{l}$ is the vector of fixed effects of the phenomic features. We also used a linear mixed model, equivalent to Equation 2.6, which treats the phenomic features as mixed effects for PP.

We used five-fold cross-validation to obtain the Pearson’s correlation between the predicted values and the BLUEs. We estimated the *accuracy* in the same way as in Case Study 1.

2.3.3. Case Study 3. (47) also compared GP and PP *predictive abilities* for grain yield in wheat. The dataset consisted of grain yield data for 3,771 genotypes measured from multiple replicates on the same genotype across different breeding cycles and different management practices. The authors considered each combination of breeding cycle and management practice as a unique treatment, or "site-year" as in (47). Genomic data were provided as a genomic relationship matrix

calculated from 34,900 SNPs. Phenomic data consisted of hyperspectral reflectances collected with a hyperspectral camera mounted in an aircraft. The phenomic data were collected multiple times during the growing season for each treatment. The authors provided the BLUEs of grain yield and of each hyperspectral band for each genotype within each treatment, correcting for random effects of trial, replicate, and block. This was described as the **ALL** method in (47).

For the grain yield of wheat in (47), we fit a GP model for genotype i within each treatment as follows:

$$(2.8) \quad y_i = g_i + \epsilon_i$$

where $\mathbf{g} = \{g_i\}$, $\mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{K}_g)$ is the random additive genetic effects with genomic relationship matrix $\mathbf{K}_g$ and additive genetic variance parameter σ_a^2 . Residuals $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma_e^2)$ were assumed independent and identically distributed among genotypes. This is the same model as Equation 4 in (47).

Following (47), we used the following PP model for genotype i within each treatment:

$$(2.9) \quad y_i = q_i + \epsilon_i$$

where y_i is the BLUE of phenotypic values of genotype i in each treatment, $\mathbf{q} = \{q_i\}$, $\mathbf{q} \sim N(\mathbf{0}, \sigma_H^2 \mathbf{K}_H)$ is the random effect of phenomic features on genotype i in each treatment with feature covariance matrix $\mathbf{K}_H = \frac{\mathbf{P}\mathbf{P}^T}{n_{bands}}$, where $\mathbf{P}$ is the centered and standardized BLUEs of hyperspectral bands of genotype i in each treatment, and n_{bands} is the total number of hyperspectral bands. This PP model is the same model described in Equation 5 in (47).

We used five-fold cross-validation to obtain the Pearson's correlation between the predicted values and the BLUEs within each treatment. We estimated the *accuracy* in the same way as in Case Study 1.

2.3.4. Software. All analyses were conducted in R version 4.2.3 (75). Model 2.7 was fitted using the *lm* function in Base R. Model 2.5 and 2.6 were fitted using *mixed.solve* function from the *rrBLUP* package (20). Model 2.8 and 2.9 were fitted using the *BGLR* function from *BGLR* package (63). *Accuracies* were estimated using the *estimate_gcor* function from (70) which utilizes

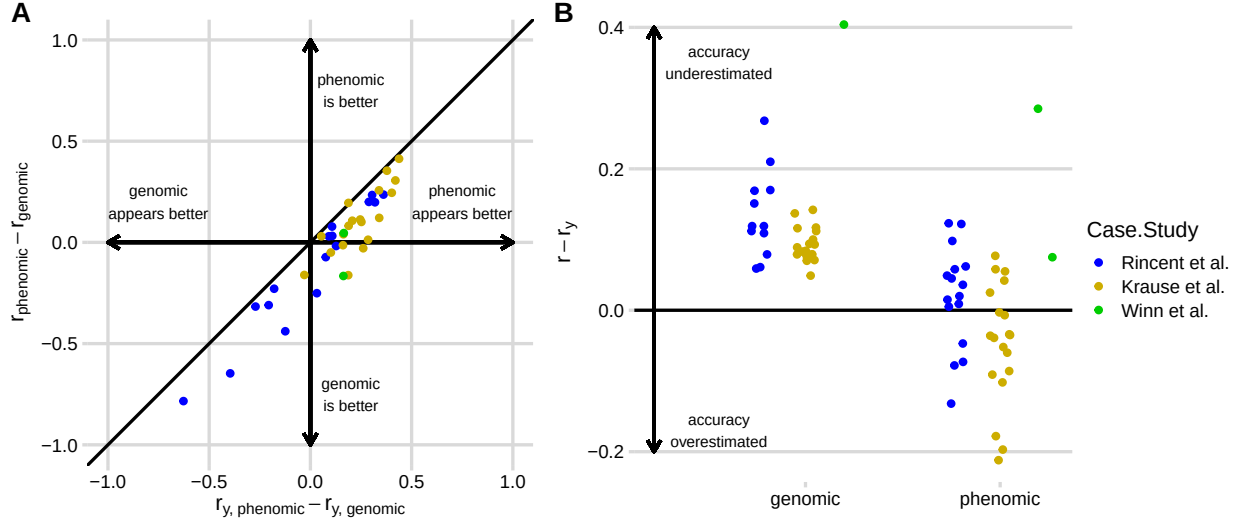


FIGURE 2.1. **Predictive ability overestimates the relative accuracy of PP vs GP models.** A) Differences between estimated PP and GP *accuracies* ($r_{\text{phenomic}} - r_{\text{genomic}}$) (y-axis) are plotted against the differences between estimated PP and GP *predictive abilities* ($r_{y, \text{phenomic}} - r_{y, \text{genomic}}$) (x-axis). The black diagonal shows the 1:1 line. Points to the right of the line indicate that the *accuracy* of PP models tends to be more over-estimated than the *accuracy* of GP models. Points in the lower-right quadrant are cases where PP models appear better than GP models when compared by *Predictive ability* but we estimate that their *accuracy* is actually lower. B) Differences between estimated *accuracies* and *predictive abilities* ($r - r_y$) (y-axis) are plotted for GP and PP models, binned by case study. Colors in both panels identify individual traits in each case study.

MCMCglmm package (35). BLUEs were calculated using the *lme4* and *emmeans* packages (4; 51). Graphs were made with the *ggplot2* and *cowplot* packages (83; 84).

2.4. Results and Theory

2.4.1. Re-analysis of Published Data. We re-analyzed the data from (65; 85; 47) to acquire the estimated *predictive abilities* ($r_y = \widehat{\text{cor}}(\hat{a}, y)$ to estimate $\rho_y = \text{cor}(\hat{a}, y)$) of both GP and PP models using methods mentioned above. We calculated the estimated *accuracies* ($r = \widehat{\text{cor}}(\hat{a}, a)$ to estimate $\rho = \text{cor}(\hat{a}, a)$) of both GP and PP models, using the method in (49) and (68), respectively. For nearly every trait across the three studies, our estimates of the difference between PP and GP *accuracies* were smaller than the reported differences in *predictive abilities*, and all but one points fell below the $y = x$ line (Figure 2.1A). Thus, by reporting *predictive abilities*, these studies overestimated the relative benefits of PP. In particular, there were eight traits, including at least

one from each of the three studies, where *predictive abilities* were higher for PP, but our estimate of *accuracy* was actually lower for PP, compared to GP. These points are in Quadrant IV of Figure 2.1A. In all cases, GP *accuracies* were underestimated by *predictive ability* (Figure 2.1B). In contrast, PP *accuracies* were frequently overestimated by *predictive abilities* (Figure 2.1B).

2.4.2. Sources of bias. We provide simplified analysis and make assumptions for illustrative purposes to demonstrate the sources of bias in the following sections. These assumptions may not be realistic in real-world breeding programs, but we emphasize their importance in conveying our theories.

To understand the source of bias in r_y relative to ρ , consider a phenotypic selection scheme on unrelated individuals where $\hat{a} = y$, and a simple path diagram (53; 19) relating phenotypes (y) to additive genetic values (a):

$$(2.10) \quad \begin{array}{ccccccc} & u & & e & & m & \\ & \searrow & & \searrow & & \searrow & \\ a & \longrightarrow & g & \longrightarrow & t & \longrightarrow & y \longrightarrow \hat{a} \end{array}$$

Here, arrows represent causal paths which are assumed to be linear. a and u are additive and non-additive genetic sources of variation, which together determine the total genetic value (g) of a candidate. An organism's actual trait value t is caused by its genetic value plus an effect of its environment (e). We make an observation (y) which is subject to measurement error (m), and use this directly as our selection criterion ($\hat{a} = y$). Based on this diagram, $\text{cor}(\hat{a}, a) \neq \text{cor}(\hat{a}, y) = 1$ because there are exogenous variables that affect $\hat{a} = y$ but do not affect a (*i.e.*, u , e , and m), which we assume are uncorrelated with each other and with a . In particular, $\text{cov}(\hat{a}, y) = \text{var}(y) \neq \text{cov}(\hat{a}, a)$ because u , e , and m are all common causes of y and $\hat{a}$, but not of a .

Next, consider a PP model $f_p(\mathbf{p}) = \hat{a}_p$ with $\mathbf{p}$ a vector of phenomic features measured on a candidate. In a phenomic selection scheme, candidates will be ranked by $\hat{a}_p$, so this is the selection criterion: $\hat{a} = \hat{a}_p$. To see that $\rho_y \neq \rho$ for $f_p(\mathbf{p})$, we expand the path diagram to include causes of the phenomic features:

$$(2.11) \quad \begin{array}{ccccccc} & a & \longrightarrow & g & \longrightarrow & t & \longrightarrow & y \\ & & & \nearrow & & \nearrow & & \nearrow \\ & & & u & & e & & m \\ & & & \nwarrow & & \nwarrow & & \\ & & & \mathbf{u}_p & & \mathbf{e}_p & & \mathbf{m}_p \\ & & & \downarrow & & \downarrow & & \downarrow \\ & \mathbf{a}_p & \longrightarrow & \mathbf{g}_p & \longrightarrow & \mathbf{t}_p & \longrightarrow & \mathbf{p} \longrightarrow \hat{a}_p \end{array}$$

Here, we distinguish the vectors of additive genetic values ($\mathbf{a}_p$), total genetic values ($\mathbf{g}_p$), and underlying traits ($\mathbf{t}_p$) for the phenomic features in $\mathbf{p}$, as well as their vectors of non-additive genetic effects ($\mathbf{u}_p$), environmental effects ($\mathbf{e}_p$), and measurement errors ($\mathbf{m}_p$), from the corresponding values and effects underlying y . Additionally, various correlations among these effects are possible, the most likely ones of which are drawn with double-headed arrows between the additive genetic effects, non-additive genetic effects, and environmental effects. Thus, even though $f_p(\mathbf{p}) = \hat{a}_p$ doesn't involve y directly, $\mathbf{p}$ is caused by exogenous variables $\mathbf{u}_p$ and $\mathbf{e}_p$ that neither cause a nor are correlated with any causes of a , but are correlated with causes of y . This means that $\hat{a}_p y \mid a$, and thus $\text{cov}(\hat{a}_p, y) \neq \text{cov}(\hat{a}_p, a)$.

Finally, consider a GP model $f_g(\mathbf{x}) = \hat{a}_g$, with $\mathbf{x}$ a vector of genotypes for a candidate. We can expand the original path diagram to include the genetic features:

$$(2.12) \quad \begin{array}{ccccccc} & & & u & & e & & m \\ & & & \searrow & & \searrow & & \searrow \\ \hat{a}_g & \longleftarrow & \mathbf{x} & \longrightarrow & a & \longrightarrow & g & \longrightarrow & t & \longrightarrow & y \end{array}$$

Here, for simplicity we assume $\mathbf{x}$ captures all causal genetic variants, although in reality there are likely causal variants that affect a but do not contribute to $\hat{a}_g$. According to this path diagram, $\hat{a}_g \perp\!\!\!\perp y \mid a$ because there are no common exogenous variables impacting $\hat{a}_g$ and y that are not also shared with a . The only shared exogenous variable is $\mathbf{x}$ itself. Therefore, $\text{cov}(\hat{a}_g, a) = \text{cov}(\hat{a}_g, y)$. However, there are additional sources of variance in y that do not affect a (u , e , and m_y), so $\text{var}(y) > \text{var}(a)$. Thus, $\text{cor}(\hat{a}_g, a) \neq \text{cor}(\hat{a}_g, y)$.

Note that in this last diagram, we did not include a causal arrow from the genotypes ($\mathbf{x}$) to the non-additive genetic effects (u). Non-additive genetic effects are caused by genetic features. But additive genetic values and dominance genetic values are uncorrelated (73; 80; 79; 23; 56; 2; 1; 87).

If we use an additive encoding of these features that is orthogonal to dominance coefficients as in (80), and use a GP model that is linear in the genetic features, like rrBLUP (20), or Bayesian Alphabet models (33), then we can assume approximately that $\text{cov}(f_g(\mathbf{x}), u) = 0$. In contrast, this independence will not hold for GP models with non-additive encodings of genotype information such as RKHS or Deep Learning models.

2.4.3. Correction factors. In the case of an additive GP model, like rrBLUP, a simple analytical correction is possible to form a consistent estimator of ρ : Because $\text{cov}(\hat{a}_g, a) = \text{cov}(\hat{a}_g, y)$, ρ and ρ_y differ only due to the sources of variance in y that are not causes of a (*i.e.*, u , e , and m_y). Thus, $\rho = \rho_y \sqrt{\text{var}(y)} / \sqrt{\text{var}(a)}$, so r_y tends to *underestimate* ρ (Figure 2.2A). This is the basis for the estimator of ρ proposed by (49): $r = r_y / \sqrt{\widehat{h^2}}$, where $\widehat{h^2}$ is the estimate of the narrow-sense heritability of y .

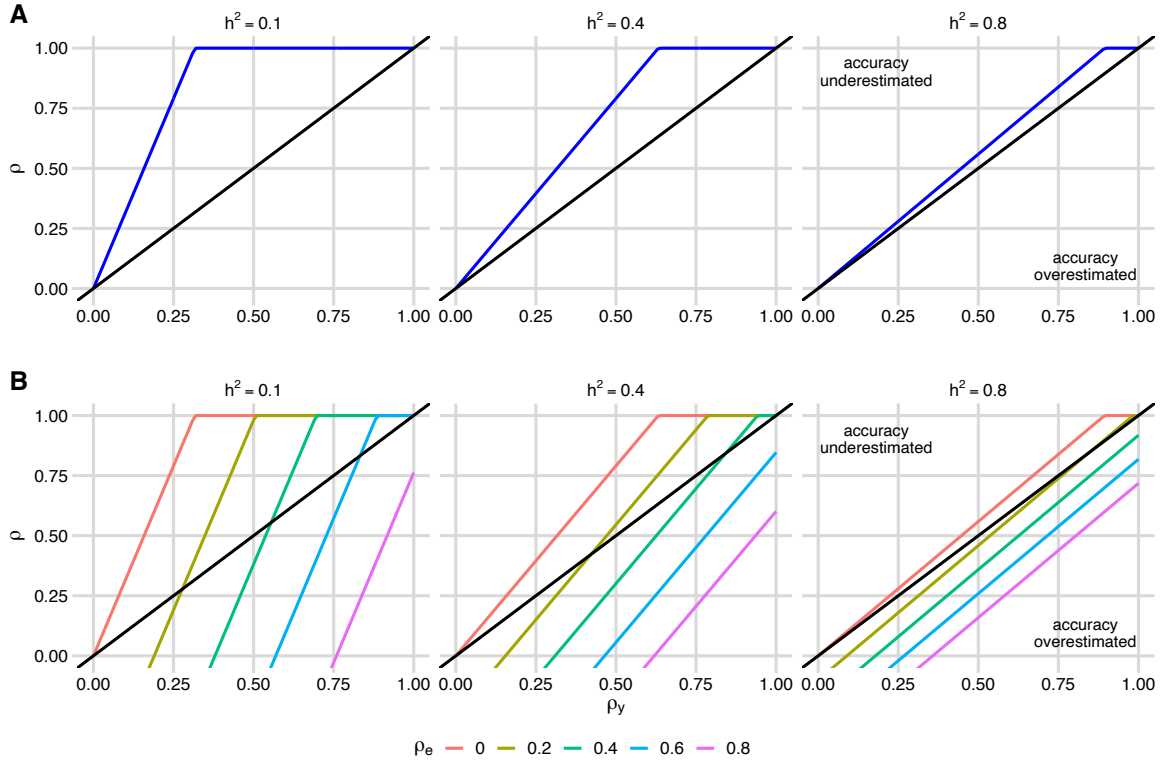


FIGURE 2.2. The relationships between *predictive ability* (ρ_y) and *accuracy* (ρ) for GP (a, blue lines) and PP models (b, colored lines) under 3 levels of heritabilities (0.1, 0.4, 0.8). The black lines indicate $y = x$. The areas above the black 1:1 lines indicate that *accuracies* are underestimated by predictive ability, whereas the areas below the blue lines indicate that *accuracies* are overestimated by predictive ability.

This correction factor is not valid for PP because $\text{cov}(\hat{a}_p, a) \neq \text{cov}(\hat{a}_p, y)$ as shown above. Instead, assuming the linear model $y = a + u + e + m_y$ and independent candidates, we can show (see Appendix A.1):

$$(2.13) \quad \rho = \frac{\rho_y - \rho_e \sqrt{1 - h^2}}{\sqrt{h^2}},$$

where $\rho_e = \text{cor}(\hat{a}_p, u + e + m_y)$. Even if phenomic features ($\mathbf{p}$) and the trait (y) are not measured on the same plots, so $\text{cov}(\hat{a}, e + m_y) = 0$, non-additive genetic variation (u) can still cause $\rho_e \neq 0$. The impact of ρ_e on the bias of $r = r_y / \sqrt{h^2}$ depends on the narrow-sense heritability (h^2), becoming stronger for low-heritability traits. Therefore, cross-validation-based calculations of r_y for a PP model can either underestimate *or overestimate* the actual *accuracy* (Figure 2.2B).

To demonstrate the impact of ρ_e on ρ , we highlight some examples (Figure 2.2)

- (1) If the heritability is low, e.g., $h^2 = 0.1$, and $\rho_y = 0.25$ for a GP model, its actual *accuracy* is much higher: $\rho = \rho_y / \sqrt{h^2} = 0.25 / \sqrt{0.1} = 0.79$. If instead $\rho_y = 0.25$ for a PP model, the actual *accuracy* could be as high as $\rho = 0.79$ if $\rho_e = \text{cor}(\hat{a}_p, u + e + m_y) = 0$. But this is unlikely unless phenotypes (y) and phenomic features $\mathbf{p}$ are measured on different plots and σ_u^2 is very small. However, ρ could be lower than ρ_y if $\text{cor}(\hat{a}_p, u + e + m_y) > 0.18$. To provide an intuition of ρ_e 's relationship with ρ , if phenomic features are direct estimates of the trait of interest (e.g., the trait of interest is plant height and drone images are used to measure plant height instead of a proxy trait), that is, $\mathbf{a} = a, \mathbf{u} = u, \mathbf{e} = e$ in Diagram 2.11, then $\frac{\rho}{\rho_e} = \sqrt{\frac{h^2}{1-h^2}}$. Equation 2.13 can be simplified to $\rho = \rho_y \sqrt{h^2}$ (see Appendix A.2), and ρ would be 0.079 for the PP model.
- (2) If the heritability is moderate, e.g., $h^2 = 0.4$, and $\rho_y = 0.5$ for a GP model, its actual *accuracy* is much higher: $\rho = \rho_y / \sqrt{h^2} = 0.5 / \sqrt{0.4} = 0.79$. If instead $\rho_y = 0.5$ for a PP model, the actual *accuracy* could be as high as $\rho = 0.79$, but would equal $\rho = 0.2$ if the correlations of predictions with a and $u + e + m_y$ are equal, and could be as low as $\rho = 0$ if the phenomic data is driven more by environmentally-sensitive traits (so ρ_e is larger). If phenomic features are direct estimates of the trait of interest, ρ would be 0.37 for the PP model.
- (3) If the heritability is high, e.g., $h^2 = 0.8$, differences between ρ_y and ρ are smaller for both GP and PP models. But even if ρ_y is very high for a PP model (e.g. $\rho_y = 0.8$), the actual

accuracy can be $\rho < 0.6$, and would be only $\rho = 0.72$ if phenomic features were direct estimates of the trait of interest.

While the difference between ρ_y and ρ have been known for many years (49; 58), many studies continue to report r_y values when comparing GP models. Despite r_y tending to underestimate the actual *accuracy*, comparing r_y values between two (additive) GP models (say model (1) is RRBLUP and model (2) is BayesC π) is valid *if the comparison is based on the ratio* $\frac{r_y^{(1)}}{r_y^{(2)}}$ (49):

$$\frac{r_y^{(1)}}{r_y^{(2)}} = \frac{r^{(1)}/\sqrt{\widehat{h^2}}}{r^{(2)}/\sqrt{\widehat{h^2}}} = \frac{r^{(1)}}{r^{(2)}}.$$

Therefore, the conclusion of the superiority of one model over the other would be the same whether r_y or r were used. The same is not true if models are compared based on the *difference* of r_y estimates:

$$\begin{aligned} r_y^{(1)} - r_y^{(2)} &= r^{(1)}/\sqrt{\widehat{h^2}} - r^{(2)}/\sqrt{\widehat{h^2}} \\ &= (r^{(1)} - r^{(2)})/\sqrt{\widehat{h^2}} \\ &\neq r^{(1)} - r^{(2)}, \end{aligned}$$

although the *sign* of the comparison remains the same. However, as shown in Figure 2.2, if a GP and PP model are compared either as a ratio (r_y^p/r_y^g) or a difference ($r_y^p - r_y^g$), both the sign and the magnitude of the result can differ from the true values (ρ^p/ρ^g or $\rho^p - \rho^g$, respectively). This means that conclusions about the superiority of PP over GP based on comparisons of *predictive ability* (ρ_y) are unreliable.

2.4.4. Alternative estimates of accuracy. (48) proposed an alternative estimator of *accuracy* called the "LR method". Unfortunately this method is not appropriate for PP models because the method assumes that $\text{cor}(\hat{a}, u + e) = 0$ which is generally violated for PP models, as discussed above.

(68) demonstrated three alternative measures of *accuracy* more appropriate for PP models, or models that combined phenomic and genomic features. Of these, the "parametric" approach is the most generally applicable, although it requires genetic or pedigree relationships among testing

individuals to compute. Specifically, a consistent estimator of ρ can be calculated as (68; 52):

$$(2.14) \quad r = \widehat{\text{cor}}_g(\hat{a}, y) \sqrt{\widehat{h}_a^2},$$

where $\widehat{\text{cor}}_g(\hat{a}, y)$ is an estimate of the genetic correlation between predictions ($\hat{a}$) and phenotypes (y) from a bivariate linear mixed model (see Appendix A.3) fit to testing candidates not used for model training, and $\widehat{h}_a^2$ is the estimated heritability of $\hat{a}$. Large sets of testing candidates are needed to estimate this parameter accurately. We used this method above to re-estimate the *accuracies* of PP models using data from three published studies because it is easier to estimate than equation 2.13.

2.5. Discussion

Our main goal with this article is to clearly demonstrate that comparing *predictive ability* estimates (r_y , the Pearson correlation coefficient between predicted genetic values and observed traits) between PP and GP models is not an appropriate way to assess the relative utility of phenomic selection and genomic selection for breeding programs.

As we show analytically and using previously published empirical results, the *predictive ability* is a biased and inconsistent estimate of the *accuracy* of genetic values (ρ), a key parameter of the Breeder's equation that determines the genetic gain after selection. While the *predictive ability* of an (additive) GP model is always less than its *accuracy*, the *predictive ability* of a PP model can be greater than its *accuracy*. For example, under reasonable, though simplified, conditions, a *predictive ability* of $\rho_y = 0.5$ for a GP model can indicate an *accuracy* of $\rho = 0.79$ (with $h^2 = 0.4$). But under the same conditions, a *predictive ability* of $\rho_y = 0.5$ for a PP model might indicate an *accuracy* of $\rho = 0.2$ (if the correlations of predictions with a and with $u + e + m_y$ are equal), or even lower (if predictions are dominated by non-genetic sources of variation). If clonal plants are grown in different plots, and phenomic data can be collected from one plot to predict the performance in another plot, the covariance between non-genetic components ($\mathbf{e}_p$ and e , Diagram 2.11) could be broken. However, the covariance between non-additive genetic components ($\mathbf{u}_p$ and u , Diagram 2.11) remains. Therefore, comparisons of *predictive abilities* between GP and PP models will tend to give the false impression that PP models are more accurate, even when they are not. Note that in real breeding programs, Equation 2.13, and the Breeder's equation itself, may not be strictly appropriate due to correlations among selection candidates and correlated errors due to fixed effects

and spatial effect terms in GP or PP models (59; 11). Equation 2.14 by (68) generally offers a better solution to estimate PP *accuracy*, which we have used in our results, but doesn't account for effects of correlations among candidates in calculations of selection intensity. Nevertheless, these simplified scenarios help intuitively demonstrate the reason that *predictive ability* fails as a fair comparison statistic between PP and GG models. We also presented examples of comparisons between PP and GP models from the recent literature, using the same datasets to directly estimate *accuracy* using an alternative method, and show that the conclusions about the superiority of one method over the other frequently change when the more appropriate statistic is used to compare models.

However, even if appropriate methods are used to estimate *accuracies* (ρ) for both GP and PP models, comparing these *accuracy estimates* directly is still not sufficient to compare genomic and phenomic selection breeding schemes. Breeding schemes should be evaluated based on their expected genetic gains over a specific time interval, and using a specific budget. Comparisons between estimates of ρ alone are insufficient for this because genomic and phenomic selection schemes are likely to impact the other three parameters of the Breeder's equation differently, which can overcome differences in accuracy. The other three parameters of the breeder's equation are L , the cycle length, σ_A , the additive genetic standard deviation, and i , the standardized selection intensity.

The most consequential parameter in the Breeder's equation is generally L , the cycle length. In many breeding programs, each cycle takes multiple years. For example, in (10; 77) (wheat), the cycle is $\approx 6-7$ years. In the University of California strawberry breeding program, we estimated the average cycle length to be 10 years (61). In its most aggressive form, genomic selection can be used to shorten breeding cycles to $\ll 1$ year, using the two-part GS scheme proposed by (31), in which candidates are selected based on seed or seedling genotypes without any phenotypic evaluation. This strategy is generally not possible for phenomic selection because most PP models require plants to be grown to mature stages to collect phenomic data. Most apparently successful PP models for yield and quality-related traits use features collected in field trials of candidates close in time to when yield measures are taken (85; 47). Applied this way, phenomic selection would not shorten cycle lengths significantly. Thus, a genomic selection scheme like that proposed by (31) would be preferred over a phenomic selection scheme even if the GP *accuracy* was lower than the PP *accuracy*.

The aggressive two-part GS scheme proposed by (31) is not currently widely used. More commonly, GS only slightly shortens breeding cycles, for example, using GP to select at the F5 or PYT stages of a wheat breeding scheme instead of the AYT stage (77). In this case, differences in L may not be large between a genomic selection and phenomic selection scheme. However, one of the principle arguments for phenomic selection is that is that phenomic data can be collected more cheaply and in higher throughput than genomic data. If so, and if each field plot is sufficiently inexpensive, PPs could be made on much larger numbers of candidates, enabling breeders to increase selection intensity. The relationship between population size and selection intensity can be approximated as:

$$i \approx \varphi(\Phi^{-1}(1 - k/n))(n/k)$$

where n is the number of candidates, k is the target number of selections, and $\varphi(\cdot)$ is the density function and $\Phi^{-1}(\cdot)$ is the inverse of the cumulative density function of the standard normal distribution. Figure 2.3 shows that i is a function of both n and k/n , the proportion selected. Larger n increases i but at a relatively slow rate, as discussed by (18). Increasing the number of candidates from GP to PP two-fold (*i.e.*, 100%), from $n_g = 100$ for a genomic selection scheme to $n_p = 200$ for a phenomic selection scheme, only increases selection intensity $\approx 25\%$. Increasing the size of the phenomic selection scheme 10-fold (*i.e.*, 900%) to $n = 1,000$ increases intensity $\approx 70\%$ (Figure 2.3). Both numbers decreases if the genomic selection scheme is larger (*e.g.* $n_g = 1,000$, compared to phenomic selection schemes of $n_p = 2,000$ or $n_p = 10,000$, respectively). Therefore, exponentially larger numbers of candidates must be evaluated for the increase in i enabled by phenomic selection to compensate for decreased *accuracy*. Such massive population sizes are often infeasible. Population sizes can be limited by field space and field costs, so it may not be common for the number of selection candidates to be significantly limited when applying genomic selection. Therefore, phenomic selection schemes are unlikely to have a major advantage over genomic selection solely based on increased intensity.

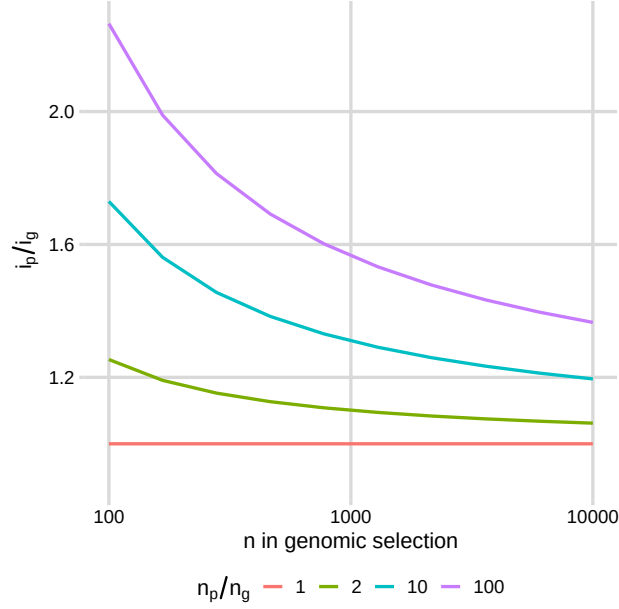


FIGURE 2.3. Relationship between selection intensity and population size.

The effects of genomic selection and phenomic selection on the maintenance of genetic variance σ_A^2 are less clear than for the other parameters of the Breeder's equation. (41) and (34) argued that genomic selection should maintain genetic variance longer than pedigree-based selection because genetic markers can be used to select individuals with good alleles at QTL but which are otherwise unrelated. However, others have argued that genomic selection can rapidly reduce genetic variance, and multiple strategies have been suggested to mitigate against rapid genetic variance loss during genomic selections. Phenomic selection is really a form of phenotypic selection, thus the consequences of phenomic selection on genetic variance are likely similar. In either strategy, careful management of genetic variance will be necessary to maintain long-term gains.

Lastly, comparisons of PP and GP predictive ability generally use training datasets that are the same size in an attempt to make the comparison fair. However, this choice itself likely biases the comparison between these approaches. Even if r could be estimated accurately, the actual *accuracy* using either strategy will depend on the size and composition of the training data for each model. Since phenomic data are generally cheaper to acquire than genomic data, an optimized phenomic selection scheme would likely permit larger training data sets than an optimized genomic selection scheme. Therefore, PP and GP models should be compared only when each model is trained on an appropriately sized training dataset.

As a final note, our discussions throughout assume that the Breeder’s equation is sufficient to predict genetic gain, and thus that accuracy (ρ) is proportional to the rate of genetic gain. But, as discussed by (11), this is only exact when the selection index (here: $\hat{a}$) and breeding values are linearly related (*e.g.*, bivariate normal). In cases of extreme departures from this assumption, it is mathematically possible for a model with a lower accuracy (ρ) to lead to faster gain than a model with higher accuracy, controlling for i , σ_a , and L . In these cases, the Price Equation, or Robertson–Price identity, will better predict the rate of genetic gain. Nevertheless, these conditions are likely fairly unusual, and we believe that if used carefully, accuracy can still be a useful statistic for studying changes to breeding schemes.

Our goal is not to suggest that phenomic selection cannot be useful for breeding. When PP is used to increase the precision or throughput of phenomic evaluation of candidates, this can certainly improve breeding outcomes. In particular, if PP is used to improve the phenotype data available to train GP models, this can greatly improve breeding outcomes. In this application, it is actually the *predictive ability* of a PP model that matters - the ability of that model to produce high-accuracy replacements for the trait data needed to train an accurate GP model. Therefore, we suggest that PP studies should report *predictive ability*, while GP studies should report *accuracy*. But these two numbers should not be directly compared - they measure different things.

2.6. Conclusions

To summarize, we argue that comparisons between *predictive abilities* of GP and PP models are largely uninformative for breeders. These comparisons use a statistic that is not directly relevant to breeding goals, and neglect the impact of each strategy on other equally important consequences of the implementation of genomic and phenomic selection schemes.

2.7. Data Availability

Scripts used in this study are available at https://github.com/Faye-Wong-stat/breeder_equation

2.8. Acknowledgements

The authors would like to acknowledge helpful comments from Hans-Peter Piepho and feedback from Renaud Rincent and Zachary Winn. UC Davis is located at Davis, California, which is located on land that was the home of the Patwin people for thousands of years.

Evaluating the Performance of a Two-step Prediction Model by Incorporating Phenomic Data into Genomic Prediction for Strawberry Fruit Quality Traits

3.1. abstract

Genomic prediction has accelerated genetic improvement in plant breeding programs, but its effectiveness remains constrained by the availability of high-quality phenotypic data. High-throughput phenotyping technologies, such as near-infrared spectroscopy (NIRS), provide a potential solution through rapid and non-destructive collection of phenomic data. In this study, we evaluated multiple strategies for incorporating NIRS data into genomic prediction models for strawberry fruit quality traits, including Brix and titratable acidity. We compared univariate and multivariate genomic prediction models, models incorporating NIRS data as secondary traits, and a two-step framework in which NIRS data were first used to predict phenotypes, which were subsequently incorporated into genomic prediction models.

The benefit of incorporating NIRS data as secondary traits was trait-dependent and inconsistent across years. Prediction accuracies improved for titratable acidity in both years and for several traits in 2023, but decreased for Brix in 2022. Fourier transformation of NIRS data provided little overall improvement in predictive performance. In contrast, the two-step framework produced more consistent improvements, particularly for titratable acidity and Brix, suggesting that NIRS data are more effective when treated as indirect phenotypes rather than auxiliary secondary traits.

We further examined the relationship between phenomic feature heritability and variable importance and observed no apparent association between the two, demonstrating that highly heritable phenomic features are not necessarily the most informative predictors. Together, these results clarify the role of phenomic data in genomic prediction and support the use of high-throughput phenotyping as an indirect phenotyping strategy for traits that are costly or labor-intensive to measure directly.

3.2. Introduction

The rate of genetic gain in breeding programs is increasingly limited by the ability to accurately and efficiently measure phenotypes. In the breeder’s equation, $\Delta g = i\rho\sigma_A/L$, the genetic gain, Δg , is proportional to selection intensity (i), prediction accuracy (ρ), additive genetic standard deviation (σ_A), and is inversely proportional to generation interval (L) (24). When we increase the number of phenotyped individuals, we increase the number of potential candidates that can be selected in the breeding program. Genetic gain increases linearly with the increase of selection intensity, but logarithmically with the increase of the number of phenotyped individuals in genomic selection (69). While advances in high-throughput genotyping have made it possible to evaluate thousands of individuals at relatively low cost, phenotyping remains labor-intensive, time-consuming, and often destructive. As a result, the phenotyping has been widely recognized as a major bottleneck in crop improvement (29; 43).

Genomic prediction (55) was developed to address some of these limitations by enabling selection on predicted genetic values rather than directly observed phenotypes. By using genome-wide markers, genomic prediction can reduce generation intervals, increases selection intensity, and can improve selection accuracy. However, despite these advantages, genomic prediction fundamentally depends on the availability of high-quality phenotypic data for model training. Consequently, it alleviates, but does not eliminate, the phenotyping bottleneck.

Recent advances in high-throughput phenotyping technologies have introduced phenomic data as a potential solution to this challenge. Techniques such as near-infrared spectroscopy (NIRS) allow rapid, non-destructive, and relatively low-cost measurement of plant tissues. These phenomic measurements have been used to construct relationship matrices analogous to genomic relationship matrices, giving rise to “phenomic prediction” as an alternative to genomic prediction (65; 47; 30; 13; 85; 54). In some cases, phenomic prediction has been reported to achieve predictive abilities comparable to or exceeding those of genomic prediction (65; 47; 85).

However, the interpretation and optimal use of phenomic prediction remain uncertain. Direct comparisons between phenomic and genomic prediction can be misleading, as the two approaches rely on fundamentally different sources of information and may capture distinct components of phenotypic variation with different values in breeding programs (81). Rather than viewing phenomic data as a replacement for genomic data, an alternative perspective is to treat phenomic measurements

as indirect (and cheaper) observations of target traits (43). Under this view, phenomic data can be used to "measure" phenotypes for individuals that have not been directly measured, thereby increasing the effective size of the phenotyped population (43; 82).

This perspective motivates a two-step prediction model in which phenomic data are first used to predict phenotypes, and these predicted phenotypes are subsequently incorporated into genomic prediction models. Conceptually, this approach separates the problem of phenotype measurement from that of genetic prediction, using phenomic data as a high-throughput surrogate for traditional phenotyping. Compared to a genomic prediction model, incorporating phenomic data potentially captures non-additive genetic effects, environment effects, and G by E effects, increasing predictive ability and improving predictions conditional on the environment. In contrast to single-step models that incorporate phenomic data as secondary traits, the two-step framework can potentially extract genetic information from improved phenotype estimates.

A key challenge in using phenomic data lies in understanding which features contribute meaningfully to prediction. A common assumption is that phenomic features with high heritability are more informative for predicting target traits. If a phenomic feature is highly related to the trait of interest, we expect that it would be predictable of the trait. The covariance between a phenomic feature and a target trait can be partitioned into genetic and non-genetic covariance. The heritability of a phenomic feature is related to the genetic covariance between a feature and a trait. However, they do not have a linear or monotone relationship.

In this study, we evaluate multiple strategies for incorporating phenomic data into genomic prediction for strawberry fruit quality traits. Specifically, we compare univariate and multivariate genomic prediction models, models that include NIRS data as secondary traits, and a two-step genomic prediction model that uses NIRS data to generate predicted phenotypes. Our objectives are to (i) assess the extent to which NIRS data can alleviate the phenotyping bottleneck, (ii) determine whether a two-step modeling approach improves prediction accuracy relative to single-step methods, and (iii) provide insight into the relationship between NIRS feature properties and their utility in prediction. Through this work, we aim to clarify the role of phenomic data in breeding programs and provide guidance for their effective integration into genomic prediction pipelines.

3.3. Materials and Methods

3.3.1. Genetic, Fruit Quality, and NIRS Data. 2916 strawberry accessions were genotyped using a 50K SNP array in 2022. 3757 accessions were genotyped with a 1650 SNP array and the 50K SNP were imputed in 2023. There were 6 common genotypes from both years. A genomic relationship matrix for each year was calculated. Fruit quality traits, including titratable acidity, brix, anthocyanin, ascorbic acid, and firmness data were measured on 2952 unique genotypes in 2022. Multiple measurements were taken for some genotypes. In 2023, the same fruit quality traits were measured on 3561 unique genotypes over 2 harvest periods. Multiple measurements were taken for some genotypes during the same harvest period. Near-infrared spectroscopy (NIRS) data (a type of phenomic data) were collected using a MicroNIR OnSite-W. 3803 samples from 2929 unique genotypes were collected in 2022 and 3546 samples from 3165 unique genotypes were measured for NIRS over 2 harvest periods in 2023. We performed Multiplicative Scatter Correction as a pre-processing step for the NIRS data to remove undesirable scattering of light in the samples (32; 66).

3.3.2. Fourier Transform and BLUEs Calculation. We hypothesized that Fourier transform on NIRS data would increase prediction accuracy, because it is suited for wave data analysis by decomposing complex wave information into subcomponents (12). We performed Fourier transforms on the 3561 and 3546 NIRS samples from 2022 and 2023, respectively. We calculated the best linear unbiased estimators (BLUEs) on fruit quality traits, NIRS, and Fourier-transformed data for each genotype from 2022. We calculated the BLUEs for each genotype from 2023, using harvest time as a random effect. In total, 2929 genotypes from 2022 and 3163 genotypes from 2023 were measured for both fruit quality traits and NIRS data and used in subsequent analysis.

3.3.3. Models.

3.3.3.1. Univariate Linear Genomic Prediction. The univariate linear mixed model is widely used in genomic prediction, and it will be used as a base comparison for the following models. Briefly, we fitted a linear mixed model, defined as:

$$(3.1) \quad \mathbf{y}_j = \mathbf{g}_j + \boldsymbol{\epsilon}_j,$$

where $\mathbf{y}_j = [y_{1j}, y_{2j}, \dots, y_{nj}]^T$, $\mathbf{g}_j = [g_{1j}, g_{2j}, \dots, g_{nj}]^T \sim N(\mathbf{0}, \sigma_{gj}^2 \mathbf{K}_j)$, and $\boldsymbol{\epsilon}_j = [\epsilon_{1j}, \epsilon_{2j}, \dots, \epsilon_{nj}]^T \stackrel{\text{iid}}{\sim} N(\mathbf{0}, \sigma_{\epsilon j}^2 \mathbf{I})$ was the BLUEs, genetic, and environmental values of the j th fruit quality trait for n genotypes. $\mathbf{K}_j$ is the genomic relationship matrix for the j th trait. This model was fit independently for ...

3.3.3.2. *Multivariate Linear Genomic Prediction.* We used the following for multivariate linear genomic prediction models:

$$(3.2) \quad \mathbf{Y} = \mathbf{G} + \mathbf{E},$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_j, \dots, \mathbf{y}_t]$ were the BLUEs of t fruit quality traits, and $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_j, \dots, \mathbf{g}_t] \sim \text{MN}(\mathbf{0}, \mathbf{K}, \mathbf{H})$, and $\mathbf{E} = [\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \dots, \boldsymbol{\epsilon}_j, \dots, \boldsymbol{\epsilon}_t] \sim \text{MN}(\mathbf{0}, \mathbf{I}, \mathbf{R})$ were the genetic and environmental values of t fruit quality traits. $j \in \{1, 2, \dots, t\}$.

3.3.3.3. *Univariate Linear Genomic Prediction with NIRS data.* We fitted a similar multivariate normal model, for each target trait at a time, with NIRS data as secondary traits:

$$(3.3) \quad \mathbf{W} = \mathbf{G} + \mathbf{E},$$

where $\mathbf{W} = [\mathbf{y}_j, \mathbf{P}] = [\mathbf{y}_j, \mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_l]$, and $\mathbf{y}_j$ was the the BLUEs of j th trait and $\mathbf{P}$ were the BLUEs of NIRS data across l wavelenghtes. $\mathbf{G}$ and $\mathbf{E}$ followed the same multivariate normal distributions in Equation 3.2. This model is univariate in the sense that the model is fitted for each trait of interest at a time, but mathmatically it is a multivariate linear normal model as there are multiple response variables. The same model was used with Fourier-transformed NIRS data instead of NIRS data.

3.3.3.4. *Multivariate Linear Genomic Prediction with NIRS data.* We fitted the following multivariate normal model for multivariate linear genomic prediction with NIRS data:

$$(3.4) \quad \mathbf{W} = \mathbf{G} + \mathbf{E},$$

where $\mathbf{W} = [\mathbf{Y}, \mathbf{P}] = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_j, \dots, \mathbf{y}_t, \mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_k, \dots, \mathbf{p}_l]$, and similarly, $\mathbf{Y}$ were the BLUEs of t traits and $\mathbf{P}$ were the BLUEs of NIRS data. Again, the same model was used with Fourier-transformed NIRS data instead of NIRS data. We hypothesize that the Multivariate Linear Model with NIRS Data will achieve the higher prediction accuracy compared to the Univariate Linear Model. To understand if the high accuracy was caused by modeling the relationships among response variable or by incorporating NIRS data, we will also evaluate the *Multivariate Linear Model* and the *Univariate Linear Model with NIRS Data*.

3.3.3.5. *Two-step Genomic Prediction.* The BLUEs of each genotype for the data from year 2022 were used in the two-step genomic prediction. However, for the data from year 2023, the BLUEs of each genotype for each harvest were calculated and used for the two-step genomic prediction. The first step of a two-step genomic prediction model was to fit the following univariate models for each fruit quality trait:

$$(3.5) \quad \mathbf{y}_j = f(\mathbf{P}),$$

where $\mathbf{y}_j = [y_{1j}, y_{2j}, \dots, y_{nj}]^T$ was the BLUEs of j th trait for n genotypes given a particular harvest, and $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_l]$ was the BLUEs for n genotypes for both harvests across l wavelengths, for the data from year 2023. For the year 2022, the same model was used on the BLUEs of each genotype. We also used Fourier-transformed NIRS data in $\mathbf{P}$.

We used 5 models for $f(\cdot)$, including partial least square (PLS), support vector machine (SVM) with a linear kernel, support vector machine with a radial basis function kernel (RBF), random forest, and neural network. We hypothesized that the non-linear models would better capture the non-linear and non-additive relationship between fruit quality traits and NIRS data. We calculated the BLUEs of the predictive phenotypic values for each genotype for data from 2022. We also calculated the BLUEs for each genotype using the harvest as a random effect for data from 2023. In the second step, we used the observed ($\mathbf{y}_j$) and predicted ($\hat{\mathbf{y}}_j$) phenotypic values ($\hat{\mathbf{Y}}_j$) of trait j to build bivariate genomic prediction models:

$$(3.6) \quad \hat{\mathbf{Y}}_j = \mathbf{G}_j + \mathbf{E}_j,$$

where $\hat{\mathbf{Y}} = [\mathbf{y}_j, \hat{\mathbf{y}}_j]$ is the observed and predicted phenotypic values for fruit quality trait j . We considered the observed and predicted values of the same trait as two traits, because we assumed that the genetic relationship between them is not 1.

For multivariate genomic prediction, we used the following:

$$(3.7) \quad \hat{\mathbf{Y}} = \mathbf{G} + \mathbf{E},$$

where $\hat{\mathbf{Y}} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_j, \dots, \mathbf{y}_t, \hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_j, \dots, \hat{\mathbf{y}}_t]$. The four previous models (univariate, multivariate with NIRS data, etc.) are restricted that they can only model the linear relationships between genotypes and phenotypes, and between traits of interest and NIRS data. $f(\cdot)$ in the two-step genomic prediction model is more flexible and is able to model complex, non-linear relationships. We also hypothesize that the non-linearity and non-additivity of $f(\cdot)$ can better predict the phenotypes used for genomic prediction in the second step.

3.3.4. Cross-Validation. We used a five-fold cross-validation to evaluate model accuracies. The basic idea of cross-validation is to separate the model fitting and tuning process from the model evaluation process by using separate datasets for each (38). This ensures that the trained models can be generalized for unseen data instead of being overfitted to the training data.

We calculated the *predictive abilities* r_y of models as Pearson's correlation between predicted additive genetic values and observed phenotypic values: $r_y = \text{cor}(\hat{\mathbf{g}}, \mathbf{y})$. The univariate and multivariate genomic prediction *accuracies* r were estimated as in

$$(3.8) \quad r = r_y / \sqrt{\widehat{h^2}},$$

where $\widehat{h^2}$ is the estimated heritability of the trait. For both the univariate and multivariate genomic prediction models with NIRS data or with Fourier-transformed NIRS data (Equations 3.3 and 3.4), we estimated the model *accuracies* r as following (68; 52; 81):

$$(3.9) \quad r = \widehat{\text{cor}}_g(\hat{\mathbf{g}}, \mathbf{y}) \sqrt{\widehat{h_{\hat{\mathbf{g}}}^2}},$$

where $\widehat{\text{cor}}_g(\hat{\mathbf{g}}, \mathbf{y})$ is the estimated genetic correlation between predictions ($\hat{\mathbf{g}}$) and phenotypes ($\mathbf{y}$) from a bivariate linear mixed model, and $\widehat{h}_{\hat{\mathbf{g}}}^2$ is the estimated heritability of $\hat{\mathbf{g}}$.

We randomly divide each year’s genotypes into 5 folds. Four folds of the genotypes, along with their corresponding BLUEs of phenotypes were used for training the Univariate and Multivariate Linear Genomic Prediction models (Equation 3.1 and 3.2). Narrow-sense heritability $\widehat{h}^2$ for each trait was calculated by fitting a new model on the remaining testing fold. Predicted additive genetic values ($\hat{g}_j$) of the j th trait on the testing fold were calculated from the trained models. This was repeated for all five folds.

For both the Univariate and Multivariate genomic prediction models with NIRS data or with Fourier-transformed NIRS data (Equations 3.3 and 3.4), we used the same five-fold cross-validation to calculate model accuracy. All five folds of fruit quality traits, NIRS data (or Fourier-transformed NIRS data), and their corresponding genotypes were used for training genomic prediction models. However, the phenotypic values of fruit quality traits from the testing fold were masked (assigned NA values) during training, and predictions were made on them.

For the two-step genomic prediction, the same four folds of fruit quality and NIRS data (or Fourier-transformed NIRS data) were used to train the models in Equation 3.5. Predicted phenotypic values ($\hat{y}_j$) of the j th fruit quality trait from all five folds were calculated from the trained models. A bivariate genomic prediction model was trained on the fruit quality traits and the predicted fruit quality traits on all five folds with the testing fold of fruit quality traits masked. This cross-validation method, which masks the phenotype values of the testing fold, is referred to as CV2 in the literature (16; 68; 27).

3.3.5. Variable Importance in Phenomic Prediction and Its Relationship with h^2 .

To examine the association between variable importances and phenomic feature heritabilities in phenomic prediction, we divided the NIRS wavelengths into 10 neighboring groups. The reason for this is that the continuous and adjacent wavelengths are highly correlated and share similar information (40). We built a full phenomic prediction model using ridge regression with all of the wavelengths and calculated its R^2 value. We then built a phenomic prediction model omitting each of the 10 wavelength groups and calculated the diminished R^2 value. The difference between the R^2 values would be used as the metric for variable importance for the corresponding group. We also calculated the average heritability for the wavelengths within each group.

3.3.6. Software. We used *PolyBreedR* (22) for imputation of the 50K SNP markers for accession in 2023. We performed Fourier transform using *mvfft()* function from base *R* (64), and we calculated BLUEs using *lme4* and *emmeans* packages (3; 51). We calculated genomic relationship matrices using *A.mat()* function from *rrBLUP* package (21). We used *mix.solve()* function from *rrBLUP* package to perform univariate genomic prediction. For multivariate genomic prediction, univariate genomic prediction with NIRS data, and multivariate genomic prediction with NIRS data, we used *MCMCglmm* and *MegaLMM* package in *R* (36; 70). We evaluated Equation 3.9 with *estimate_gcor()* from

3.4. Results and Discussion

We evaluated multiple ways to incorporate NIRS data into prediction models to improve selection accuracy. We treated NIRS data as secondary traits in genomic prediction models to predict fruit quality traits either one-at-a-time (Equation 3.3) or simultaneously (Equation 3.4). For comparison, we also built univariate (Equation 3.1) and multivariate (Equation 3.2) genomic prediction models using the same fruit quality traits. We expected that incorporating NIRS data would increase accuracy by utilizing the relationships among traits, compared to the models without NIRS data. Because Fourier transform is suited for wave data (12), we examined the models with post-Fourier-transformed NIRS data as secondary traits as well. All the above models are referred to as "single-step" models. We also evaluated a "two-step" genomic prediction model, which first modeled the relationship between fruit quality traits and NIRS data, and then used the predicted outcomes as a secondary trait in a bivariate genomic prediction model. We expected the "two-step" genomic prediction models would have higher accuracies compared to the "single-step" models, by modeling the complex relationship between NIRS data and fruit quality traits.

Overall, we observed several patterns: i) using NIRS data as secondary traits in genomic prediction provides trait-specific benefit; ii) the two-step approach offers a more consistent benefit; iii) feature importance does not have a linear relationship with feature heritability.

The benefit of using multivariate vs. univariate, of using NIRS data as secondary traits, and of using Fourier transformed NIRS data as secondary traits was case-specific (Figure 3.1). The benefit of multivariate genomic prediction cannot be generalized to all fruit quality traits and to all years. It is possible likely due to the low correlation among the fruit quality traits ($-0.42 \sim 0.17$ in

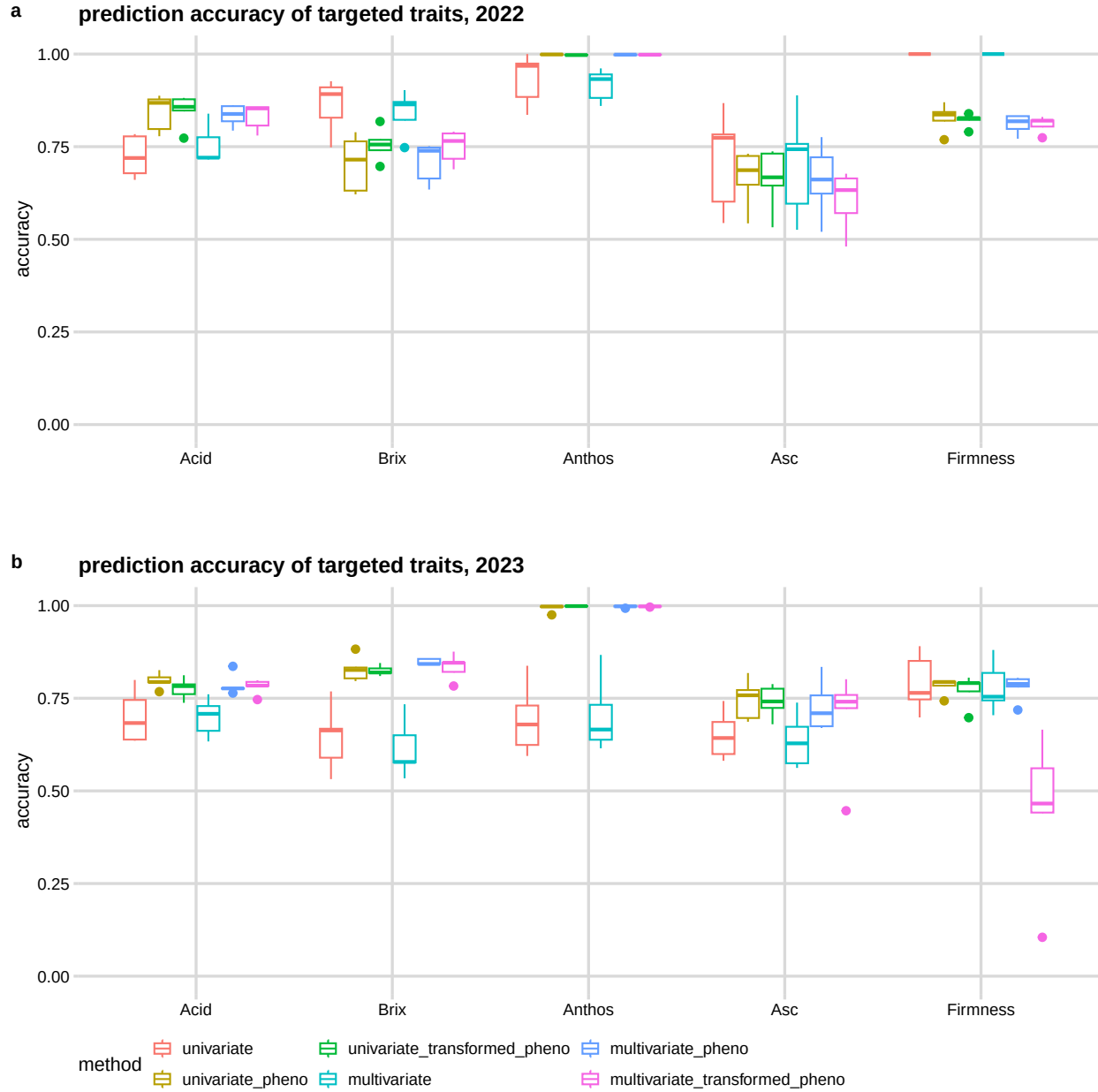


FIGURE 3.1. *Boxplots of prediction accuracies of target fruit quality traits (x-axis). Panels are separated by year. "Univariate" indicates that each fruit quality trait was modeled individually, whereas "multivariate" indicates that all five fruit quality traits were modeled in a multivariate model. "Pheno" indicates that NIRS data were included in the model as secondary traits. "Transformed" means that Fourier transform was performed on NIRS data first before being included in the models.*

2022, $-0.21 \sim 0.45$ in 2023) The incorporation of NIRS data (whether as individual wavelengths or post-Fourier-tranformation) increased prediction accuracies for acidity in both years, and for Brix and ascorbic acid in 2023. However, both decreased prediction accuracy for Brix in 2022. The

prediction accuracies of anthocyanin were close to 1 when using NIRS data as secondary traits. This may reflect instability in the estimation of residual variance and genetic correlation when the residual variance approaches zero. In some folds for firmness (in 2022), the accuracies were estimated to be over 1 by Equation 3.8. This could happen if the estimated heritability was small and/or the correlation between predicted breeding values and observed phenotypic values was large (48). Prediction accuracies greater than 1 are not theoretically meaningful. Therefore, accuracies were truncated at 1.

We calculated the predictive abilities of predicted phenotypic values from the first step (Equation 3.5, Figure 3.2) of the "two-step" genomic predictions. We believed that various machine learning models can capture the complex, non-linear and non-additive relationship between fruit quality traits and NIRS data. If correct, we should see high predictive abilities of predicted phenotypic values in this first step. For Brix, most of the models achieved predictive abilities close to 1 in both years. This is not surprising because Brix itself is measured using a light refraction-based assay, with similar physical properties to NIRS. For acidity, anthocyanin, and ascorbic acid, both PLS and SVM models had higher predictive abilities (in both years, for acidity with or without Fourier transform, for the others only without Fourier transform). Contrary to our expectation, nonlinear models did not consistently outperform linear methods, suggesting that the associations between NIRS data and fruit quality traits are predominantly linear.

Fourier transform on NIRS data was expected to increase prediction accuracy, because it is suited for wave data analysis by decomposing complex wave information into subcomponents (12). However, transforming the NIRS data provided little benefit in accuracy in almost all models when incorporating NIRS data as secondary traits, with the exception of firmness in 2023 (Figure 3.1), in which Fourier transform decreased model accuracy. We also did not observe an overall benefit of Fourier transform in predicting fruit quality traits from NIRS data in either year (Figure 3.2). These results suggest that the predictive information in the NIRS spectra was already adequately represented in the original wavelength space, and that frequency decomposition did not reveal additional biologically relevant structure.

NIRS data may act as a proxy for certain traits through a two-step genomic prediction model. Overall, the accuracies from the two-step genomic prediction models were higher than or similar to those from the single-step multivariate genomic prediction models with NIRS data as secondary

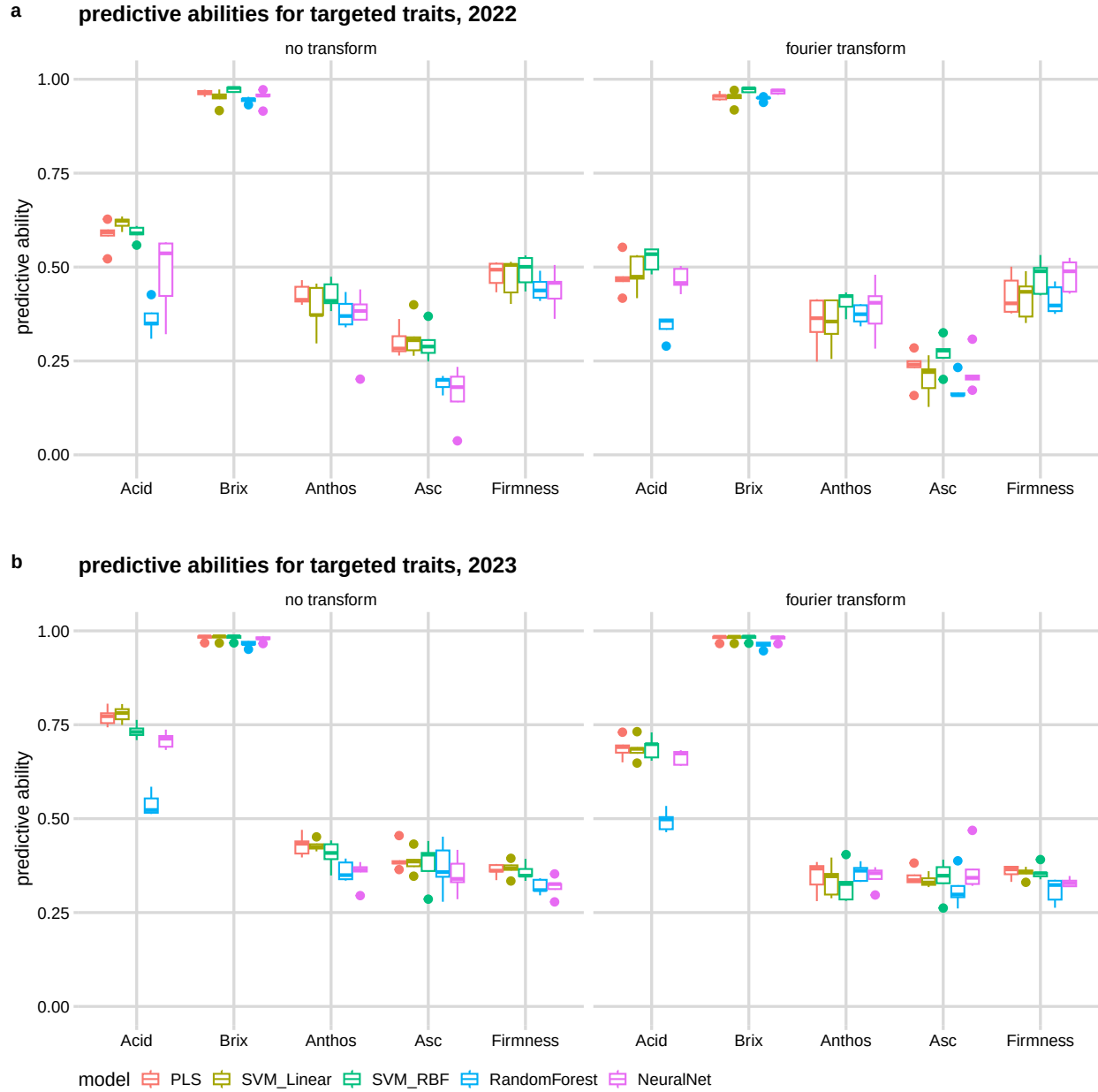


FIGURE 3.2. Boxplots of predictive abilities of target fruit quality traits (x-axis) were made with a five-fold cross-validation. Panels are separated by year. Different colors represented different models. "PLS" represents partial least square. "SVM_Linear" represents support vector machine with a linear kernel, whereas "SVM_RBF" represents support vector machine with a radial basis function kernel.

traits, with the exception of acidity in 2022 (Figure 3.3). We also observed a similar result in the "two-step" genomic prediction models using Fourier transformed NIRS data (Figure C.1). We observed the greatest increase in accuracy for acidity in 2023 and Brix in both years (Figure 3.3).

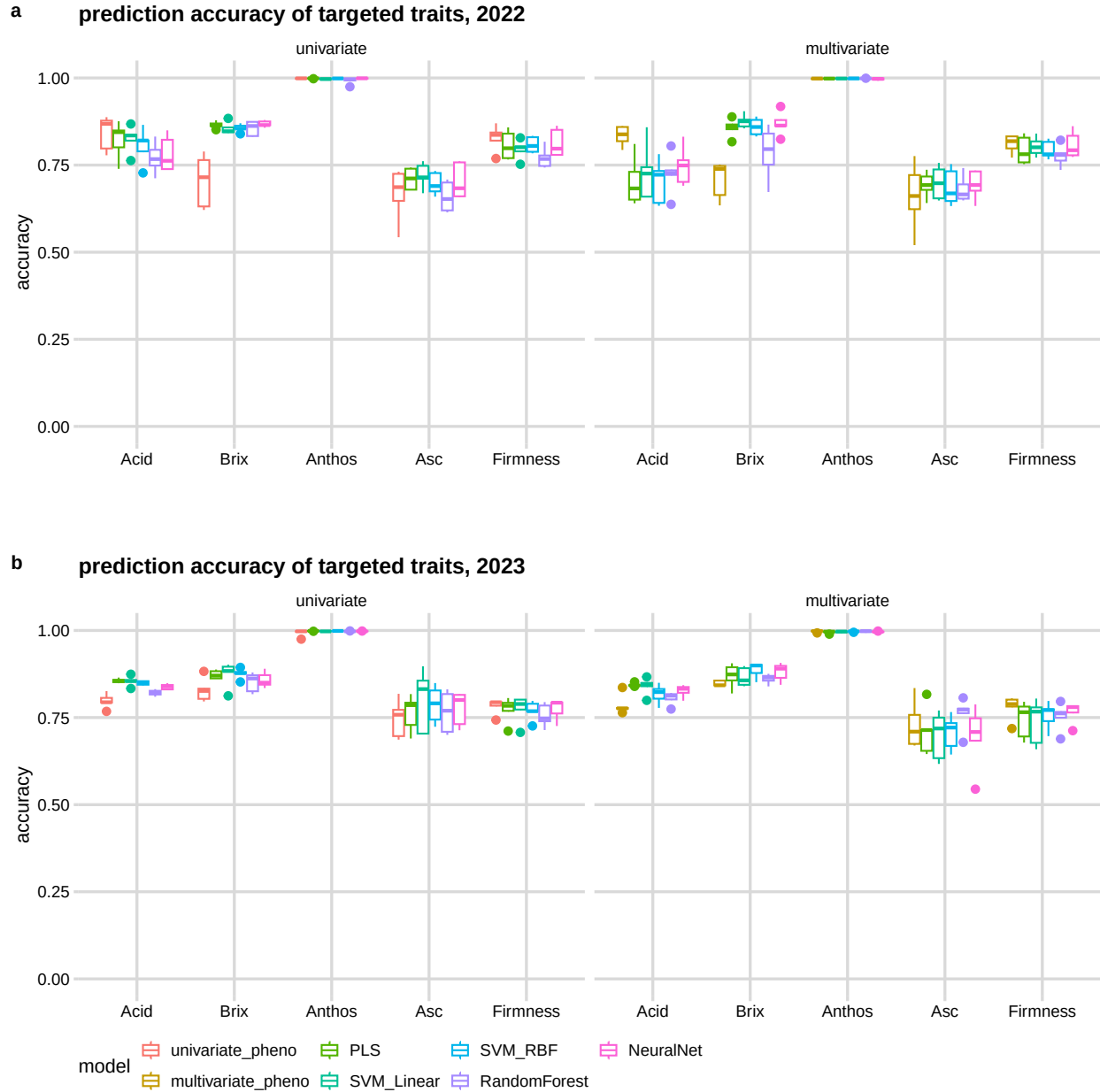


FIGURE 3.3. Boxplots of prediction accuracies of target fruit quality traits (x-axis) were made with a five-fold cross-validation. Panels are separated by year. "univariate_pheno" and "multivariate_pheno" boxes were calculated from single step genomic prediction models using NIRS data as secondary traits, which were plotted in Figure 3.1. They are plotted here again for comparison between single step and two-step genomic prediction models.

These results support the interpretation of NIRS data as indirect phenotypes rather than simply auxiliary covariates in genomic prediction models. By first estimating phenotypes from spectral

data and subsequently incorporating these estimates into genomic prediction models, the two-step framework more effectively leveraged the information contained in NIRS data.

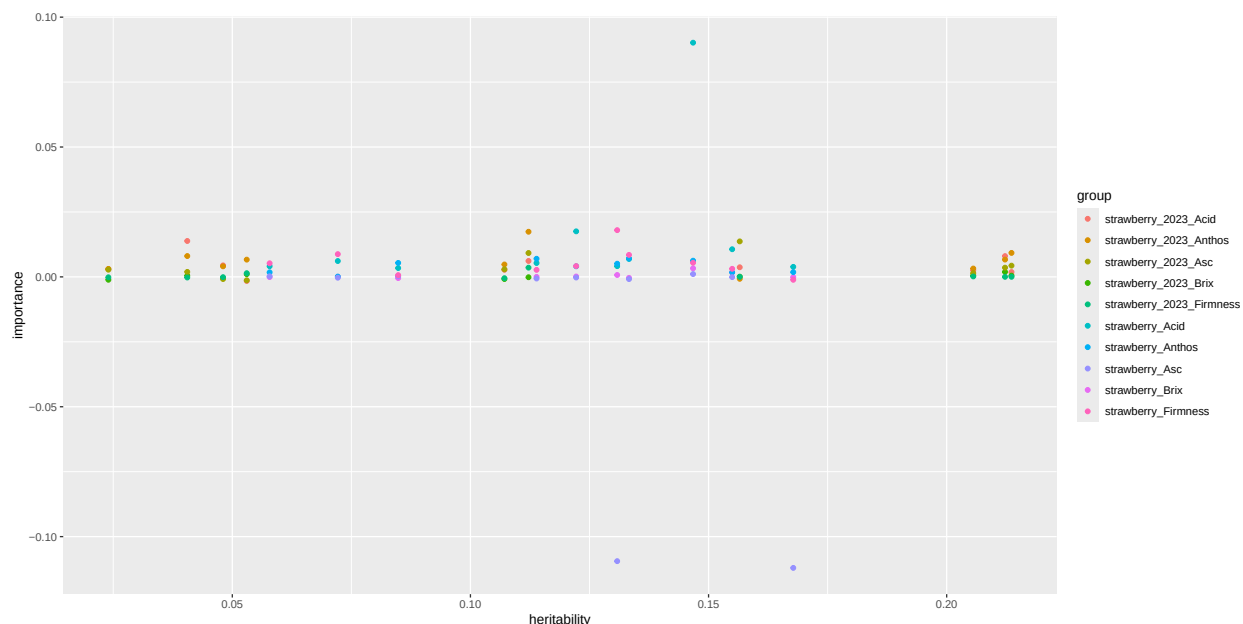


FIGURE 3.4. Scatter plot of phenomic feature heritability and variable importance. Different color represents different trait-year combination.

We used the decrease in predictive R^2 as a proxy for variable importance for grouped wavelengths. We plotted the variable importances against the average wavelength heritability (Figure 3.4). There was no apparent relationship between the two. These results demonstrate that highly heritable phenomic features are not necessarily the most informative predictors of target traits. Variable importance depends on the covariance structure between phenomic features and the target trait, rather than on feature heritability alone. Consequently, selecting phenomic features based solely on heritability ignores the relationship between it and target traits, and it may also overlook features with substantial predictive value.

In this study, we used a five-fold cross validation. Four folds of data were used for training models and the remaining one fold was used for testing. This process was repeated for each of the five folds. However, when NIRS data were incorporated, the training data included all five folds of NIRS data, only excluding the target trait data from the testing fold. The goal of this setup was analogous to a real life scenario where we observe parts of the traits of interest but all of the phenomic data. Incorporating NIRS data led to increased prediction accuracy for titratable acidity, a result that is

beneficial for strawberry breeding, as this trait is time-consuming and labor-intensive to measure. The benefit of incorporating NIRS data depended strongly on the target trait and modeling strategy. The most consistent improvements were observed for titratable acidity and Brix with the two-step prediction method, supporting the use of NIRS data as indirect phenotypes for traits that are costly or labor-intensive to measure directly. The relatively modest gains observed for some traits may reflect the limited spectral resolution of the dataset or weak biological associations between NIRS measurements and the target traits. Future studies incorporating broader spectral ranges or traits more directly related to chemical composition, such as aroma or volatile compounds, may further improve the utility of phenomic data in genomic prediction.

3.5. acknowledgements

The authors would like to thank Dominique Pincot for imputing the markers of genotypes for year 2023, and acknowledge helpful comments from Jean-Luc Jannink. UC Davis is located in Davis, California, which is located on the land that was the home of the Patwin people for thousands of years.

3.6. conflict of interest

The authors declare no conflicts of interest.

APPENDIX A

Appendix of Chapter 2

A.1. Derivation of correction factor for ρ_y

Let $f(\mathbf{w}) = \hat{y}$ be a prediction model that takes the feature vector $\mathbf{w}$. Assuming the following decomposition of y into additive (a), nonadditive genetic (u) and residual (e) variation (where we combine e and m_y into the same term for simplicity):

$$(A.1) \quad y = a + u + e.$$

We find:

$$(A.2) \quad \begin{aligned} \frac{\rho_y - \rho_e \sqrt{1 - h^2}}{\sqrt{h^2}} &= \left\{ \frac{\text{cov}(\hat{y}, y)}{\sqrt{\text{var}(\hat{y})} \sqrt{\text{var}(y)}} - \frac{\text{cov}(\hat{y}, u + e)}{\sqrt{\text{var}(\hat{y})} \sqrt{\text{var}(u + e)}} \sqrt{\frac{\text{var}(u + e)}{\text{var}(y)}} \right\} \sqrt{\frac{\text{var}(y)}{\text{var}(a)}} \\ &= \frac{\text{cov}(\hat{y}, y)}{\sqrt{\text{var}(\hat{y})} \sqrt{\text{var}(a)}} - \frac{\text{cov}(\hat{y}, u + e)}{\sqrt{\text{var}(\hat{y})} \sqrt{\text{var}(a)}} \\ &= \frac{\text{cov}(\hat{y}, a)}{\sqrt{\text{var}(\hat{y})} \sqrt{\text{var}(a)}} \\ &= \rho. \end{aligned}$$

If we can assume that $\rho_e = \text{cor}(\hat{y}, u + e) = 0$, then the quantity on the left simplifies to the familiar $\rho_y / \sqrt{h^2} = \rho$. This is often a reasonable assumption for additive genomic prediction models, but is not a reasonable assumption for any phenomic prediction model because $\text{cor}(\hat{y}, u) \neq 0$. Furthermore, whenever the features $\mathbf{w}$ are measured on the same individuals as y , $\text{cor}(\hat{y}, e) \neq 0$.

(?) provided a similar "semi-parametric accuracy estimate" for a prediction model $g(\mathbf{x}, \mathbf{w})$ that takes in both genetic ($\mathbf{x}$) and phenomic ($\mathbf{w}$) features. This estimate also accounts for non-independence among the selection candidates through a genomic relationship matrix $\mathbf{K}$.

A.2. Derivation of ρ_e when PP measures y directly

If phenomic features are measurements of the targeted trait, for example, plant height, and we assume there is little measurement error on the phenotype (*i.e.* m_y becomes 0) but some measurement error on the phenomic prediction, then Diagram 2.11 becomes:

$$(A.3) \quad \begin{array}{ccccc} & & m_p & & \\ & & \searrow & & \\ & & \hat{a}_p & & \\ & & \nearrow & & \\ a & \longrightarrow & g & \longrightarrow & t \\ u & \nearrow & & \nearrow & \searrow \\ & & e & & y \end{array}$$

Here, we use $\hat{a}_p$ instead of $\mathbf{P}$ for simplification. In the simplest case, $y = a + u + e$ and $\hat{a}_p = a + u + e + m_p$. ρ and ρ_e become:

$$\rho = \frac{\text{cov}(\hat{a}_p, a)}{\sqrt{\text{var}(\hat{a}_p)\text{var}(a)}},$$

$$\rho_e = \frac{\text{cov}(\hat{a}_p, u + e)}{\sqrt{\text{var}(\hat{a}_p)\text{var}(u + e)}},$$

and $\frac{\rho}{\rho_e}$ becomes:

$$\begin{aligned} \frac{\rho}{\rho_e} &= \frac{\text{cov}(\hat{a}_p, a)}{\text{cov}(\hat{a}_p, u + e)} \frac{\sqrt{\text{var}(\hat{a}_p)\text{var}(u + e)}}{\sqrt{\text{var}(\hat{a}_p)\text{var}(a)}} \\ &= \frac{\text{var}(a)}{\text{var}(u + e)} \frac{\sqrt{\text{var}(u + e)}}{\sqrt{\text{var}(a)}} \\ &= \sqrt{\frac{\text{var}(a)}{\text{var}(u + e)}} \\ &= \sqrt{\frac{h^2}{1 - h^2}}, \\ \rho_e &= \rho \sqrt{\frac{1 - h^2}{h^2}}. \end{aligned}$$

We plug this into Equation 2.13:

$$\begin{aligned}
\rho &= \frac{\rho_y - \rho \sqrt{\frac{1-h^2}{h^2}} \sqrt{1-h^2}}{\sqrt{h^2}} \\
&= \frac{\rho_y}{\sqrt{h^2}} - \rho \frac{1-h^2}{h^2}, \\
\rho(1 + \frac{1-h^2}{h^2}) &= \frac{\rho_y}{\sqrt{h^2}}, \\
\rho &= \rho_y \frac{1}{\sqrt{h^2}} \frac{1}{1 + \frac{1-h^2}{h^2}} \\
&= \rho_y \sqrt{h^2}.
\end{aligned}$$

A.3. Derivation of estimating accuracy with bivariate linear mixed model

In PP, we can consider $\mathbf{Y} = [\mathbf{y}, \hat{\mathbf{a}}_p]$ an $n \times 2$ matrix. Although we calculated $\hat{\mathbf{a}}_p$ using phenomic data, we assume that it can still be written as an additive linear model with genetic ($\mathbf{a}_p$) and environmental ($\mathbf{e}_p$) components, and we can write a bivariate linear model as following:

$$\begin{aligned}
\mathbf{Y} &= \mathbf{X}\beta + \mathbf{A} + \mathbf{E}, \quad \mathbf{A} \sim \text{MN}(0, \mathbf{K}, \mathbf{G}), \quad \mathbf{E} \sim \text{MN}(0, \mathbf{I}, \mathbf{R}) \\
\mathbf{G} &= \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix},
\end{aligned}
\tag{A.4}$$

where $\mathbf{X}$ is a $n \times 2$ matrix of 1's, and β a 2×1 vector of intercepts, $\mathbf{A} = [\mathbf{a}, \mathbf{a}_p]$ and $\mathbf{E} = [\mathbf{e}, \mathbf{e}_p]$, $\text{MN}(\cdot, \cdot, \cdot)$ is the matrix normal distribution. By fitting Model A.4, we can estimate $\rho_g = \frac{g_{12}}{\sqrt{g_{11}g_{22}}}$ and $h_a^2 = \frac{g_{22}}{g_{22}+r_{22}}$, and plug these into equation 2.14 to estimate PP accuracy. In the analyses in the main text we fit the model with `estimate_gcor()` function from https://github.com/deruncie/MegaLMM_analyses/blob/master/Wheat/Estimate_gcor_prediction.R which uses a Gibbs sampler implemented in the `MCMCglmm` function of the `MCMCglmm` package (36) with the following parameters:

```

prior = list(R = list(V = diag(c(.5,.01),2),nu = 3),
             G = list(G1 = list(V = diag(c(.5,.5),2),
                                   nu = 3,
                                   alpha.mu = rep(0,2),

```

```
alpha.V = diag(1,2)))
```

```
nitt = 30000
```

```
thin = 50
```

```
burn = 5000
```

to approximate a flat prior on ρ_g . For $\mathbf{y}$, we pre-processed all phenotype data at once and then split the processed data into testing and training partitions for cross-validation.

APPENDIX B

Supplementary Materials of Chapter 1

The supplementary materials in Chapter 1 can be found at <https://doi.org/10.1093/g3journal/jkae296>

APPENDIX C

Supplementary Materials of Chapter 3

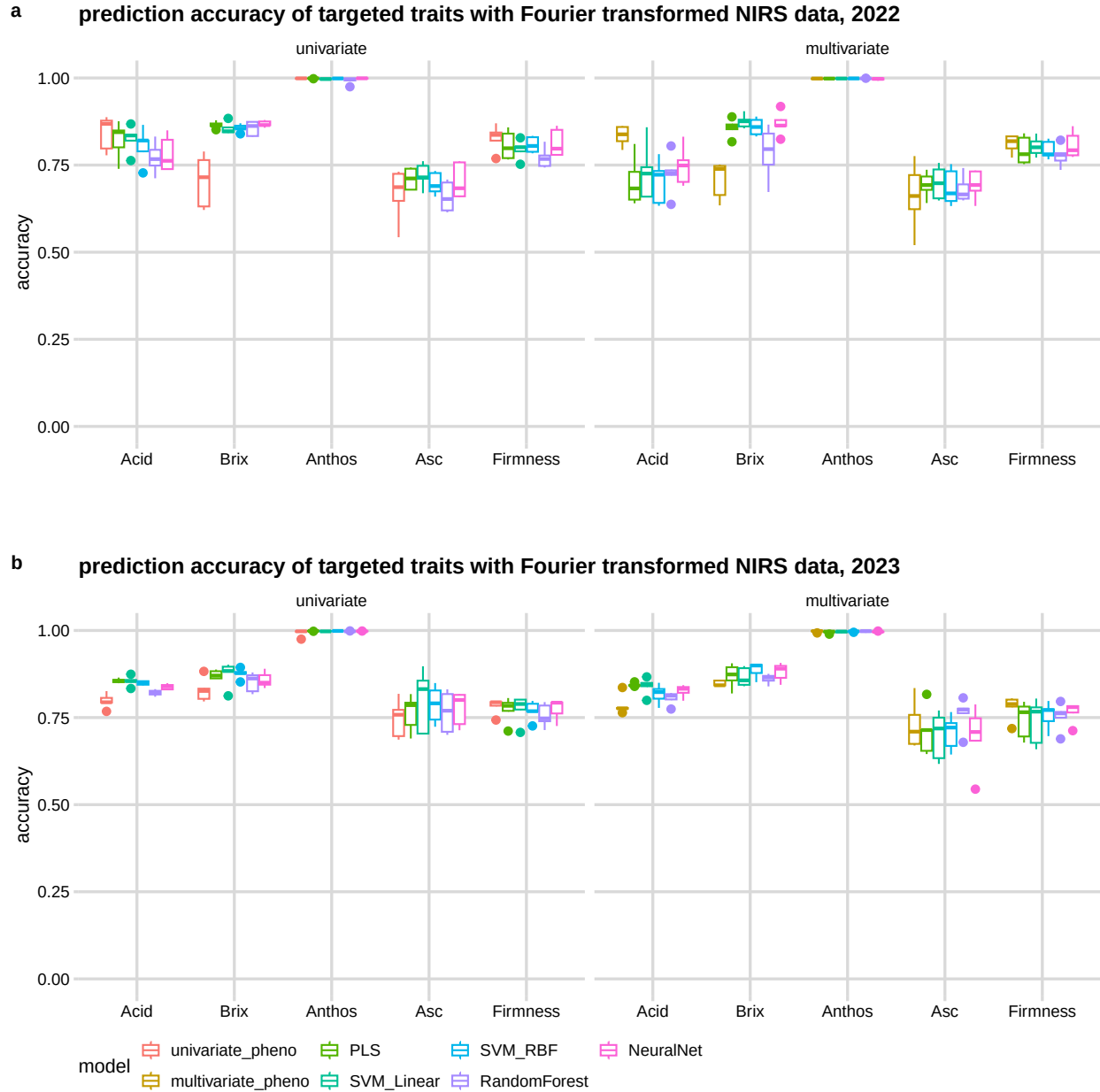


FIGURE C.1. Boxplots of "two-step" genomic prediction accuracies of target fruit quality traits (x-axis) were made with a five-fold cross-validation. Panels are separated by year. Fourier transformed NIRS data were used in the first step. "univariate_pheno" and "multivariate_pheno" boxes were calculated from single step genomic prediction models using NIRS data as secondary traits, which were plotted in Figure 3.1. They are plotted here again for comparison between single step and two-step genomic prediction models.

Bibliography

- [1] H. ALILOO, J. E. PRYCE, O. GONZÁLEZ-RECIO, B. G. COCKS, M. E. GODDARD, AND B. J. HAYES, *Including nonadditive genetic effects in mating programs to maximize dairy farm profitability*, Journal of Dairy Science, 100 (2017), pp. 1203–1222.
- [2] H. ALILOO, J. E. PRYCE, O. GONZÁLEZ-RECIO, B. G. COCKS, AND B. J. HAYES, *Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits*, Genetics Selection Evolution, 48 (2016), p. 8.
- [3] D. BATES, M. MÄCHLER, B. BOLKER, AND S. WALKER, *Fitting linear mixed-effects models using lme4*, Journal of Statistical Software, 67 (2015), pp. 1–48.
- [4] D. BATES, M. MÄCHLER, B. BOLKER, AND S. WALKER, *Fitting Linear Mixed-Effects Models Using lme4*, Journal of Statistical Software, 67 (2015), pp. 1–48.
- [5] J. BERLINGERI, A. FUENTES, E. RANARIO, H. YUN, E. Y. RIM, O. GARRETT, A. HOWARD, M.-F. LAPORTE, S. LO, D. PAULI, J. HERSHBERGER, M. EARLES, A. VAN DEYNZE, E. C. BRUMMER, R. MICHELMORE, C. Y. S. WONG, T. S. MAGNEY, P. C. RONALD, D. E. RUNCIE, B. N. BAILEY, AND C. H. DIEPENBROCK, *Integration of crop modeling and sensing into molecular breeding for nutritional quality and stress tolerance*, Theoretical and Applied Genetics, 138 (2025), p. 205.
- [6] R. BERNARDO, *Essentials of Plant Breeding*, Stemma Press, 2014.
- [7] ———, *Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE*, Heredity, 125 (2020), pp. 375–385.
- [8] ———, *Outliers and their distribution in breeding populations*, Crop Science, 62 (2022), pp. 1107–1114.
- [9] R. BERNARDO, L. MOREAU, AND A. CHARCOSSET, *Number and Fitness of Selected Individuals in Marker-Assisted and Phenotypic Recurrent Selection*, Crop Science, 46 (2006), pp. 1972–1980. __eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.2135/cropsci2006.01-0057>.

- [10] J. A. BHAT, S. ALI, R. K. SALGOTRA, Z. A. MIR, S. DUTTA, V. JADON, A. TYAGI, M. MUSHTAQ, N. JAIN, P. K. SINGH, G. P. SINGH, AND K. V. PRABHU, *Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding*, *Frontiers in Genetics*, 7 (2016), p. 221.
- [11] P. BIJMA, *The Price equation as a bridge between animal breeding and evolutionary biology*, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375 (2020), p. 20190360.
- [12] R. BRACEWELL, *The Fourier Transform & Its Applications*, McGraw-Hill Science/Engineering/Math, Boston, 2000.
- [13] C. BRAULT, J. LAZERGES, A. DOLIGEZ, M. THOMAS, M. ECARNOT, P. ROUMET, Y. BERTRAND, G. BERGER, T. PONS, P. FRANÇOIS, L. LE CUNFF, P. THIS, AND V. SEGURA, *Interest of phenomic prediction as an alternative to genomic prediction in grapevine*, *Plant Methods*, 18 (2022), p. 108.
- [14] B. L. BROWNING, Y. ZHOU, AND S. R. BROWNING, *A One-Penny Imputed Genome from Next-Generation Reference Panels*, *The American Journal of Human Genetics*, 103 (2018), pp. 338–348.
- [15] S. R. BROWNING AND B. L. BROWNING, *Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering*, *The American Journal of Human Genetics*, 81 (2007), pp. 1084–1097.
- [16] J. BURGUEÑO, G. DE LOS CAMPOS, K. WEIGEL, AND J. CROSSA, *Genomic Prediction of Breeding Values when Modeling Genotype $\times$ Environment Interaction using Pedigree and Dense Molecular Markers*, *Crop Science*, 52 (2012), pp. 707–719.
__eprint: <https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2135/cropsci2011.06.0299>.
- [17] A. CABALLERO, *Quantitative Genetics*, Cambridge University Press, Apr. 2020.
- [18] J. N. COBB, R. U. JUMA, P. S. BISWAS, J. D. ARBELAEZ, J. RUTKOSKI, G. ATLIN, T. HAGEN, M. QUINN, AND E. H. NG, *Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder’ s equation*, *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 132 (2019), pp. 627–645.
- [19] J. C. M. DEKKERS, *Prediction of response to marker-assisted and genomic selection using selection index theory*, *Journal of Animal Breeding and Genetics*, 124 (2007), pp. 331–341.
__eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1439-0388.2007.00701.x>.

- [20] J. B. ENDELMAN, *Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP*, The Plant Genome, 4 (2011).
- [21] J. B. ENDELMAN, *Ridge regression and other kernels for genomic selection with r package rrblup*, Plant Genome, 4 (2011), pp. 250–255.
- [22] J. B. ENDELMAN, M. KANTE, H. LINDQVIST-KREUZE, A. KILIAN, L. M. SHANNON, M. V. CARAZA-HARTER, B. VAILLANCOURT, K. MAILLOUX, J. P. HAMILTON, AND C. R. BUELL, *Targeted genotyping-by-sequencing of potato and data analysis with R/polyBreedR*, The Plant Genome, 17 (2024), p. e20484.
- [23] J. ERTL, A. LEGARRA, Z. G. VITEZICA, L. VARONA, C. EDEL, R. EMMERLING, AND K.-U. GÖTZ, *Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle*, Genetics Selection Evolution, 46 (2014), p. 40.
- [24] D. S. FALCONER AND T. F. C. MACKAY, *Introduction to Quantitative Genetics*, Pearson Education, 1996. Google-Books-ID: 7ASZNAEACAAJ.
- [25] M. J. FELDMANN, D. D. A. PINCOT, G. S. COLE, AND S. J. KNAPP, *Genetic gains underpinning a little-known strawberry Green Revolution*, Nature Communications, 15 (2024), p. 2468.
- [26] M. J. FELDMANN, D. D. A. PINCOT, M. V. VACHEV, R. A. FAMULA, G. S. COLE, AND S. J. KNAPP, *Accelerating genetic gains for quantitative resistance to verticillium wilt through predictive breeding in strawberry*, The Plant Genome, 17 (2024), p. e20405.
- [27] C. C. F. FILHO, M. H. M. L. ANDRADE, J. A. R. NUNES, D. H. JARQUIN, AND E. F. RIOS, *Genomic prediction for complex traits across multiples harvests in alfalfa (Medicago sativa L.) is enhanced by enviromics*, The Plant Genome, 16 (2023), p. e20306. __eprint: <https://acsess.onlinelibrary.wiley.com/doi/pdf/10.1002/tpg2.20306>.
- [28] B. FORSTER, B. TILL, A. GHANIM, H. HUYNH, H. BURSTMAYR, AND P. CALIGARI, *Accelerated plant breeding.*, CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources, 9 (2015).
- [29] R. T. FURBANK AND M. TESTER, *Phenomix –technologies to relieve the phenotyping bottleneck*, Trends in Plant Science, 16 (2011), pp. 635–644.
- [30] R. J. GALÁN, A.-M. BERNAL-VASQUEZ, C. JEBSEN, H.-P. PIEPHO, P. THORWARTH, P. STEFFAN, A. GORDILLO, AND T. MIEDANER, *Integration of genotypic, hyperspectral, and*

- phenotypic data to improve biomass yield prediction in hybrid rye*, Theoretical and Applied Genetics, 133 (2020), pp. 3001–3015.
- [31] R. C. GAYNOR, G. GORJANC, A. R. BENTLEY, E. S. OBER, P. HOWELL, R. JACKSON, I. J. MACKAY, AND J. M. HICKEY, *A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines*, Crop Science, 57 (2017), pp. 2372–2386. [_eprint: https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2135/cropsci2016.09.0742](https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2135/cropsci2016.09.0742).
 - [32] P. GELADI, D. MACDOUGALL, AND H. MARTENS, *Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat*, Applied Spectroscopy, 39 (1985), pp. 491–500.
 - [33] D. GIANOLA, *Priors in Whole-Genome Regression: The Bayesian Alphabet Returns*, Genetics, 194 (2013), pp. 573–596.
 - [34] M. E. GODDARD, B. J. HAYES, AND T. H. E. MEUWISSEN, *Using the genomic relationship matrix to predict the accuracy of genomic selection*, Journal of Animal Breeding and Genetics = Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie, 128 (2011), pp. 409–421.
 - [35] J. D. HADFIELD, *MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package*, Journal of Statistical Software, 33 (2010), pp. 1–22.
 - [36] J. D. HADFIELD, *Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package*, Journal of Statistical Software, 33 (2010), pp. 1–22.
 - [37] M. A. HARDIGAN, M. J. FELDMANN, J. CARLING, A. ZHU, A. KILIAN, R. A. FAMULA, G. S. COLE, AND S. J. KNAPP, *A medium-density genotyping platform for cultivated strawberry using DArTag technology*, The Plant Genome, 16 (2023), p. e20399.
 - [38] T. HASTIE, R. TIBSHIRANI, AND J. FRIEDMAN, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York, NY, 2009.
 - [39] E. L. HEFFNER, M. E. SORRELLS, AND J. JANNINK, *Genomic Selection for Crop Improvement*, Crop Science, 49 (2009), pp. 1–12.
 - [40] A. HÖSKULDSSON, *Variable and subset selection in PLS regression*, Chemometrics and Intelligent Laboratory Systems, 55 (2001), pp. 23–38.
 - [41] J.-L. JANNINK, *Dynamics of long-term genomic selection*, Genetics Selection Evolution, 42 (2010), p. 35.
 - [42] J.-L. JANNINK, A. J. LORENZ, AND H. IWATA, *Genomic selection in plant breeding: from theory to practice*, Briefings in Functional Genomics, 9 (2010), pp. 166–177.

- [43] Y. JIANG AND C. LI, *Convolutional Neural Networks for Image-Based High-Throughput Plant Phenotyping: A Review*, Plant Phenomics, 2020 (2020), p. 4152816.
- [44] N. P. JIMÉNEZ, M. J. FELDMANN, R. A. FAMULA, D. D. A. PINCOT, M. BJORNSEN, G. S. COLE, AND S. J. KNAPP, *Harnessing underutilized gene bank diversity and genomic prediction of cross usefulness to enhance resistance to Phytophthora cactorum in strawberry*, The Plant Genome, 16 (2023), p. e20275.
- [45] S. J. KNAPP, G. S. COLE, D. D. PINCOT, C. M. LÒPEZ, O. A. GONZALEZ-BENITEZ, AND R. A. FAMULA, ‘*UC Eclipse*’, a Summer Plant-adapted Photoperiod-insensitive Strawberry Cultivar, HortScience, 58 (2023), pp. 1568–1572.
- [46] S. J. KNAPP, G. S. COLE, D. D. A. PINCOT, C. J. DILLA-ERMITA, M. BJORNSEN, R. A. FAMULA, T. R. GORDON, J. M. HARSHMAN, P. M. HENRY, AND M. J. FELDMANN, *Transgressive segregation, hopeful monsters, and phenotypic selection drove rapid genetic gains and breakthroughs in predictive breeding for quantitative resistance to Macrophomina in strawberry*, Horticulture Research, 11 (2024), p. uhad289.
- [47] M. R. KRAUSE, L. GONZÁLEZ-PÉREZ, J. CROSSA, P. PÉREZ-RODRÍGUEZ, O. MONTESINOS-LÓPEZ, R. P. SINGH, S. DREISIGACKER, J. POLAND, J. RUTKOSKI, M. SORRELLS, M. A. GORE, AND S. MONDAL, *Hyperspectral Reflectance-Derived Relationship Matrices for Genomic Prediction of Grain Yield in Wheat*, G3 Genes|Genomes|Genetics, 9 (2019), pp. 1231–1247.
- [48] A. LEGARRA AND A. REVERTER, *Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method*, Genetics Selection Evolution, 50 (2018), p. 53.
- [49] A. LEGARRA, C. ROBERT-GRANIÉ, E. MANFREDI, AND J.-M. ELSÉN, *Performance of Genomic Selection in Mice*, Genetics, 180 (2008), pp. 611–618. [_eprint: https://academic.oup.com/genetics/article-pdf/180/1/611/42096631/genetics0611.pdf](https://academic.oup.com/genetics/article-pdf/180/1/611/42096631/genetics0611.pdf).
- [50] C. LEHERMEIER, G. DE LOS CAMPOS, V. WIMMER, AND C.-C. SCHÖN, *Genomic variance estimates: With or without disequilibrium covariances?*, Journal of Animal Breeding and Genetics = Zeitschrift Fur Tierzucht Und Zuchtungsbiologie, 134 (2017), pp. 232–241.
- [51] R. V. LENTH AND J. PIASKOWSKI, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2026. R package version 2.0.3.

- [52] M. LOPEZ-CRUZ, E. OLSON, G. ROVERE, J. CROSSA, S. DREISIGACKER, S. MONDAL, R. SINGH, AND G. D. L. CAMPOS, *Regularized selection indices for breeding value prediction using hyper-spectral image data*, Scientific Reports, 10 (2020), p. 8195.
- [53] M. LYNCH AND B. WALSH, *Genetics and Analysis of Quantitative Traits*, Oxford University Press, Oxford, New York, Jan. 1998.
- [54] A. MAGGIORELLI, N. BAIG, V. PRIGGE, J. BRUCKMÜLLER, AND B. STICH, *Using drone-retrieved multispectral data for phenomic selection in potato breeding*, Theoretical and Applied Genetics, 137 (2024), p. 70.
- [55] T. H. E. MEUWISSEN, B. J. HAYES, AND M. E. GODDARD, *Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps*, Genetics, 157 (2001), pp. 1819–1829.
- [56] P. R. MUÑOZ, M. F. R. RESENDE, JR, S. A. GEZAN, M. D. V. RESENDE, G. DE LOS CAMPOS, M. KIRST, D. HUBER, AND G. F. PETER, *Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices*, Genetics, 198 (2014), pp. 1759–1768.
- [57] T. OSTHUSHENRICH, M. FRISCH, AND E. HERZOG, *Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines*, PLOS ONE, 12 (2017), p. e0188839.
- [58] S. B. OULD ESTAGHVIROU, J. O. OGUTU, T. SCHULZ-STREECK, C. KNAAK, M. OUZUNOVA, A. GORDILLO, AND H.-P. PIEPHO, *Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding*, BMC Genomics, 14 (2013), p. 860.
- [59] H.-P. PIEPHO AND J. MÖHRING, *Computing Heritability and Selection Response From Unbalanced Plant Breeding Trials*, Genetics, 177 (2007), pp. 1881–1888.
- [60] D. D. A. PINCOT, M. J. FELDMANN, M. A. HARDIGAN, M. V. VACHEV, P. M. HENRY, T. R. GORDON, M. BJORNSEN, A. RODRIGUEZ, N. COBO, R. A. FAMULA, G. S. COLE, G. L. COAKER, AND S. J. KNAPP, *Novel Fusarium wilt resistance genes uncovered in natural and cultivated strawberry populations are found on three non-homoeologous chromosomes*, Theoretical and Applied Genetics, 135 (2022), pp. 2121–2145.
- [61] D. D. A. PINCOT, M. LEDDA, M. J. FELDMANN, M. A. HARDIGAN, T. J. POORTEN, D. E. RUNCIE, C. HEFFELFINGER, S. L. DELLAPORTA, G. S. COLE, AND S. J. KNAPP, *Social network analysis of the genealogy of strawberry: retracing the wild roots of heirloom and modern cultivars*, G3 Genes|Genomes|Genetics, 11 (2021), p. jkab015.

- [62] J. A. POLAND AND T. W. RIFE, *Genotyping-by-Sequencing for Plant Breeding and Genetics*, The Plant Genome, 5 (2012).
- [63] P. PÉREZ AND G. DE LOS CAMPOS, *Genome-Wide Regression and Prediction with the BGLR Statistical Package*, Genetics, 198 (2014), pp. 483–495.
- [64] R. CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021.
- [65] R. RINCENT, J.-P. CHARPENTIER, P. FAIVRE-RAMPANT, E. PAUX, J. LE GOUIS, C. BASTIEN, AND V. SEGURA, *Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar*, G3 Genes|Genomes|Genetics, 8 (2018), pp. 3961–3972.
- [66] . RINNAN, F. V. D. BERG, AND S. B. ENGELSEN, *Review of the most common pre-processing techniques for near-infrared spectra*, TrAC Trends in Analytical Chemistry, 28 (2009), pp. 1201–1222.
- [67] H. RITCHIE AND L. RODÉS-GUIRAO, *Peak global population and other key findings from the 2024 UN World Population Prospects*, Our World in Data, (2024).
- [68] D. RUNCIE AND H. CHENG, *Pitfalls and Remedies for Cross Validation with Multi-trait Genomic Prediction Methods*, G3 Genes|Genomes|Genetics, 9 (2019), pp. 3727–3741.
- [69] D. E. RUNCIE, *The use of cross-validation has overestimated the value of genomic selection in plant breeding*, Jan. 2026. ISSN: 2692-8205 Pages: 2026.01.05.697784 Section: New Results.
- [70] D. E. RUNCIE, J. QU, H. CHENG, AND L. CRAWFORD, *MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits*, Genome Biology, 22 (2021), p. 213.
- [71] F. W. SCHNELL AND H. F. UTZ, *F1-leistung und elternwahl euphy-der züchtung von selbstbefruchtern*, Bericht über die Arbeitstagung der Vereinigung Österreichischer Pflanzenzüchter., (1975), pp. 243–248.
- [72] M. F. SCOTT, O. LADEJOBI, S. AMER, A. R. BENTLEY, J. BIERNASKIE, S. A. BODEN, M. CLARK, M. DELL’ACQUA, L. E. DIXON, C. V. FILIPPI, N. FRADGLEY, K. A. GARDNER, I. J. MACKAY, D. O’ SULLIVAN, L. PERCIVAL-ALWYN, M. ROORKIWAL, R. K. SINGH, M. THUDI, R. K. VARSHNEY, L. VENTURINI, A. WHAN, J. COCKRAM, AND R. MOTT, *Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding*, Heredity, 125 (2020), pp. 396–416.

- [73] G. SU, O. F. CHRISTENSEN, T. OSTERSEN, M. HENRYON, AND M. S. LUND, *Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers*, PLOS ONE, 7 (2012), p. e45293.
- [74] K. SWARTS, H. LI, J. A. ROMERO NAVARRO, D. AN, M. C. ROMAY, S. HEARNE, C. ACHARYA, J. C. GLAUBITZ, S. MITCHELL, R. J. ELSHIRE, E. S. BUCKLER, AND P. J. BRADBURY, *Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants*, The Plant Genome, 7 (2014), p. plantgenome2014.05.0023.
- [75] R. C. TEAM, *R: A Language and Environment for Statistical Computing*, tech. rep., R Foundation for Statistical Computing, Vienna, Austria, 2020.
- [76] F. TECHNOW, *R Package hypred : Simulation of Genomic Data in Applied Genetics*, 2011.
- [77] B. B. TESSEMA, H. LIU, A. C. SØRENSEN, J. R. ANDERSEN, AND J. JENSEN, *Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat*, Frontiers in Genetics, 11 (2020), p. 578123.
- [78] P. M. VANRADEN, *Efficient Methods to Compute Genomic Predictions*, Journal of Dairy Science, 91 (2008), pp. 4414–4423.
- [79] Z. G. VITEZICA, L. VARONA, J.-M. ELSÉN, I. MISZTAL, W. HERRING, AND A. LEGARRA, *Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs*, Genetics Selection Evolution, 48 (2016), p. 6.
- [80] Z. G. VITEZICA, L. VARONA, AND A. LEGARRA, *On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope*, Genetics, 195 (2013), pp. 1223–1230.
- [81] F. WANG, M. J. FELDMANN, AND D. E. RUNCIE, *Do not benchmark phenomic prediction against genomic prediction accuracy*, The Plant Phenome Journal, 8 (2025), p. e70029. `_eprint:` <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ppj2.70029>.
- [82] Y.-H. WANG AND W.-H. SU, *Convolutional Neural Networks in Computer Vision for Grain Crop Phenotyping: A Review*, Agronomy, 12 (2022), p. 2659.
- [83] H. WICKHAM, *ggplot2, Use R!*, Springer International Publishing, Cham, 2016.
- [84] C. O. WILKE, *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2020. R package version 1.1.1.

- [85] Z. J. WINN, A. L. AMSBERRY, S. D. HALEY, N. D. DEWITT, AND R. E. MASON, *Phenomic versus genomic prediction—A comparison of prediction accuracies for grain yield in hard winter wheat lines*, The Plant Phenome Journal, 6 (2023), p. e20084. [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ppj2.20084](https://onlinelibrary.wiley.com/doi/pdf/10.1002/ppj2.20084).
- [86] M. D. WOLFE, A. W. CHAN, P. KULAKOW, I. RABBI, AND J.-L. JANNINK, *Genomic mating in outbred species: predicting cross usefulness with additive and total genetic covariance matrices*, Genetics, 219 (2021), p. iyab122.
- [87] T. XIANG, O. F. CHRISTENSEN, Z. G. VITEZICA, AND A. LEGARRA, *Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs*, Genetics Selection Evolution, 48 (2016), p. 92.
- [88] S. ZHONG, J. C. M. DEKKERS, R. L. FERNANDO, AND J.-L. JANNINK, *Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study*, Genetics, 182 (2009), pp. 355–364.
- [89] S. ZHONG AND J.-L. JANNINK, *Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance*, Genetics, 177 (2007), pp. 567–576.