NEW DEVELOPMENTS FOR EMBRACING GENOMIC SELECTION IN BREEDING APPLICATIONS

EDITED BY: Diego Jarquin, Waseem Hussain and Shiori Yabe
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-434-3
DOI 10.3389/978-2-88974-434-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

NEW DEVELOPMENTS FOR EMBRACING GENOMIC SELECTION IN BREEDING APPLICATIONS

Topic Editors:
Diego Jarquin, University of Nebraska-Lincoln, United States
Waseem Hussain, International Rice Research Institute (IRRI), Philippines
Shiori Yabe, Institute of Crop Science (NARO), Japan

Citation: Jarquin, D., Hussain, W., Yabe, S., eds. (2022). New Developments for Embracing Genomic Selection in Breeding Applications. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-434-3

Table of Contents

frontiers in Genetics

METHODS
published: 18 December 2020
doi: 10.3389/fgene.2020.599510

Predicting Rice Heading Date Using an Integrated Approach Combining a Machine Learning Method and a Crop Growth Model

*Tai-Shen Chen[1], Toru Aoike[1], Masanori Yamasaki[2], Hiromi Kajiya-Kanegae[1] and Hiroyoshi Iwata[1]**

[1] Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo, Japan, [2] Food Resources Education and Research Center, Graduate School of Agricultural Science, Kobe University, Kasai, Hyogo, Japan

Accurate prediction of heading date under various environmental conditions is expected to facilitate the decision-making process in cultivation management and the breeding process of new cultivars adaptable to the environment. Days to heading (DTH) is a complex trait known to be controlled by multiple genes and genotype-by-environment interactions. Crop growth models (CGMs) have been widely used to predict the phenological development of a plant in an environment; however, they usually require substantial experimental data to calibrate the parameters of the model. The parameters are mostly genotype-specific and are thus usually estimated separately for each cultivar. We propose an integrated approach that links genotype marker data with the developmental genotype-specific parameters of CGMs with a machine learning model, and allows heading date prediction of a new genotype in a new environment. To estimate the parameters, we implemented a Bayesian approach with the advanced Markov chain Monte-Carlo algorithm called the differential evolution adaptive metropolis and conducted the estimation using a large amount of data on heading date and environmental variables. The data comprised sowing and heading dates of 112 cultivars/lines tested at 7 locations for 14 years and the corresponding environmental variables (day length and daily temperature). We compared the predictive accuracy of DTH between the proposed approach, a CGM, and a single machine learning model. The results showed that the extreme learning machine (one of the implemented machine learning models) was superior to the CGM for the prediction of a tested genotype in a tested location. The proposed approach outperformed the machine learning method in the prediction of an untested genotype in an untested location. We also evaluated the potential of the proposed approach in the prediction of the distribution of DTH in 103 F_2 segregation populations derived from crosses between a common parent, Koshihikari, and 103 cultivars/lines. The results showed a high correlation coefficient (ca. 0.8) of the 10, 50, and 90th percentiles of the observed and predicted distribution of DTH. In this study, the integration of a machine learning model and a CGM was better able to predict the heading date of a new rice cultivar in an untested potential environment.

Keywords: crop growth model, bayesian inference, differential evolution adaptive metropolis, machine learning, Markov chain Monte-Carlo

OPEN ACCESS

Edited by:
Waseem Hussain, International Rice Research Institute (IRRI), Philippines

Reviewed by:
Saeed Khaki, Iowa State University, United States
Miguel Perez-Enciso, Autonomous University of Barcelona, Spain

***Correspondence:**
Hiroyoshi Iwata
hiroiwata@g.ecc.u-tokyo.ac.jp

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *27 August 2020*
Accepted: *26 November 2020*
Published: *18 December 2020*

Citation:
Chen T-S, Aoike T, Yamasaki M, Kajiya-Kanegae H and Iwata H (2020) Predicting Rice Heading Date Using an Integrated Approach Combining a Machine Learning Method and a Crop Growth Model. Front. Genet. 11:599510. doi: 10.3389/fgene.2020.599510

INTRODUCTION

Heading date is a critical trait for the adoption of a rice cultivar to target cultivation area and cropping season (Yano et al., 1997). Improvement in the understanding and modeling of rice phenology could benefit production and breeding. However, it has been challenging to model a complex trait such as heading date, which is usually influenced by genotype, environment, and their interaction. In the past, when available data were limited, simple models such as those based on growing degree days have been used widely, but these have good predictability for the specific genotype in environments with few variabilities. With more available data and knowledge and the improvement in computing power, more complex models such as crop growth models (CGMs) have been developed to simulate the performance of genotypes in a wide range of environments, mainly variable temperature and photoperiod. A CGM is implemented as a process-based mathematical set of equations describing the growth process of a crop plant, and enables the prediction of growth and production under environmental, management, and physiological input variables. The physiological parameters in the CGM equations account for the among-genotype differences and are usually regarded as environment-independent genotypic characteristics (Yin et al., 2000). This allows the predictions to be unrestricted to environments where the model parameters are calibrated/estimated (Yin et al., 2003).

As genetic marker information becomes available, the genetic control of the response to environments can be revealed via the dissection of the variation in the CGM parameters into the effects of discrete genetic loci—quantitative trait loci (QTLs). The relevant studies include the research on flowering time in barely (Yin et al., 2005), rice (Nakagawa et al., 2005), *Brassica oleracea* (Uptmoor et al., 2008), and wheat (Bogard et al., 2014). These studies suggest the possibility of predicting the performance of a given genotype in an untested environment by plugging in the parameters that are predicted for the genotype based on the estimated QTL effects into a CGM. As an example, Bogard et al. (2014) predicted days to heading (DTH) of wheat based on the estimated QTL effects and found that the root mean square error (RMSE) between the observed and predicted values was 6.3 days. The approach of integrating a gene-based or QTL-based model with a CGM has been advocated by several studies (White and Hoogenboom, 1996; Chapman et al., 2002a,b; Letort et al., 2008). However, further refinement is required for linking the CGM parameters with genotypes of markers or genes.

For the integrated approach, we must first estimate the parameters of the CGM using the phenotypic and environmental data collected in field experiments. Owing to several reasons, such as the lack of sufficient input data for estimating many parameters, difficulties in defining the criteria for validating the predicted accuracy of a CGM, and the diverse structure of input data, the estimation of CGM parameters remains a rather open field (Seidel et al., 2018). The estimation methods can be classified as frequentist or Bayesian. The frequentist approach assumes that the parameter is a fixed effect and does not include the prior information of the parameter in the model. The Bayesian approach assumes that the parameter is a random variable and the prior information is built into the model. A comprehensive introduction of this topic can be found in Makowski et al. (2006). Although the better choice among Bayesian and frequentist approaches is not clear, the Bayesian approach could provide further information regarding the parameters, such as the uncertainty of the estimates, when the main interest is in interpreting the biological meaning of estimated parameter values instead of optimizing the predicted accuracy of the CGM. In several studies (Iizumi et al., 2009; Jones et al., 2011) a Bayesian approach with the Markov chain Monte Carlo (MCMC) technique has been applied for estimating CGM parameters. The commonly used MCMC method, such as the Metropolis–Hastings algorithm, however, has slow convergence in practice. Dumont et al. (2014) and Iizumi et al. (2014) suggested the use of an advanced MCMC technique, such as the differential evolution adaptive metropolis (DREAM) algorithm, which can automatically tune the scale and orientation of the proposed distribution during the search and overcome the problems of heavy-tailed and multimodal posteriors.

Another consideration is how an integrated framework connecting the CGM to markers or genes can be built for predicting complex traits. A straightforward approach is the two-step approach that first computes the estimates of the CGM parameters and then uses the statistical models developed for QTL analysis or genomic prediction (Meuwissen et al., 2001) to predict the CGM parameters. A unified predictive system has also been proposed by Technow et al. (2015) and Onogi et al. (2016) Their framework applied different Bayesian approaches, but both based their system on a single hierarchical model instead of the two-stage approach to predict complex traits such as yield in maize and heading in rice, respectively. Although the integration of genomic prediction with CGM has shown good potential in previous studies, another modeling paradigm, such as machine learning, could also have great potential as a candidate method for modeling the non-linear, complicated interaction between the gene and the environment.

Unlike statistical models that focus more on the extraction of information on the underlying mechanism producing the data, the machine learning method is concerned with the accuracy of prediction (Breiman, 2001b). As a result of the big data era, machine learning has shown unprecedented predictive power against traditional statistical models. However, there were very few studies applying the machine learning method in predicting crop growth, which could stem from the lack of appropriate data and unfamiliarity with this method in the relevant community. In this study, we collected the heading data of 112 rice cultivars/lines tested in multiple locations from 2004 to 2017. This large amount of heading data combined with environmental data and genetic marker data allowed us to train a robust machine learning model and to compare its predictability with that of other methods. We also collected the heading of 103 F_2 segregating populations created from the crosses of cultivars/lines, which were selected from the 112 cultivars/lines. This F_2 population data helps validate the model performance in predicting DTH of a simulated genotype in a new environment. In addition to training a single machine learning model, building an integrated

framework combining a CGM and a machine learning model to predict the complex trait could also be a promising method that has not been attempted earlier.

In this study, we explored the potential use of the machine learning method and proposed an integrated approach that could be superior in an interpolation scenario. We implemented a Bayesian method for the estimation of CGM parameters. Although many powerful machine learning methods have been proposed, there is no single best method that can outperform others on all fronts, such as the so-called "no free lunch" theorem. In this study, we evaluated three representative methods: two decision tree-based approaches [random forest (RF) and eXtreme gradient boosting (XGB)], and a neural network-based approach [extreme learning machine (ELM)]. We compared the predictive performance of different modeling methods, including a CGM [developmental rate (DVR) model], three machine learning methods, and the proposed integrated framework, which combines machine learning and CGM using a two-stage approach to predict the DTH in rice. The comparison was performed under three cross-validation schemes. We also examined the ability of the proposed integrated framework in predicting the distribution of DTH in 103 F_2 segregation populations and demonstrated the superiority of the proposed approach in predicting the heading date of a new genotype in a new environment.

MATERIALS AND METHODS

Rice Heading Data

Two datasets of experiments evaluating DTH in rice cultivars/lines were analyzed in this study. The first was the dataset of 112 cultivars/lines, and the other was the dataset of F_2 segregation populations derived from crosses between a Japanese leading cultivar as a common parent, Koshihikari, and 103 cultivars/lines. The 112 cultivars/lines dataset comprised 7,098 observations of sowing, transplanting, and heading dates of the 112 cultivars/lines evaluated in eight locations in Japan from 2004 to 2017 (64 combinations of locations and years in total, **Supplementary Table S1**). The 112 cultivars/lines were chosen from those developed in different regions of Japan (**Supplementary Table S2**). The experiments were conducted in one location (Tsukubamirai) in the middle of Japan in the first 2 years, and then gradually expanded to other locations distributed from the north to the south of Japan in the following years. All 112 cultivars/lines were sown and transplanted at the same time in a single experiment at each location, and more than one experiment (sowing and transplanting on different dates) was conducted at some locations. We defined the heading date as the date when more than 50% of individuals reached the heading stage. The number of plants evaluated for each cultivar/line was different among the experiments and ranged from 7 to 30. DTH was calculated as the difference between the heading date and sowing date. In 70 of 7,168 cases, cultivars/lines did not reach the heading stage before the end of the experiment. Thus, 70 cases were removed from the analysis. The dataset of the F_2 segregation population was created by crossing Koshihikari and 103 of the 112 cultivars/lines. In 2007 and 2008, we evaluated 73 and 30 F_2 populations, respectively, in Kasai, Hyogo. Each population was evaluated using 96 F_2 plants (genotypes). The distribution of DTH in each segregation population was obtained by recording the heading date of each plant individually.

Meteorological Data

Temperature and photoperiod (day length) are the two most influential meteorological factors affecting the phenological development (e.g., flowering) of rice. We downloaded the daily average temperature data from the Agro-Meteorological Grid Square Data, National Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization, Japan. We computed the theoretical day length based on the latitude and longitude of each location according to the CBM model (Forsythe et al., 1995).

Genotype Marker Data

We used two sets of genotypic marker data from 112 cultivars/lines in this study. The first was the genotype data of 14 SNPs in five heading date-related genes, *Hd1* (Yano et al., 2000), *Ghd7* (Xue et al., 2008), *Hd6* (Takahashi et al., 2001), *Hd16* (Hori et al., 2013), and *Hd17* (Matsubara et al., 2012). The other was the genotype data of 1,594 markers, which included the 14 heading date-related SNPs and other SNPs and Simple-sequence repeats (SSRs) markers. We generated 1,000 simulated genotypes of the 14 heading date-related SNPs as simulated progeny from each F_2 population. The simulation was performed based on the linkage map positions of the SNPs and genotype marker data of parents of an F_2 population.

Methods for Predicting Rice Heading

We compared three methods in the prediction of the heading date of rice. CGM, a machine learning method, and the proposed integrated models. The three methods are described in **Table 1** with the type of input data and the type of cross-validation schemes, which are explained in section "Cross-Validation."

DVR Model

A CGM named the DVR model was modified from a three-stage beta model (Yin et al., 1997), as proposed by Nakagawa et al. (2005). The model assumes that the pre-flowering development of a rice plant is divided into three subphases: (1) the juvenile phase, when the plant is insensitive to the flowering stimulus; (2) the "photoperiod sensitive phase," when the plant starts to respond to the photoperiodic flowering stimulus; and (3) the "post- photoperiod sensitive phase," the period after the completion of the photoperiod sensitive phase. The progress of developmental stages (DVS) from seedling emergence (DVS0), flowering (DVS1) to maturation (DVS2) is quantified as 0, 1, and 2, respectively, and is calculated by integrating the growth rate of the i-th day DVR_i as:

$$DVS_d = \sum_{i=0}^{d} DVR_i$$

where d is the number of days since seedling emergence. DVR is modeled as the multiplicative function of a temperature response function and a photoperiod response function, and is defined as follows:

$$DVR_i = \begin{cases} \frac{f(T_d)}{G} & if\ DVS_d < DVS_1\ or\ DVS_d > DVS_2 \\ f(T_d)g(P_d)/G & if\ DVS_1 < DVS_d < DVS_2 \end{cases}$$

where T_d and P_d are the daily mean temperature (°C) and the photoperiod (h) of the d-th day, respectively, f and g denote the temperature response function and photoperiod function, respectively, and G ($G > 0$) denotes the earliness of flowering under the optimal condition. DVS_1 and DVS_2 represent the ends of the juvenile and photosensitive phases, respectively. The functions f and g are given by

$$f(T_d) =$$

$$\begin{cases} \left[\left(\frac{T_d - T_b}{T_o - T_b}\right)\left(\frac{T_c - T_d}{T_c - T_o}\right)^{(T_c - T_o)/(T_o - T_b)}\right]^{\alpha} & if \quad T_b \le T_d \le T_c, \\ 0 & otherwise \end{cases}$$

$$g(P_d) = \begin{cases} \left[\left(\frac{P_d - P_b}{P_o - P_b}\right)\left(\frac{P_c - P_d}{P_c - P_o}\right)^{\frac{P_c - P_o}{P_o - P_b}}\right]^{\beta} & if\ P_o \le P_d, \\ 1 & otherwise \end{cases}$$

where, T_b, T_c, and T_o are the base, ceiling, and optimum temperatures (in the unit of degree Celsius), respectively, and P_b, P_c, and P_o represent the base, ceiling, and optimum photoperiods (in the unit of hours), respectively. The values of T_b, T_o, T_c, P_b, P_o, and P_c, are fixed at 8, 30, 42, 0, 10, and 24 according to Nakagawa et al. (2005). The parameter α (α > 0) is the temperature-sensitivity coefficient, whereas β (β > 0) is the photoperiod sensitivity coefficient. To minimize the number of parameters, DVS_1 and DVS_2 are defined as

$$DVS_1 = 0.145 + 0.005G$$

$$DVS_2 = 0.345 + 0.005G$$

according to Nakagawa et al. (2005). Parameters α, β, and G remain in the DVR model and are assumed to be able to quantify genetic differences in phenological responses to environmental factors.

Parameter Estimation

We used the advanced MCMC algorithm to estimate the posterior distribution of the parameters (α, β, and G). The details of the implemented DREAM algorithm are provided in **Supplementary Material**. The DREAM algorithm runs multiple chains instead of a single chain. The number of chains should be larger than twice the number of parameters (three in the DVR model) and was set to 10. The number of iterations, the number of samples discarded during burn-in, and the number of selected samples were set as 50,000, 10,000, and 10,000, respectively. For the parameters in the DVR model, the normal priors and ranges of the parameters assumed in the study are summarized in **Table 2**. We developed a program in the language Julia to implement the DREAM algorithm for parameter estimation of the DVR model. The source code is available from the authors upon request.

TABLE 1 | Prediction methods used in the study.

Method[a]			Input[b]		Cross-validation[c]	
Type	**Name**	**Description**	**E**	**G**	**LOGO**	**LOGLO**
CGM	DVR	DVR model with Bayesian DREAM MCMC algorithm	✓			
Machine learning	ELM	Extreme learning machine	✓	✓		
	XGB	Gradient boosting	✓	✓	✓	✓
	RF	Random forest	✓	✓		
Integrated model	CGM-ELM	DVR-Bay -> ELM	✓	✓	✓	✓
	CGM-XGB	DVR-Bay -> GB	✓	✓	✓	✓
	CGM-RF	DVR-Bay -> RF	✓	✓	✓	✓

[a] *Methods used for predicting days to heading in rice.*
[b] *Input data used for prediction: E indicates environmental data, including daily mean temperature and daily photoperiod; G indicates the genotype marker data.*
[c] *LOGO represents leave-one-genotype-out cross-validation. LOGLO is a leave-one-combination-of-genotype-and-location-out cross-validation.*

TABLE 2 | DVR model parameters and their prior information.

Parameter	Definition	Prior N (μ, σ)	Range*	Unit
alpha	Sensitivity of temperature	N (3, 1)	0–20	–
beta	Sensitivity of photoperiod	N (4, 1)	0–25	–
G	Earliness of flowering under optimal photoperiod and temperature	N (35, 2)	30–120	Day

**A proposal that fell out of the range was discarded during Markov chain Monte-Carlo sampling.*

Machine Learning Methods

We implemented RF, XGB, and ELM to predict the heading date of rice. The same training data, with the environmental data and genotypic data as inputs and DTH as outputs, were prepared for the three machine learning methods. The environmental data of each observation consisted of daily temperature from the date of sowing to 199 days later and the daily photoperiod at the sowing day, and 100 and 200 days after sowing. As the theoretical photoperiod has a bell-shaped curve determined only by latitude and longitude, the photoperiod of three representative days was used to avoid multicollinearity in the input variables. As described in section "Genotype Marker Data," we used two types of genotype marker data. The data were converted to dummy variables and combined with environmental data as input.

RF is an ensemble learning method that combines de-correlated trees and aggregates their predictions by averaging (Breiman, 2001a). It has been successful as a general-purpose classification and regression method and is involved in various practical problems (Biau and Scornet, 2016). We implemented RF using the R package "randomForest" (Liaw and Wiener, 2002) with hyperparameters set as the default values, except for the following parameters: the number of trees $ntree = 500$ and the number of variables randomly sampled $mtry = p/3$, where p is the number of columns in the input matrix.

The gradient tree boosting proposed by Friedman (2002) is an effective and popular machine learning method. Chen and Guestrin (2016) and Sagi and Rokach (2018) implemented a scalable end-to-end tree boosting system, called XGB, which includes innovations such as a novel tree learning algorithm and a theoretically justified weighted quantile sketch procedure. XGB has won competitions for machine learning on Kaggle (Ziêba et al., 2016) and has been proven to be a versatile and effective tool in regression and classification problems. We implemented XGB using the R package "XGBoost" (Chen and He, 2015) with hyperparameters set as their default values except the following parameters: the maximum depth of a tree = 6, learning rate = 0.1, and the number of iterations = 200.

An ELM is a single hidden layer neural network that randomly assigns the hidden node learning parameters and analytically determines the network output weights by solving the linear square system using the least squares method (Huang et al., 2006). ELM can save time in the training process compared to a feedforward neural network that adjusts weights through a back-propagation method. We implemented ELM using the R package "elmNNRcpp" and set the hyperparameter for the number of hidden nodes as 100 based on the result of a grid search for 25, 50, 100, 200, and 400 nodes.

Integrated Approach

The proposed integrated approach aimed to link the genotypic effect on phenological growth using the concept shown in **Figure 1**. Data with a large variation in phenological growth among diverse genotypes tested in multiple environments are essential for the success of the proposed approach. The approach is basically a two-step model (Nakagawa et al., 2005; Bogard et al., 2014) that first links the gene effect to the parameters in the CGM through a machine learning method, and then predicts the heading date of a genotype in a target environment through the CGM. In step 1a, we estimated the model parameters (α, β, and G) for each genotype using the Bayesian method. The posterior distributions of the parameters are obtained via the Bayesian methods. The mean values of the posterior distributions were chosen as the estimates of the parameters in the Bayesian method. To link the effect of markers to the model parameters, we used 112 cultivars/lines for 14 heading-related markers or 1,594 markers as the input of a machine learning model for predicting the parameter values. Estimates of the parameters were used as the output for training a machine learning model (step 1b). Then, we connected the genetic effect on the parameters to the crop model in step 2 and predicted the heading date of a given marker genotype under the target environment.

Cross-Validation

Three types of cross-validation (CV) were performed to compare the prediction ability of the different methods. The first is a fivefold CV that was applied to compare the performance of the CGM with the machine learning methods when information on all genotypes and locations are available. This is a scheme used to validate the accuracy of prediction for tested genotypes under tested locations. In a breeding program, we usually do not test the full set of genotypes across all the environments. The prediction under this scheme therefore allows breeders to predict the DTH of "untested combinations" of tested genotypes under the tested locations. Based on the prediction, breeders can evaluate the potential adaptation of a tested genotype to a tested target location.

The second is the leave-one-genotype-out (LOGO) CV. In this scheme, from among the 112 genotypes, one genotype is removed from the data and the model is trained to predict the DTH for the removed genotype. The process is repeated until each genotype has been removed and predicted once. The predictions under this scheme allow breeders to predict the DTH of new lines (or even simulated marker genotypes) under the tested locations. Based on the prediction, breeders can evaluate the potential adaptation of an untested genotype (e.g., lines under development) to a tested target location based on the marker genotype of the untested genotype. The LOGO CV was only applied to machine learning methods, and the integrated approach as the crop model requires the data of the target genotype to estimate the model parameters.

The third is the leave-one-combination-of-genotype-and-location-out (LOGLO) CV. In this scheme, one of the eight locations and one of the 112 cultivars/lines were removed from the data, and the DTH of the removed genotype in the removed location is predicted using the prediction model derived from the data comprising 111 genotypes in 7 locations. This is a scheme to validate the accuracy of prediction of untested genotypes under untested locations. The prediction

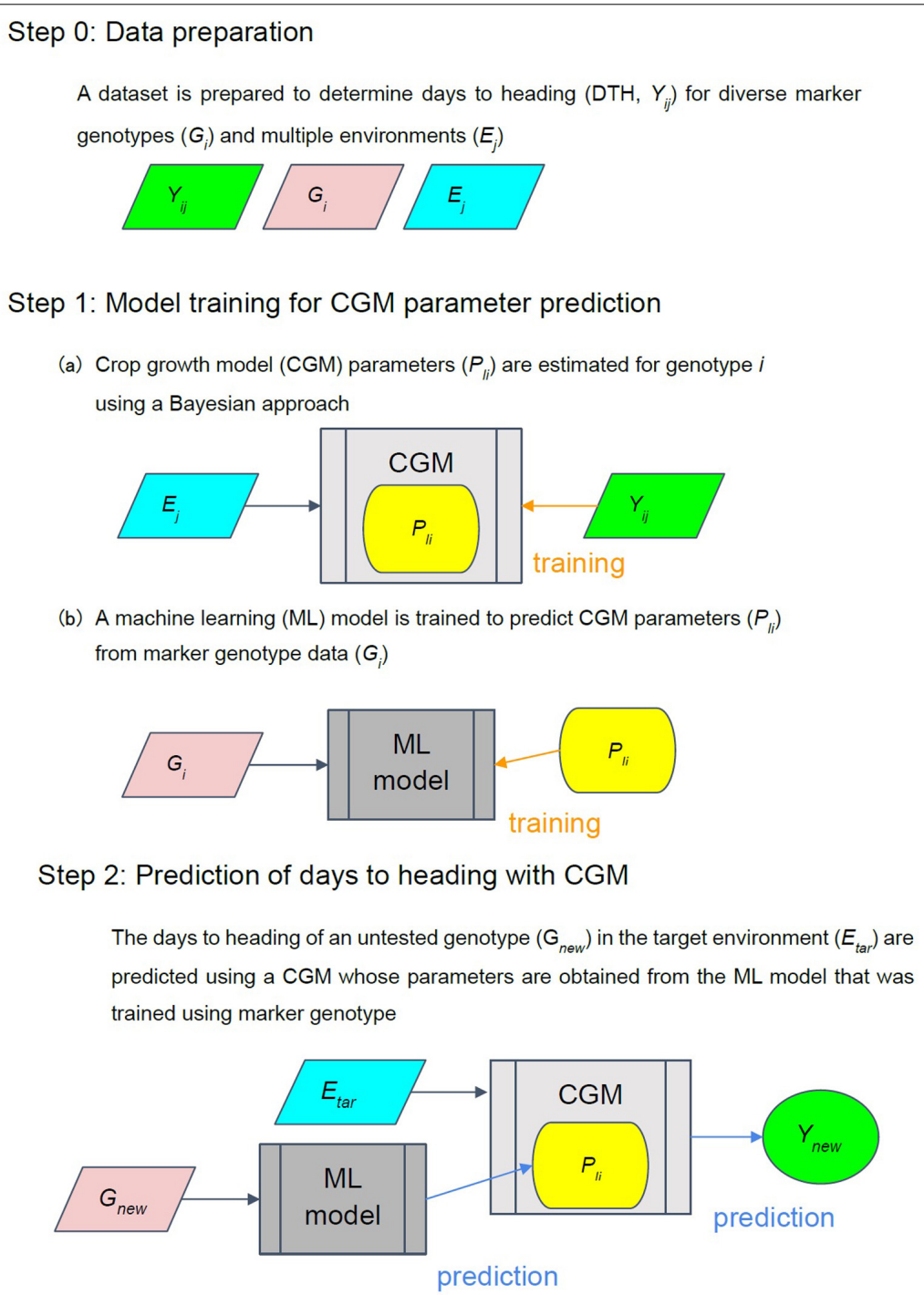

FIGURE 1 | Integrated approach concept. E_j is the environmental data that comprises the daily average temperature and daily photoperiod in the j-th environment (from the date of seeding to the date of heading + 70 days). Y_{ij} is the observed phenotypic trait (heading date) for the i-th genotype in j-th environment ($j = 1, 2, \ldots, n$). P_{li} is the l-th CGM parameters for the i-th cultivars/line. G_i is a vector of marker genotypes of the i-th cultivar/line.

under this scheme allows us to predict the DTH of new breeding lines (or simulated marker genotypes, as demonstrated in this study) to an untested target environment (e.g., expected environmental conditions in the future) based on marker genotypes of the untested genotype and environmental data of the untested environment.

In each CV scheme, the predicted DTH was obtained for each genotype in each environment (combination of

location, year, and treatment of different sowing dates). We then compared the prediction ability among the modeling methods based on the RMSE between the predicted and observed DTH.

Prediction of DTH in the F_2 Segregation Populations

The selection of a good parental combination that has a high probability of generating offspring with desired characteristics is important in breeding (Iwata et al., 2013). The better prediction of DTH in the F_2 segregation population can help the breeder to choose the best parental combination to generate progeny with the desired DTH prior to crossing. This can greatly reduce the cost and increase the efficiency of breeding. To demonstrate the potential of the integrated approach, we implemented the integrated model CGM-XGB trained by parents' data (the same data of 112 cultivars) and predicted the DTH in the derived F_2 segregation populations (created from the crossing of selected parental combinations of 112 cultivars) grown under an untested location, Kasai. In Kasai, 103 F_2 segregation populations derived from a common parent, Koshihikari, and 103 cultivars/lines were planted in 2008 and 2009 with 73 and 30 populations, respectively. We evaluated 96 F_2 individuals of each segregation population and measured their heading date to determine the distribution of DTH in the population. To predict the distribution of DTH in the segregation populations, we simulated the genotype marker data of F_2 segregation populations, and then predicted the heading date of simulated genotypes at an untested location with environmental data. The genotype marker data of progeny in an F_2 segregation population can be simulated from the genotype marker data of their parents and the estimated recombination rates between markers. In this study, we simulated 1,000 progeny for 14 markers of heading date-related genes, and applied the genotype data to the CGM-XGB model constructed based on the data of 112 cultivars/lines to predict the segregation distribution of DTH in the F_2 population. We considered the range of DTH of the F_2 segregation population for the selection of progeny with a reasonable value. We therefore compared the 10, 50, and 90th quantiles of the predicted and observed DTH.

RESULTS

Estimation of the DVR Model Parameter

We implemented a Bayesian method for estimating the CGM parameters in this study. The Bayesian method provided us with an approximated posterior distribution that was more informative than the point estimation obtained from the frequentist method. **Table 3** shows the average of the posterior mean, median, and mode of the CGM parameters (α, β, and G) among the cultivars/lines from each origin. The average of the median and mean values of the cultivars/lines from the same origin are similar; however, the average of the mode occasionally deviates from the average of the mean. This tendency is mainly because of the multimodal posteriors induced by the correlation between the CGM parameters. Therefore, the mean of the approximated posterior distribution could be more appropriate to describe the phenological features of a genotype. The genotypes from the high latitude origins, such as Hokkaido, Tohoku, and Hokuriku, have less photoperiod sensitivity and are expected to have a smaller estimated value of β. In contrast, a larger β should be observed for the photoperiod sensitive genotypes, mostly from the south of Japan, such as Kinki, Chugoku, and Kyushu. The average values of the posterior mean of β were approximately 1.01–1.48 for high latitude origins and 4.76–5.66 for low latitude origins. For α, we can find fewer differences between the average value of the posterior mean among origins (all average values were approximately 0.8–1.2). This probably reflects that the heading date in rice is more sensitive to the variation in photoperiod rather than the variation in temperature under the usual conditions. For parameter G, the smallest average posterior mean could be observed for the genotype of Northeast origins, Hokkaido (47.6 days) and a larger average value for the genotypes from Tohoku and Hokuriku (61.4 and 59.6 days), which are south of Hokkaido and in the north of Japan. The genotypes from the other origins had a similar average posterior mean of around 52.9–55.4 days.

Comparison of Prediction Ability Between the Methods

The comparison between the prediction ability of the CGM, the machine learning approach, and the integrated approach for the heading date of rice is summarized in **Table 4**. We first evaluated the prediction of the heading date of a tested cultivar/line in a tested location through a fivefold CV process. The machine learning approach using XGB had a smaller RMSE (4.372 and 2.653 for the model using the environmental data with 14 heading date-related markers and 1,594 markers data, respectively, as input) than the CGM (5.711 for the DVR with the parameters estimated by the Bayesian approach with the DREAM algorithm). This shows that the use of environmental data and genetic data combined with the powerful machine learning method can better predict the heading date of the tested cultivar/line in a tested environment than the CGM alone. We then evaluated the prediction of the heading date of an untested genotype in a tested location using the LOGO CV process. As described in "Materials and Methods" section, the CGM requires parameters that are genotypic specific and is unable to make such predictions. The machine learning method XGB had a better predictive ability (RMSE = 5.02 and 4.468 with 14 heading date-related markers and 1,594 markers, respectively) in LOGO CV than the integrated approach (RMSE = 6.47 and 9.05 with 14 heading date-related markers and 1,594 markers, respectively). This shows that the single machine learning method could be a better predictor when the environmental data is included and the genotypic data are removed from the training data. The integrated approach achieved the prediction of the heading date of an untested genotype via the estimation of the CGM parameters and then via the fitting of the CGM with the estimates. Both the bias in predicting the CGM parameters and the adoption of relatively simple functions in the CGM compared to the more complex and flexible machine learning methods could be responsible for the relatively poor predictability in the integrated approach.

TABLE 3 | Average of the posterior statistics among 112 cultivars/lines from seven different origins.

Parameters in DVR	Posterior	Cultivar origins						
		Hokkaido (9)	Tohoku (26)	Kanto and Tokai (24)	Hokuriku (14)	Kinki and Chugoku (9)	Kyushu (11)	Landrace and others (19)
α	mean	0.792	1.062	1.141	1.224	0.944	0.891	0.976
	median	0.789	1.063	1.149	1.225	0.953	0.911	0.977
	mode	0.781	1.071	1.031	1.219	0.735	0.592	0.853
β	mean	1.478	1.015	3.994	1.336	4.764	5.652	4.499
	median	1.419	0.917	3.853	1.282	4.537	5.144	4.361
	mode	1.019	0.427	5.948	1.116	6.632	9.810	5.315
G	mean	47.559	61.432	55.236	59.62	55.435	54.788	52.916
	median	47.653	61.619	55.693	59.727	56.195	56.497	53.18
	mode	47.430	62.403	45.357	55.98	50.004	42.844	49.600

The parameters α, β, and G in the DVR model represent the temperature sensitivity coefficient, the photoperiod coefficient, and the earliness of flowering under the optimal condition, respectively.

TABLE 4 | Root mean square errors (RMSE) of the three prediction methods used.

	Crop growth model	Machine learning		Integrated approach	
	DVR[a]	XGB[b]		CGM-XGB[c]	
		14H	1,594	14H	1,594
Fivefold[d]	5.711	4.372	2.653		
LOGO[e]		5.025	4.468	6.471	9.050
LOGLO[f]		9.361	8.573	7.690	9.793

[a]*DVR: DVR model with Bayesian DREAM Markov chain Monte-Carlo algorithm.*
[b]*XGB: gradient boosting method.*
[c]*CGM-XGB: the integrated approach combining DVR with XGB.*
[d]*Fivefold: fivefold cross-validation.*
[e]*LOGO: leave-one-genotype-out cross-validation.*
[f]*LOGLO: leave-one combination-of-genotype-and-location-out cross-validation. 14H represents 14 heading-related markers. 1,594 represents 1,594 markers, including the 14 heading-related markers.*

The integrated approach shows its superiority in predicting the untested genotype in the untested location in the LOGLO CV process. The integrated approach adopted the Bayesian approach for the estimation of CGM parameters and trained an XGB model for predicting the parameters from genotype markers in step 1. Then, the heading date was predicted with the CGM of the predicted parameters in step 2. The procedure of this prediction, abbreviated as CGM-XGB, had the best predictive ability (RMSE = 7.69 when using 14 heading-related markers in machine learning) compared to a simple XGB model (RMSE = 9.361 and 8.537 for the model using the environmental data with 14 heading-related markers and 1,594 markers data, respectively, as input). In LOGLO CV, the information of the tested genotype and the tested location are removed from the training data, leading the predictor trained by the machine learning method to be more specific to the involved regions only. In contrast, the CGM quantifies the response of a plant to environmental factors using non-linear mechanical equations, which are more simplified but could be more robust in the prediction under a more uncertain condition.

TABLE 5 | Root mean square errors (RMSE) of the integrated approaches involving three different machine learning methods.

Methods in step 1b of the integrated approaches[b]	Methods in step 1a of the integrated approaches[a]		
	Bayesian with normal dist. prior		
	Marker	LOGO[c]	LOGLO[d]
ELM	14H	6.566	7.731
XGB	14H	6.574	7.776
RF	14H	6.817	8.038
ELM	1,594	18.627	19.087
XGB	1,594	9.552	10.658
RF	1,594	7.716	8.528

[a]*Step 1 in the integrated approaches was to estimate the crop growth model parameters using the rice heading date data and the environmental data.*
[b]*Step 2 in the integrated approaches was to train a machine learning model to predict the CGM parameters of an unknown genotype.*
[c]*LOGO: leave-one-genotype-out cross-validation.*
[d]*LOGLO: leave-one-combination-of-genotype-and-location-out cross-validation. ELM, extreme learning machine; XGB, eXtreme gradient boosting; RF, random forest.*

Table 5 shows the results of the integrated approaches that were implemented with the combinations of three machine learning methods (RF, XGB, and ELM), and two sets of genotype marker data (14 heading-related markers and 1,594 markers). First, we found that the model with 14 heading date-related markers had better prediction ability than the model with 1,594 markers, which also included the 14 heading-related markers. The lower prediction ability in the model with a larger number of markers could be attributed to the inclusion of markers irrelevant to phenological growth and the lack of training data for the target genotype. $\alpha\beta G$ Second, the adoption of a different machine learning method could affect the prediction ability. The rank of the ability in the model was XGB > ELM > RF with 14 heading date-related markers, and RF > XGB > ELM with the 1,564 markers. It reveals

that XGB and ELM could be a better predictor of CGM parameters when less noise is present in the input data (14 heading date-related markers), whereas RF is relatively robust to the input data with noise. ELM could be greatly affected by the noise in the input data and even provided a highly deviated estimation of the parameters. Such problems could be found in the especially large RMSE of ELM in the model with 1,594 markers. Among all combinations of methods in the integrated approaches, XGB in step 1b in the integrated approach had the best prediction ability in both the LOGO and LOGLO CV processes.

Predicting DTH Distribution in F_2 Segregation Populations

We examined the ability of the proposed integrated model CGM-XGB in predicting the distribution of DTH in 103 F_2 segregation populations. **Figure 2** shows the scatterplot of the 10, 50, and 90th percentiles of the observed distributions and predicted distributions. The predicted RMSE, correlation coefficient, and absolute mean difference are also shown in **Figure 2**. The percentiles of the predicted DTH distribution tended to be underestimated in comparison to the percentiles of the observed DTH distribution for most populations. The correlation coefficients were mostly over 0.8, and showed that the integrated approach could be useful in predicting the rank of the percentiles of distribution in DTH between different F_2 segregation populations. A slightly better prediction was found in the 30 populations tested in 2009 than in the 73 populations tested in 2008, although the reason for this is unclear. Histograms of the observed and predicted distributions in DTH for each segregation population are shown in **Supplementary Figures S1–S3**.

DISCUSSION

In this study, we proposed a potential integrated approach that combines machine learning methods and a CGM to improve the modeling of physiological growth of rice plants. We emphasize the importance of the training data for the successful building of the model. Phenotypic and environmental data consisting of a wide range of genotypes grown in multiple environments is a prerequisite for the proposed approach. In this study, 112 cultivars/lines were selected from among those adapted to different ecological regions in Japan (Yamasaki and Ideta, 2013) and had been evaluated at these locations for more than 10 years. Such comprehensive data allows us to estimate the phenological parameters in a CGM with less estimation bias and mitigates the bias induced by the location effect and makes it possible to associate the marker effect with the model parameters. In addition, the real data of the F_2 segregating populations presents opportunities to validate the predictability of the model for predicting the potential of a cross to develop a new cultivar/line for a new environment. In previous studies, this validation was mostly conducted using a simulation study or cross-validation that might not reflect the true performance of the proposed model. The power of machine learning methods is in addition to the quality of the training data. Because more information can be collected from high-throughput phenotyping, genotyping, environmental sensing, and omics analyses, more attention can be paid to the data rather than only the methodologies.

Estimating the parameters of the CGM appropriately is essential for the prediction accuracy of a model and for further inference that utilizes the predicted model parameters. Despite prior knowledge, the estimation method can influence the results; therefore, the best strategy to conduct such estimation remains open for discussion. For parameter estimation, we implemented both frequentist and Bayesian approaches and showed no obvious difference in the prediction accuracy of the CMG (results not shown here). This might mainly result from the substantial and complete heading data collected in this study, which provides sufficient information for parameter estimation.

The parameters α, β, and G in the DVR model represent the temperature sensitivity coefficient, the photoperiod coefficient, and the earliness of flowering under the optimal condition, respectively. In this study, we obtained not only the point estimated value but also the approximated posterior distribution of these three parameters for 112 Japanese rice cultivars. This information allowed us to first examine the phenological characteristics of the most representative cultivars quantitatively and use them in building the integrated model. The characteristics of most Japanese rice cultivars, including the tendency of photoperiod sensitivity, can be found in a database[1]. We compared the tendency of photoperiod sensitivity of the tested cultivars/lines to the posterior mean of β, and the results were mostly matched (results not shown here). In addition, parameter estimation using the Bayesian method also matched our knowledge regarding the character of a genotype and might better quantify the indirect features of the phenological growth of rice.

As shown in **Table 3**, the β of cultivars originating in high latitude regions was close to 1, indicating the strong tendency of photoperiod insensitivity and vice versa. The results are consistent with those of a previous study (Okumoto et al., 1996) and rice photoperiod sensitivity is generally diverse (Hori et al., 2016). All five heading date-related genes in this study are associated with the rice photoperiodic pathway (Yano et al., 2000; Takahashi et al., 2001; Xue et al., 2008; Matsubara et al., 2012; Hori et al., 2013). Although temperature is also an essential factor in predicting rice growth, the variation in α was small, suggesting that the diversity of the thermal reaction among the 112 cultivars may be small. In combination with the estimation of α and G, it presented the possible coordination between thermal reaction, photoperiod sensitivity, and the earliness of flower initiation that helps the corresponding cultivar to adapt to the target environment.

Compared to the results achieved by machine learning methods in other fields in agriculture, such as crop management

[1] https://ineweb.narcc.affrc.go.jp/

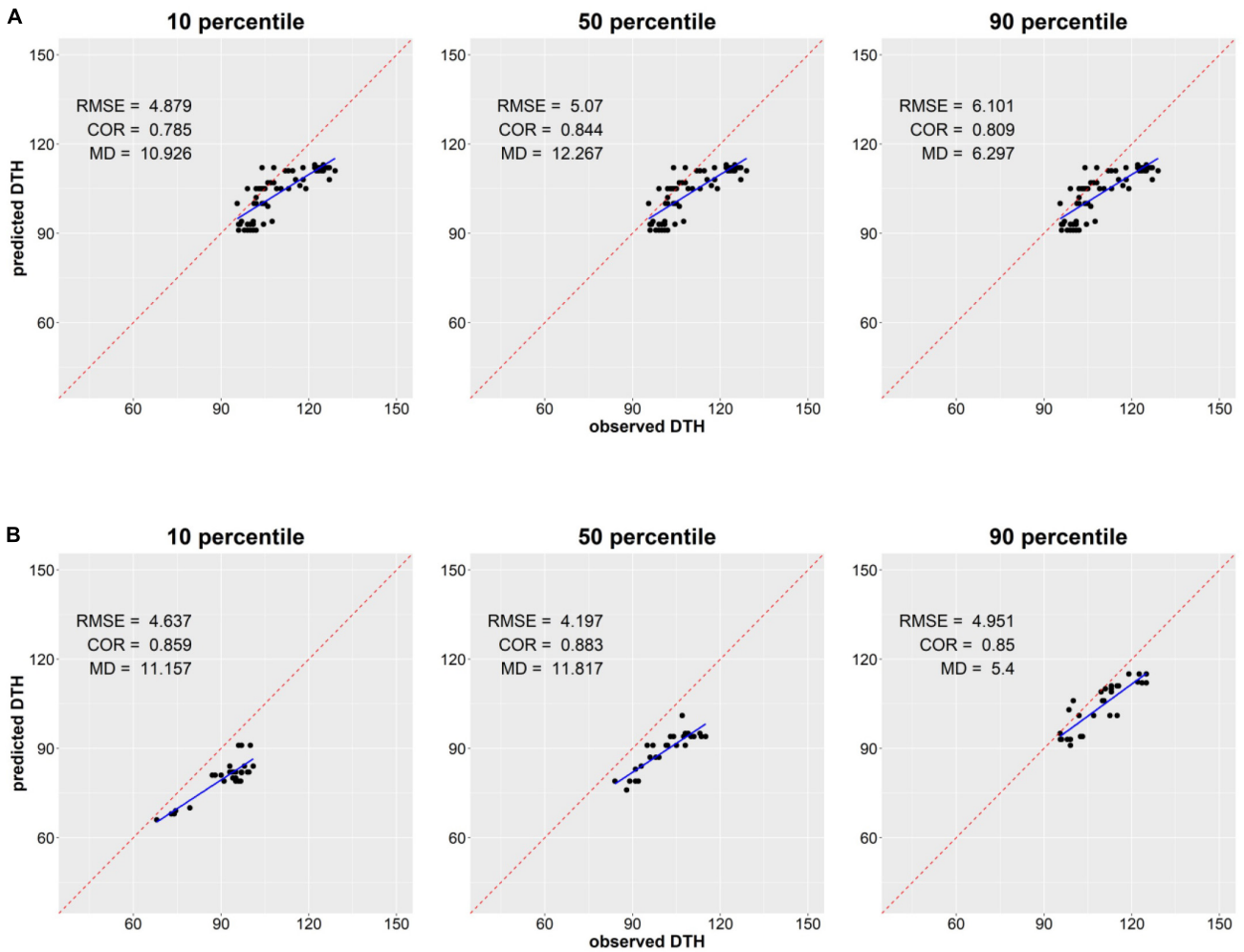

FIGURE 2 | Scatterplot of the 10, 50, and 90th percentiles of observed and predicted distribution of days to heading (DTH) of segregation populations in 2008 **(A)** and in 2009 **(B)**. The observed DTH was obtained from the 96 plants of each population, and the predicted DTH was generated from the proposed integrated approach CGM-XGB. RMSE: root mean square error; COR: correlation coefficient; MD: mean absolute difference.

and water management (Liakos et al., 2018), examples of successful applications in crop breeding and genetics are still relatively rare. The fundamental reason is not only the complexity of the genotype × environment × management interaction, but also the unfamiliarity of the method, the lack of adequate data, and the few experts who are familiar with both fields. We compared the use of a CGM, machine learning models, and integrated approaches in predicting rice heading. The results showed that the machine learning model with the genotypic marker was more accurate than the CGM in predicting the heading of a tested cultivar/line in a tested location. We also compared the predictability of three machine learning methods: RF (a popular ensemble learning method), ELM (a feed-forward neural network), and XGB (a modified gradient boosting method), and showed the advantages of applying the newly developed algorithm. It is not surprising that the machine learning methods were capable of better capturing the complex and non-linear association between complicated traits and genetic and environmental variables. However, at the same time, a machine learning method could yield worse predictions than a mechanistic CGM if the training data is limited or full of noise. We also showed that the machine learning models were less applicable for predicting the extrapolation problem, such as the prediction of the heading of untested genotypes in an untested location that can be predicted better by the proposed integrated approach. However, both CGM and the machine learning model could be useful for cultivation management, such as supporting the decision on the suitable sowing timing for an optimal heading date.

The three machine learning methods (RF, XGB, and ELM) compared in this study have proven their superiority in many machine learning challenges, and the implemented packages are

already available to run on many platforms. Although XGB and ELM had slightly better predictability than RF in our results, there is no guarantee that one method could outperform others in a different scenario. The experimental design, training data, and setting of hyperparameters sometimes play an important role in practical applications. In addition, factors such as (1) suitability to a given setting, (2) computational cost, (3) software availability, and (4) usability, may be considered when selecting the best method (Sagi and Rokach, 2018). In addition, the lack of interpretability in most machine learning methods could be an issue when we apply them to biological problems. For example, the machine learning model in our integrated approach could not provide an intuitive understanding of the underlying gene regulation of rice heading. The development of interpretable machine learning methods might be helpful in the future when both predictability and interpretability are needed.

Using 112 cultivars, the integrated model CGM-XGB simulated and predicted the distributions of DTH in 103 F_2 segregation populations. The predicted distributions of DTH were generally similar to those observed in the real data. Based on the prediction of DTH in a segregating population in an environment and management system before producing crosses, breeders can consider the optimum cross combinations to develop a novel cultivar. In addition, a recent serious event, high temperature during the rice ripening period resulted in deterioration of the grain quality in Japan (Morita, 2009). The models explored in this study can propose the ideal heading date and sowing timing in a cultivar to avoid such damage.

CONCLUSION

The capability of the proposed integrated approach in predicting the heading of a new genotype in a new environment was demonstrated, and this could prove useful in suggesting the locally adapted ideotype for rice phenology. We also revealed that the machine learning model could outperform the crop growth model (CGM) (phenological model without genotypic data) in predicting the heading of a tested cultivar/line in a tested environment and could be replaced with a phenological model when higher accuracy is preferred. However, the machine learning model is highly dependent on the given data and is usually less capable of extrapolating, as demonstrated by the Leave-one-genotype-out cross-validation (LOGLO CV) results. It is also difficult to dissect the machine learning model and reveal the explanatory mechanisms underneath the model, as can be done with the CGM. The CGM models the key physiological processes of crop growth, and the inclusion of CGM into the modeling platform can reduce the uncertainty when simulating crop growth. This study confirmed that the integrated approach improved the prediction of the complex trait for a new genotype in a new location and may benefit crop selection.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The dataset analyzed in this study is not publicly available due to parallel studies but may be available from the corresponding author upon reasonable request. Requests to access these datasets should be directed to Hiroyoshi Iwata, hiroiwata@g.ecc.u-tokyo.ac.jp.

AUTHOR CONTRIBUTIONS

T-SC developed the methodology, analyzed the data, wrote the manuscript, and developed the software. MY designed and conducted the experiments and provided the rice heading data and marker data. HK-K prepared and provided marker data. TA conducted the pioneer study and was involved in methodology development. HI was involved in the conceptualization, methodology development, and supervision of the study. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the MEXT KAKENHI, Japan, Grant Nos. JP25252005 and 15H04436, and partly by JST CREST, Japan, Grant Nos. JPMJCR16O2 and JPMJCR17O3. This study was partially funded by the Indo-Japan DST-JST SICORP program "Data Science-based Farming Support System for Sustainable Crop Production under Climatic Change" from the Japan Science and Technology Agency.

ACKNOWLEDGMENTS

We sincere thanks to the following members: Osamu Ideta, Tomomori Kataoka, Narifumi Yokogami, Ryota Kaji, Hideo Maeda, Kazumasa Murata, and Hiroshi Nakagawa for helping with the phenotyping work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.599510/full#supplementary-material

REFERENCES

Biau, G., and Scornet, E. (2016). A random forest guided tour. *Test* 25, 197–227.

Bogard, M., Ravel, C., Paux, E., Bordes, J., Balfourier, F., Chapman, S. C., et al. (2014). Predictions of heading date in bread wheat (*Triticum aestivum* L.) using QTL-based parameters of an ecophysiological model. *J. Exp. Bot.* 65, 5849–5865. doi: 10.1093/jxb/eru328

Breiman, L. (2001a). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: the two cultures. *Stat. Sci.* 16, 199–215. doi: 10.1214/ss/1009213726

Chapman, S. C., Cooper, M., and Hammer, G. L. (2002a). Using crop simulation to generate genotype by environment interaction effects for sorghum in water-limited environments. *Aust. J. Agric. Res.* 53, 379–389. doi: 10.1071/AR01070

Chapman, S. C., Hammer, G., Podlich, D., and Cooper, M. (2002b). "Linking biophysical and genetic models to integrate physiology, molecular biology and plant breeding," in *Quantitative Genetics, Genomics and Plant Breeding*, ed. M. S. Kang (Wallingford: CABI), 167–187. doi: 10.1079/9780851996011.0167

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 785–794. doi: 10.1145/2939672.2939785

Chen, T., and He, T. (2015). *Xgboost: eXtreme Gradient Boosting. R Package Version 0.4-2*. 1–4.

Dumont, B., Leemans, V., Mansouri, M., Bodson, B., Destain, J. P., and Destain, M. F. (2014). Parameter identification of the STICS crop model, using an accelerated formal MCMC approach. *Environ. Model. Softw.* 52, 121–135. doi: 10.1016/j.envsoft.2013.10.022

Forsythe, W. C., Rykiel, E. J. Jr., Stahl, R. S., Wu, H. I., and Schoolfield, R. M. (1995). A model comparison for daylength as a function of latitude and day of year. *Ecol. Model.* 80, 87–95. doi: 10.1016/0304-3800(94)00034-F

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38, 367–378.

Hori, K., Matsubara, K., and Yano, M. (2016). Genetic control of flowering time in rice: integration of Mendelian genetics and genomics. *Theor. Appl. Genet.* 129, 2241–2252. doi: 10.1007/s00122-016-2773-4

Hori, K., Ogiso-Tanaka, E., Matsubara, K., Yamanouchi, U., Ebana, K., and Yano, M. (2013). *H D16*, a gene for casein kinase I, is involved in the control of rice flowering time by modulating the day-length response. *Plant J.* 76, 36–46. doi: 10.1111/tpj.12268

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: algorithm, theory and applications. *Neurocomputing* 70, 489–501. doi: 10.1016/j.neucom.2005.12.126

Iizumi, T., Tanaka, Y., Sakurai, G., Ishigooka, Y., and Yokozawa, M. (2014). Dependency of parameter values of a crop model on the spatial scale of simulation. *J. Adv. Model. Earth Syst.* 6, 527–540. doi: 10.1002/2014MS000311

Iizumi, T., Yokozawa, M., and Nishimori, M. (2009). Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: application of a Bayesian approach. *Agric. For. Meteorol.* 149, 333–348. doi: 10.1016/j.agrformet.2008.08.015

Iwata, H., Takeshi, H., Shingo, T., Norio, T., Toshihiro, S., and Toshiya, Y. (2013). Genomic prediction of trait segregation in a progeny population: a case study of Japanese pear (*Pyrus pyrifolia*). *BMC genetics* 14:81. doi: 10.1186/1471-2156-14-81

Jones, J. W., He, J., Boote, K. J., Wilkens, P., Porter, C. H., and Hu, Z. (2011). "Estimating DSSAT cropping system cultivar-specific parameters using Bayesian techniques," in *Methods of Introducing System Models into Agricultural Research*, Vol. 2, eds L. R. Ahuja and L. Ma (Madison, WI: SSSA), 365–394. doi: 10.2134/advagricsystmodel2.c13

Letort, V., Mahe, P., Cournède, P. H., De Reffye, P., and Courtois, B. (2008). Quantitative genetics and functional-structural plant growth models: simulation of quantitative trait loci detection for model parameters and application to potential yield optimization. *Ann. Bot.* 101, 1243–1254. doi: 10.1093/aob/mcm197

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine learning in agriculture: a review. *Sensors* 18:2674. doi: 10.3390/s18082674

Liaw, A., and Wiener, M. (2002). Classification and regression by RandomForest. *R News* 2, 18–22.

Makowski, D., Hillier, J., Wallach, D., Andrieu, B., and Jeuffroy, M.-H. (2006). "Parameter estimation for crop models," in *Working with Dynamic Crop Models*, *Evaluation, Analysis, Parameterization and Applications*, eds D. Wallach, D. Makowski, and J. W. Jones (Amsterdam: Elesevier), 55–100.

Matsubara, K., Ogiso-Tanaka, E., Hori, K., Ebana, K., Ando, T., and Yano, M. (2012). Natural variation in Hd17, a homolog of Arabidopsis ELF3 that is involved in rice photoperiodic flowering. *Plant Cell Physiol.* 53, 709–716. doi: 10.1093/pcp/pcs028

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

Morita, S. (2009). Eco-physiological analysis for high-temperature effects on rice-grain ripening. *Bull. Natl. Agric. Res. Center Kyushu Okinawa Region* 52, 1–78.

Nakagawa, H., Yamagishi, J., Miyamoto, N., Motoyama, M., Yano, M., and Nemoto, K. (2005). Flowering response of rice to photoperiod and temperature: a QTL analysis using a phenological model. *Theor. Appl. Genet.* 110, 778–786.

Okumoto, Y., Ichitani, K., Inoue, H., and Tanisaka, T. (1996). Photoperiod insensitivity gene essential to the varieties grown in the northern limit region of paddy rice (*Oryza sativa* L.) cultivation. *Euphytica* 92, 63–66. doi: 10.1007/BF00022829

Onogi, A., Watanabe, M., Mochizuki, T., Hayashi, T., Nakagawa, H., Hasegawa, T., et al. (2016). Toward integration of genomic selection with crop modelling: the development of an integrated approach to predicting rice heading dates. *Theor. Appl. Genet.* 129, 805–817.

Sagi, O., and Rokach, L. (2018). Ensemble learning: a survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 8:e1249. doi: 10.1002/widm.1249

Seidel, S. J., Palosuo, T., Thorburn, P., and Wallach, D. (2018). Towards improved calibration of crop models – where are we now and where should we go? *Eur. J. Agron.* 94, 25–35. doi: 10.1016/j.eja.2018.01.006

Takahashi, Y., Shomura, A., Sasaki, T., and Yano, M. (2001). Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the $α$ subunit of protein kinase CK2. *Proc. Natl. Acad. Sci. U.S.A.* 98, 7922–7927. doi: 10.1073/pnas.111136798

Technow, F., Messina, C. D., Totir, L. R., and Cooper, M. (2015). Integrating crop growth models with whole genome prediction through approximate Bayesian computation. *PLoS One* 10:e0130855. doi: 10.1371/journal.pone.0130855

Uptmoor, R., Schrag, T., Stützel, H., and Esch, E. (2008). Crop model based QTL analysis across environments and QTL based estimation of time to floral induction and flowering in *Brassica oleracea*. *Mol. Breed.* 21, 205–216. doi: 10.1007/s11032-007-9121-y

White, J. W., and Hoogenboom, G. (1996). Simulating effects of genes for physiological traits in a process-oriented crop model. *Agron. J.* 88, 416–422. doi: 10.2134/agronj1996.00021962008800030009x

Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L., et al. (2008). Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nat. Genet.* 40, 761–767. doi: 10.1038/ng.143

Yamasaki, M., and Ideta, O. (2013). Population structure in Japanese rice population. *Breed. Sci.* 63, 49–57. doi: 10.1270/jsbbs.63.49

Yano, M., Harushima, Y., Nagamura, Y., Kurata, N., Minobe, Y., and Sakaki, T. (1997). Identification of quantitative trait loci controlling heading date in rice using a high-density linkage map. *Theor. Appl. Genet.* 95, 1025–1032. doi: 10.1007/s001220050658

Yano, M., Katayose, Y., Ashikari, M., Yamanouchi, U., Monna, L., Fuse, T., et al. (2000). *Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene *CONSTANS*. Plant Cell 12, 2473–2483. doi: 10.1105/tpc.12.12.2473

Yin, X., Chasalow, S. D., Dourleijn, C. J., Stam, P., and Kropff, M. J. (2000). Coupling estimated effects of QTLs for physiological traits to a crop growth model: predicting yield variation among recombinant inbred lines in barley. *Heredity* 85, 539–549. doi: 10.1046/j.1365-2540.2000.00790.x

Yin, X., Kropff, M. J., Horie, T., Nakagawa, H., Centeno, H. G., Zhu, D., et al. (1997). A model for photothermal responses of flow- ering in rice I. Model

description and parameterization. *Field Crops Res.* 51, 189–200. doi: 10.1016/S0378-4290(96)03456-9

Yin, X., Stam, P., Kropff, M. J., and Schapendonk, Ad H. C. M (2003). Crop modeling, QTL mapping, and their complementary role in plant breeding. *Agron. J.* 95, 90–98. doi: 10.2134/agronj2003.9000a

Yin, X., Struik, P. C., Van Eeuwijk, F. A., Stam, P., and Tang, J. (2005). QTL analysis and QTL-based prediction of flowering phenology in recombinant inbred lines of barley. *J. Exp. Bot.* 56, 967–976. doi: 10.1093/jxb/eri090

Zięba, M., Tomczak, S. K., and Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Syst. Appl.* 58, 93–101. doi: 10.1016/j.eswa.2016.04.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 January 2021
doi: 10.3389/fgene.2020.609117

Prior Biological Knowledge Improves Genomic Prediction of Growth-Related Traits in *Arabidopsis thaliana*

*Muhammad Farooq[1,2], Aalt D. J. van Dijk[1,3], Harm Nijveen[1], Mark G. M. Aarts[4], Willem Kruijer[3], Thu-Phuong Nguyen[4], Shahid Mansoor[2] and Dick de Ridder[1]**

[1] Bioinformatics Group, Wageningen University, Wageningen, Netherlands, [2] Molecular Virology and Gene Silencing Lab, Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering (NIBGE), Punjab, Pakistan, [3] Biometris, Wageningen University, Wageningen, Netherlands, [4] Laboratory of Genetics, Wageningen University, Wageningen, Netherlands

OPEN ACCESS

Edited by:
Shiori Yabe,
Institute of Crop Science (NARO), Japan

Reviewed by:
Yongkang Kim,
University of Colorado Boulder, United States
Tian Qing Zheng,
Chinese Academy of Agricultural Sciences, China

***Correspondence:**
Dick de Ridder
dick.deridder@wur.nl

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: 22 September 2020
Accepted: 21 December 2020
Published: 20 January 2021

Citation:
Farooq M, van Dijk ADJ, Nijveen H, Aarts MGM, Kruijer W, Nguyen T-P, Mansoor S and de Ridder D (2021) Prior Biological Knowledge Improves Genomic Prediction of Growth-Related Traits in Arabidopsis thaliana. Front. Genet. 11:609117. doi: 10.3389/fgene.2020.609117

Prediction of growth-related complex traits is highly important for crop breeding. Photosynthesis efficiency and biomass are direct indicators of overall plant performance and therefore even minor improvements in these traits can result in significant breeding gains. Crop breeding for complex traits has been revolutionized by technological developments in genomics and phenomics. Capitalizing on the growing availability of genomics data, genome-wide marker-based prediction models allow for efficient selection of the best parents for the next generation without the need for phenotypic information. Until now such models mostly predict the phenotype directly from the genotype and fail to make use of relevant biological knowledge. It is an open question to what extent the use of such biological knowledge is beneficial for improving genomic prediction accuracy and reliability. In this study, we explored the use of publicly available biological information for genomic prediction of photosynthetic light use efficiency (Φ_{PSII}) and projected leaf area (PLA) in *Arabidopsis thaliana*. To explore the use of various types of knowledge, we mapped genomic polymorphisms to Gene Ontology (GO) terms and transcriptomics-based gene clusters, and applied these in a Genomic Feature Best Linear Unbiased Predictor (GFBLUP) model, which is an extension to the traditional Genomic BLUP (GBLUP) benchmark. Our results suggest that incorporation of prior biological knowledge can improve genomic prediction accuracy for both Φ_{PSII} and PLA. The improvement achieved depends on the trait, type of knowledge and trait heritability. Moreover, transcriptomics offers complementary evidence to the Gene Ontology for improvement when used to define functional groups of genes. In conclusion, prior knowledge about trait-specific groups of genes can be directly translated into improved genomic prediction.

Keywords: genomic prediction (GP), photosynthesis, phenomics data analysis, *Arabidopsis thaliana* (Arabidopsis), GBLUP, GFBLUP

INTRODUCTION

Due to breakthroughs in DNA sequencing technology over the past decade, high-throughput genotyping is now a routine practice in plant breeding (Rimbert et al., 2018). Phenotyping is undergoing a similar revolution: large phenomics facilities are being developed that can rapidly score large germplasm collections of plants in a range of different environments (Flood et al., 2016; Crain et al., 2018). These technological developments have made it possible to acquire datasets describing genotypes and phenotypes for large numbers of individuals at an extended temporal scale. Despite recent advances in phenomics it is still more expensive and laborious than genotyping. To make the most use of phenomic datasets, Genomic Selection (GS) based breeding programs aim to predict unobserved phenotypes of individuals based on genotypes alone. This has the twofold benefit of reducing breeding costs and speeding up breeding programs as plants can be genotyped in the seedling stage and selected accordingly, thus negating the need to grow large populations to maturity and scoring them all to obtain breeding values based on phenotypes. GS usually models the unobserved phenotypes as additive effects of all genetic markers (total additive genomic value or breeding value) in the test population using a genomic prediction (GP) model. This GP model is based on a reference population which has both been genotyped and phenotyped for the trait(s) of interest (Meuwissen et al., 2001). The performance of GP depends on many factors, including genetic architecture, reference population size and structure and heritability (Karaman et al., 2016). However, GP accuracy, usually defined as the correlation (Pearson's r) between observed phenotypes and predicted breeding values, is generally lower for complex traits than for simpler ones (Morgante, 2018). This is because such traits are affected by many loci with small to moderate effects, along with non-additive genetic (dominance, epistasis) and genotype-by-environment (GxE) interactions (Falconer and Mackay, 1996). Incorporating epistasis into GP models has been reported to improve performance in selfing plant species but may not work for outcrossing species; therefore, additive GP models are still the primary choice (Jiang and Reif, 2015).

In GP models, each individual's genetic or breeding value is modeled as the sum of additive marker effects. Despite advancements in phenomics, phenotyping data is still usually only available for a few traits of several hundreds of individuals (n), compared to millions of genetic markers (p). GP models tackle this curse of dimensionality ($p > n$) by regularization (Meuwissen et al., 2001). When marker effects are fixed, this comes in the form of a penalty term added to the log-likelihood, as in LASSO or ridge regression. More frequently, marker effects are considered random, and regularization is achieved through prior distributions on the marker effects. The variance in these priors is directly related to the heritability, and can be estimated either using REML, or a fully Bayesian approach. In the classical GBLUP-approach, a single normal distribution with equal variance is assumed for all marker effects (Vanraden, 2008). More recently, mixture distributions have been considered (Moser et al., 2015). The prior could e.g., be a mixture of Gaussian distributions with large and small variances, and a point mass at zero, allowing a marker to have respectively, large or small effects, or no effect at all (Macleod et al., 2016). Moreover, restrictions on the shape of the probability distribution, usually Gaussian, can be relaxed (e.g., t-distribution) to accommodate genetic architectures having a larger number of high to moderate effect sizes (Gianola, 2013) or another suitable distribution can be exploited instead. In spite of these refinements, it is usually impossible to find the true causal variants when $p > n$, which may lead to suboptimal prediction. Therefore, several authors suggested that *a priori* available biological knowledge may be incorporated in GP models, prioritizing likely causal markers, and ultimately improving prediction accuracy (Edwards et al., 2016; Ehsani et al., 2016; Wang et al., 2018).

Two types of biological knowledge have been considered in the literature: first, knowledge on biological properties of genes and their associated markers and second, knowledge in the form of secondary phenotypes. The latter typically concerns -omics data, and is modeled using additional relatedness matrices (Guo et al., 2016; Morgante, 2018; Azodi et al., 2020) or penalized selection indices (Lopez-Cruz et al., 2020). Although such -omics data can in principle be generated for the GP reference population, the use of more general publicly available information is often more feasible and cost-effective. We therefore focus on biological properties of genes and markers, such as Gene Ontology (GO) and post-GWAS QTL information. The GO provides a structured resource of functional classes of gene products based on orthology, represented into three biological domains, i.e., molecular function, cellular component and biological process (Ashburner et al., 2000). Similar functional groupings can be achieved from transcriptomic experiments based on the assumption that functionally related genes are expressed together. These clusters of co-expressed genes may be enriched in multiple GO terms or pathways. Such information can be incorporated by allowing the GP model to put more weight on either certain individual markers (Legarra and Ducrocq, 2012; Macleod et al., 2016) or groups of markers (Edwards et al., 2016). Various modeling approaches have been proposed to enable use of such data (Zhang et al., 2010; Speed and Balding, 2014; Edwards et al., 2016; Ehsani et al., 2016; Guo et al., 2016; Fragomeni et al., 2017). Here we use the Genomic Feature Best Linear Unbiased Predictor (GFBLUP) approach proposed by Edwards et al., 2016. GFBLUP extends GBLUP by partitioning the total genomic variance into two sub-components to weigh different genomic regions differently. This allows incorporating prior biological knowledge about groups of variants by treating each region as a separate random genetic effect with different variance. Subsequently, researchers applied this approach to various traits (Sarup et al., 2016; Fang et al., 2017; Rohde et al., 2017; Gebreyesus et al., 2019). While prior biological knowledge has thus been used to improve GP accuracy, the question remains what type of knowledge is most useful and how much the genetic architecture impacts the potential for improvement of particular traits.

In this study, we investigate improvement in GP performance using two sources of publicly available biological knowledge, i.e., Gene Ontology (GO) and clusters of co-expressed genes

(COEX). This information was incorporated using the GFBLUP modeling approach, grouping markers in genes according to either their predicted function or co-expression, respectively. As complex traits of study, we focused on photosynthetic light use efficiency of photosystem II (Φ_{PSII}) and projected leaf area (PLA) in *Arabidopsis thaliana*. Both of these traits are related, in the sense that the Φ_{PSII} directly illustrates the photosynthetic light use efficiency and can capture the most immediate physiological and regulatory response to varying irradiance levels (Van Rooijen et al., 2015), whereas growth in PLA is the net outcome of unit leaf photosynthetic capacity over time (Weraduwage et al., 2015; Liu et al., 2020).

RESULTS

Genomic Prediction of Complex Growth Related Traits

Previously, Van Rooijen et al. (2017) conducted a GWAS on *A. thaliana* photosynthesis. In particular, they measured the light use efficiency of photosystem II electron transport (Φ_{PSII}) for 344 accessions of the Arabidopsis HapMap population, switching from low light (100 μmol m^{-2} s^{-1}) to high light (550 μmol m^{-2} s^{-1}) irradiance at the onset of day 25. In total, they took 6 measurements before and 12 after applying light stress to identify potential QTLs during acclimation to high light. As we intend to use this population to explore the utility of biological knowledge in genomic prediction, we combined projected leaf area (PLA), another indicator of plant growth, with Φ_{PSII}. We first assessed whether GP works with reasonable performance for these complex traits. For this purpose, a classical Genomic Best Linear Unbiased Prediction (GBLUP) model was constructed to assess how well the infinitesimal modeling assumptions fit and to calculate markers-based heritability. In this model (Equation 2), all marker effects are treated as arising from a single normal distribution $N(0, \mathbf{G}\sigma_g^2)$ having one random genetic component, to regress each individual phenotype measurement over all markers simultaneously. At low light (LL) levels, mean prediction accuracy for Φ_{PSII} is lower (Pearson's r between predicted and observed phenotypic values ranging from 0.16 ± 0.02 to 0.22 ± 0.01) than at high light (HL, Pearson's r ranging from 0.40 ± 0.01 to 0.48 ± 0.01), as shown in **Figure 1A**. Prediction accuracy for PLA (**Figure 1B**) ranges from 0.06 ± 0.01 to 0.17 ± 0.01 and rises with the increase in plant size and simultaneously decreases with increase in phenotypic coefficient of variation. Genomic heritability (h^2_{GBLUP}) for Φ_{PSII} ranged from 0.08 to 0.13 under LL and 0.56 to 0.87 under HL, and 0.05 to 0.17 for PLA (**Supplementary Figure 1**). Differences in prediction accuracy for Φ_{PSII} between LL and HL are in line with differences in genomic heritability, in accordance with the observation that genomic prediction accuracy is generally positively correlated with heritabilities (Hayes et al., 2009). Moreover, for ∼1.2% of the GBLUP models for PLA, h^2_{GBLUP} was zero because of undetermined genomic variance, whereas for Φ_{PSII} ∼7% of genomic variances were estimated to be 100% ($h^2_{GBLUP} = 1$), which is clearly an over-estimation (**Supplementary Figure 2**). As reported by Kruijer et al., 2015, it was expected (based on 5000 simulated traits) that ∼10-15% of GBLUP models could have variance components that cannot be estimated for this population, so we discarded these models from our analysis.

An extension of GBLUP is MultiBLUP (Speed and Balding, 2014), using multiple random genetic components in the model (Equation 4), thus allowing differential weighting of groups of genomic markers, each having a separate kinship matrix derived from that group. We applied MultiBLUP using adjacent overlapping chromosomal partitions of 10 kb (yielding best performance when testing window sizes of 1 to 100 kb) to check if multiple kinship matrices or genomic variance decomposition improve prediction. The results (**Supplementary Figure 3**) indicate that performance was close to that of GBLUP and could not be improved further. This could be because most models ended up with only one background kinship matrix during cross-validation and many of these genomic regions did not meet the significance threshold ($p_{\text{Bonferroni}} < 0.05$) during association testing. In summary, these results show that predictive performance for these complex traits is low and there may be room for improvement by incorporating prior biological knowledge, decomposing the total genomic variance into biologically relevant subsets.

High-Level Biological Knowledge Does Not Necessarily Improve Genomic Prediction

The next question is whether predictive performance can be improved by using only markers residing within genes that are known to be linked to the traits of interest. The idea comes from previous studies, in which a subset of markers was associated to biological relevant genes and achieved a genomic value similar to the total genomic value achieved when using all SNPs (Vanraden et al., 2017; Li et al., 2018). Here, we selected 7,242 photosynthesis related genes, referred to as *PSGENES* in the text, from public repositories (see M&M) and constructed a GBLUP model based only on these. The Genomic Relationship Matrix (GRM) was constructed from all markers within the ORFs of *PSGENES*, leaving ∼17% of the total genotyped markers after filtering. Interestingly, the models performed equally well (**Figure 1**) as the GBLUP based on all markers for both traits, with a slight improvement in predictive ability for PLA (max. ∼6% increase in accuracy). Subsequently, to assess whether this pre-selected subset of markers can improve results if they are weighted differently than the rest of markers, we constructed another model using the GFBLUP modeling approach (Edwards et al., 2016) (Equation 3) having two genomic components. In this model, the markers within *PSGENES* were treated as one genomic component and the remaining markers as a second genomic component. Again, this model showed similar predictive performance as GBLUP, with some reduction in variability for PLA, but could not improve the accuracy further (**Figure 1**). From this, we conclude that prior biological knowledge-based selection of functionally relevant genes is potentially useful, but an optimal grouping may be important to improve GP further.

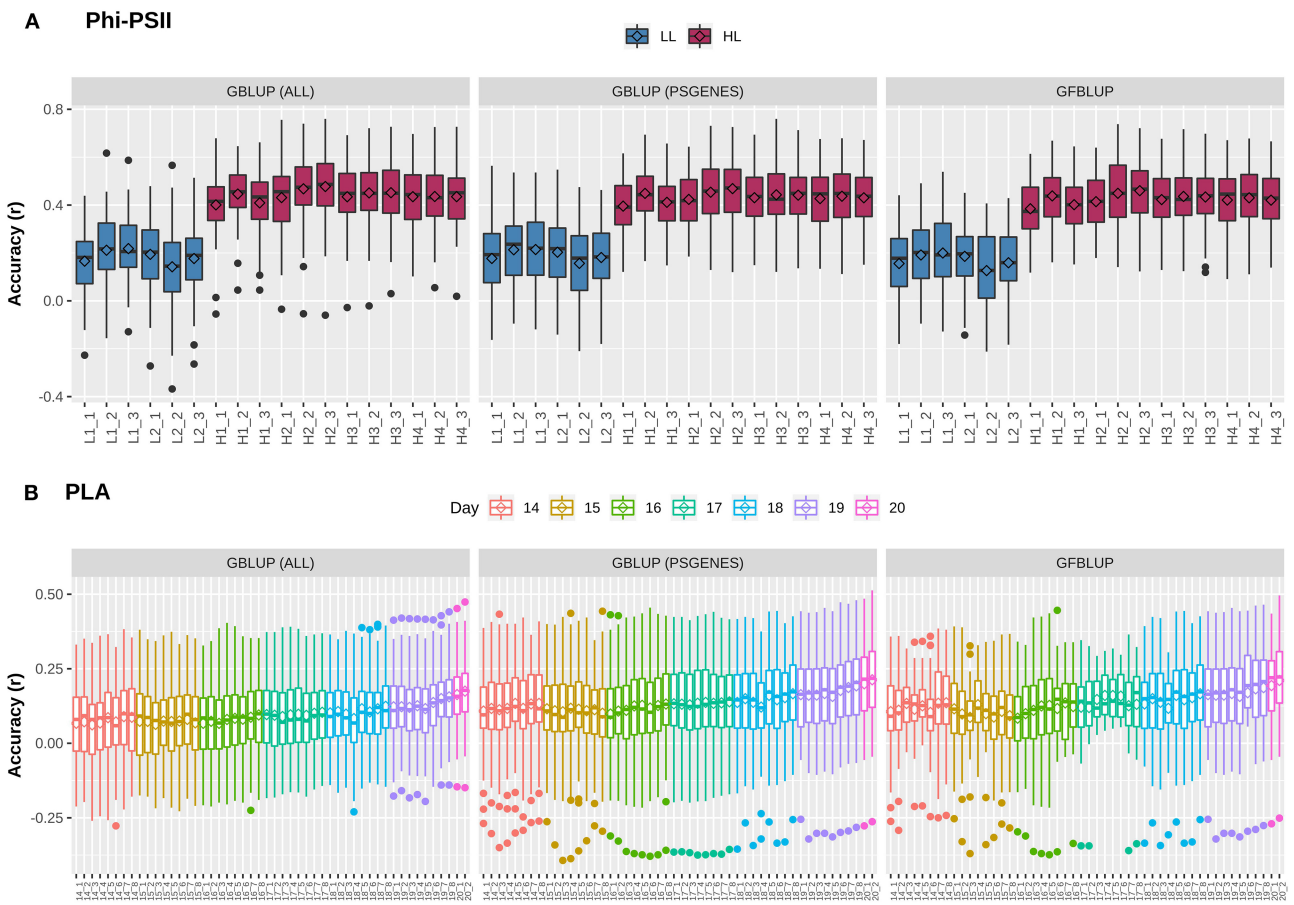

FIGURE 1 | Genomic prediction accuracy for photosynthetic light efficiency and projected leaf area. Cross-validation based assessment of prediction accuracy as Pearson's *r* between true and predicted values using three models: (i) GBLUP using all genomic markers (ALL), (ii) GBLUP using only pre-selected photosynthesis related genic markers (*PSGENES*) and (iii) GFBLUP using *PSGENES* as a genomic feature in two genomic components. **(A)** Accuracy for Φ_{PSII} for two days (6 time points) under low light irradiance levels (LL) and four days (12 time points) under high light irradiance (HL). **(B)** Accuracy for PLA measured 8 times per day from day 14 after sowing to day 20, where day 20 has only two measurements.

More Fine-Grained Biological Knowledge Is Useful for Improving Genomic Prediction

To assess whether prior information from publicly available resources can help improve GP performance, we tested grouping of genes based on Gene Ontology (GO) terms and previously reported clusters of co-expressed genes (COEX) of *Arabidopsis thaliana* in multiple tissues and developmental stages (Movahedi et al., 2011). Each of the three GO sub-ontologies, Biological Process (BP), Molecular Function (MF) and Cellular Component (CC), was used. The corresponding groups of markers in a GO or COEX group, called a genomic feature (GF), were used in GFBLUP (Equation 3) using a separate model for each feature with two genomic components, i.e., one with markers from the GF and the other with the remaining markers (rGF). The predictive performance was compared to that of the GBLUP benchmark using all markers with identical sets of 8-fold cross-validation test populations. Each group of markers based on GO or COEX was treated as a separate random effect in its respective GFBLUP model for which its contribution to the total genomic variance was calculated (see M&M). For each GF, the effects of all corresponding markers were assumed to follow a normal distribution with equal variance, but different from the remaining markers; that is, the markers in the GF are differentially weighted and prioritized from the rest.

In total, 7,297 GO terms and 12,419 disjoint COEX gene groups were linked to at least one marker. The total number of genes ranged between 1 and 24,998 for the GO features and between 1 and 3,384 for the COEX groups (**Supplementary Figure 4**, **Supplementary Table 4**); the number of markers ranged between 0 and 109,723 for the GO features and 4 and 19,621 for the COEX groups. Due to the hierarchical GO structure, the 95th percentile of the total number of genes within GO features was lower (496) as compared to COEX (2,466). Note that both GO and COEX groups may overlap, i.e., a gene can be in multiple functionally related GO/COEX groups. In the following results, the improvement in genomic prediction has been quantified in terms of percent gain in accuracy compared to the GBLUP benchmark, GFBLUP model's goodness

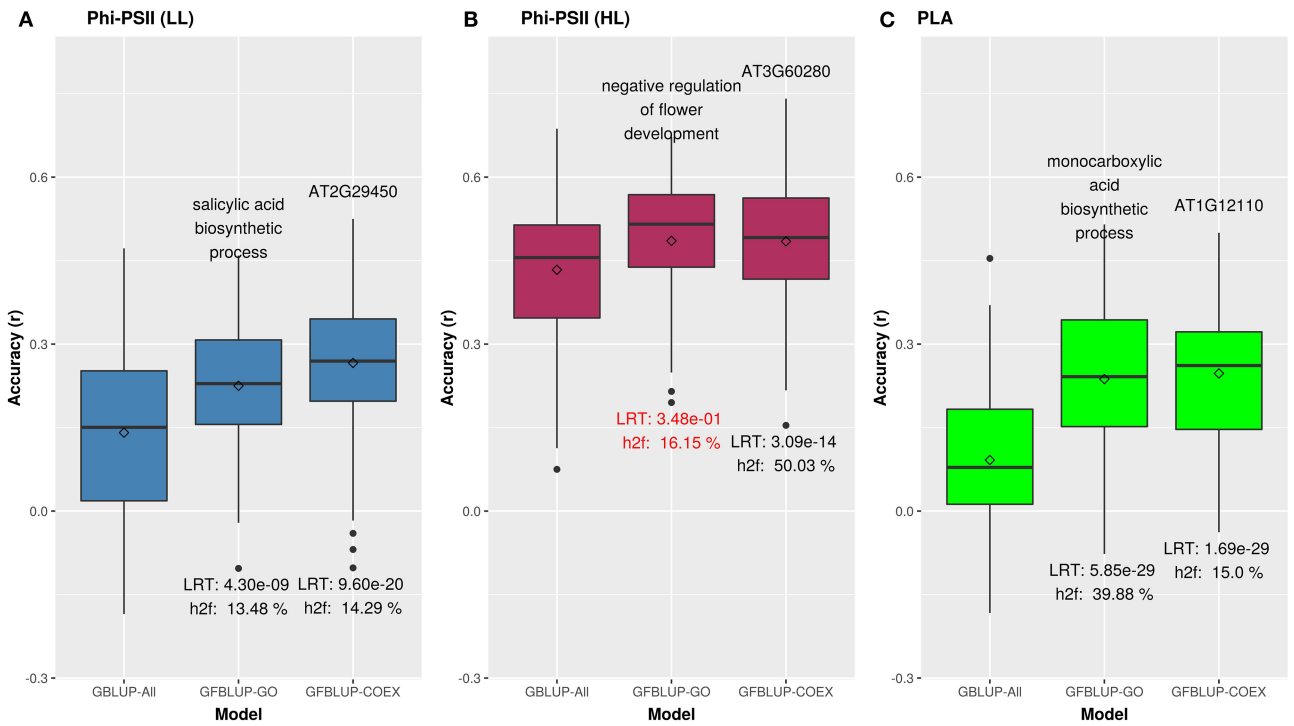

FIGURE 2 | Biological priors can help improve genomic prediction accuracy for growth related traits. Prediction accuracy of the best overall GFBLUP models using Gene Ontology (GO) and co-expression (COEX) gene groups, compared to the GBLUP benchmark (without prior biological knowledge). Since the GBLUP model was evaluated for each measurement point, so the GBLUP here is shown for the corresponding time point where improvement by GFBLUP was observed. Accuracy was calculated as Pearson's *r* between true vs. predicted values. The GBLUP-ALL model uses all markers in GBLUP; GFBLUP-GO and GFBLUP-COEX models use the top GO terms and COEX (see text for details). **(A)** Accuracy for Φ_{PSII} under low light irradiance levels (LL). **(B)** Accuracy for φ_{PSII} under high light irradiance (HL). Here, despite showing some improvement, the GFBLUP-GO model did not pass all of our model evaluation criteria (see Model Performance Evaluation). **(C)** Accuracy for PLA.

of fit measured using likelihood ratio test (LR), and genomic heritability (h^2_{GBLUP}) and proportion of genomic heritability explained by a genomic feature (h^2_f).

GO Informed Prediction

7,297 GO terms were tested with repeated 8-fold cross-validation at multiple measurements of a trait, leading to a total of ∼10 million GFBLUP model accuracies for Φ_{PSII} and ∼29 million for PLA (**Supplementary Figure 5**). The models for which variance was apparently over-estimated ($h^2_f > 0.99$) or undetermined ($h^2_f < 0.01$) were not considered for subsequent analysis. This was the case for ∼50% of the models for both traits, indicating that only selected biological groups are potentially helpful.

We initially analyzed the highest gain in prediction performance obtained by any GO term at any time point. For Φ_{PSII}, "salicylic acid biosynthesis" (BP) provided the highest increase in accuracy (∼60%), for Φ_{PSII} measurements under low light on the second day (**Figure 2**, **Supplementary Table 2A**). For the GO sub-ontologies CC and MF, "organelle outer membrane" and "phosphatase activity," respectively yielded highest gains in these categories under low light (∼43 and 37%, respectively; **Supplementary Table 2A**). None of the GO terms yielded a significant improvement after high light stress; however, some GO terms, e.g., "protein containing complex" yielded an increase in accuracy higher than the benchmark but not passing our model evaluation criteria wholly (**Figure 3**). For PLA, the largest improvement (∼197%) was obtained by the biological process "monocarboxylic acid biosynthesis" (**Figure 2**, **Supplementary Table 2B**). The best performing MF and CC terms for PLA were "exopeptidase activity" and "chloroplast part" (∼185 and ∼178%, respectively; **Figure 3**, **Supplementary Table 2B**). Interestingly, these best CC terms for both traits are directly related to photosynthesis, which lends credibility to the usefulness of the GO terms to capture relevant prior biological knowledge.

In total, 43 GO terms (BP:34, CC:6, MF:3) were potentially informative (i.e., Wilcoxon–Mann–Whitney test *p*-values < 0.05, without multiple testing correction), showing a tendency to improve Φ_{PSII} traits and yielding a significant increase in GFBLUP model accuracy (**Supplementary Figures 6A**, **7**, **Supplementary Table 2A**) compared to GBLUP. The overall gain in accuracy for these informative GO features ranged between 23 and 60%. The GO terms' hierarchical redundancy was removed using GO trimming (Jantzen et al., 2011) and the remaining 40 informative terms fell broadly into six biological clusters (**Figure 4**, **Supplementary Figure 9**): (i) hormonal regulation; (ii) cellular development; (iii) transport; (iv)

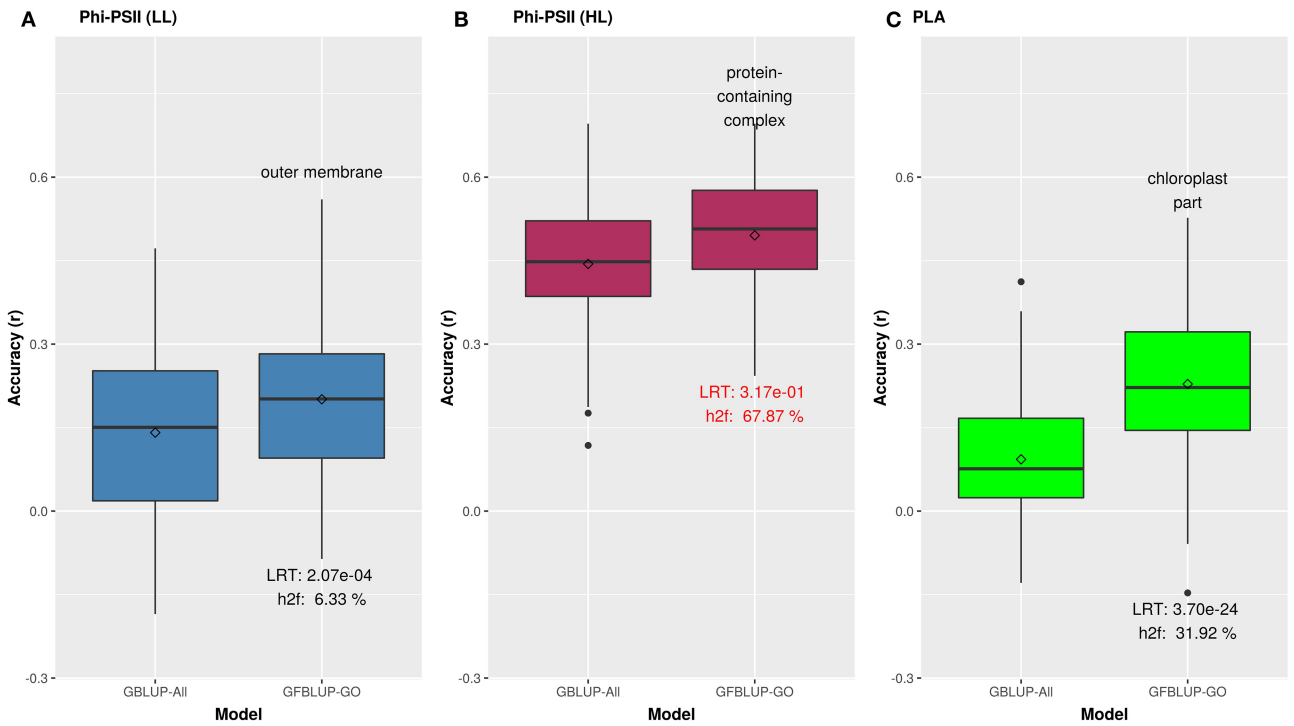

FIGURE 3 | Biological priors based on top GO cellular components improving genomic prediction accuracy for growth related traits. Prediction accuracy of the best GFBLUP models using Gene Ontology (GO) cellular components gene groups, compared to the GBLUP benchmark (without prior biological knowledge). The accuracy of benchmark model may differ within corresponding figures of **Figures 2**, **3**, because it is calculated from the corresponding time point, where improvement by GFBLUP was observed. Accuracy was calculated as Pearson correlation between true vs. predicted values. The GBLUP-ALL model uses all markers in GBLUP; GFBLUP-GO models use the top GO cellular component terms mentioned in the text above. The text in the bottom of boxplots shows the likelihood ratio test *p*-value (LRT) and proportion of genomic heritability explained (h_f^2) by corresponding GO model. **(A)** Accuracy for Φ_{PSII} under low light irradiance levels (LL). **(B)** Accuracy for Φ_{PSII} under high light irradiance (HL). Similar to **Figure 2**, the GFBLUP-GO model did not pass all of our model evaluation criteria (see Model Performance Evaluation), though showing some improvement. **(C)** Accuracy for PLA.

metabolism; (v) catabolism and (vi) macromolecular complex assembly, organization, and biogenesis. The cellular component terms were semantically clustered into organellar membranes and photosynthesis machinery sub-compartments, whereas molecular function terms were related to transmembrane transport and phosphatase activities.

For PLA, 52 GO terms (BP:41, CC:6, MF:5) resulted in significant improvement ($p_{\text{FDR}} < 0.05$) in predictive ability (**Figure 5**, **Supplementary Figure 6C**, **Supplementary Table 2B**) and the gain in accuracy ranged between 104 and 197%. After removal of hierarchical redundancy, semantic grouping of the remaining 45 GO terms showed that they involved a number of growth and developmental processes. Biological process GO terms fell into ∼8 clusters (**Figure 6**, **Supplementary Figure 10**) related to development, defense response, stress response, cell cycle regulation, metabolism, molecular biosynthesis, cellular component organization, and transport. The molecular function terms were clustered into two groups including exopeptidase and methyltransferase activities. The cellular component terms included the photosynthesis machinery (i.e., chloroplast) and endoplasmic reticulum. Comparison of average accuracy over multiple folds of GO models (**Supplementary Figures 6A,C**) indicate that many models performed better than GBLUP. Some of these passed our significance threshold (see model evaluation criteria, M&M) at a particular trait measurement but appeared to improve prediction performance for other measurement points as well.

The maximum number of genes annotated with the informative GO terms for Φ_{PSII} and significant GO terms for PLA were 1,358 and 1,245, respectively. These GO terms appeared at multiple levels of the GO hierarchical structures, including parent and child terms closely related to photosynthesis and growth (**Table 1**). Moreover, many genes were common with the pre-selected photosynthesis related *PSGENES*: 42 and 58% for Φ_{PSII} and PLA respectively, significantly more than what expected by chance ($p_{\chi^2_{df:1}} < 0.05$). Total genomic heritability (h^2_{GBLUP}) was negatively correlated with predictive gain ($r_{\Phi PSII} = -0.77$, $r_{\text{PLA}} = -0.5$). The genomic heritability explained individually (h_f^2) by the informative GO terms ranged between 6 and 31% for Φ_{PSII} and between 3 and 43% for PLA (**Supplementary Tables 2A,B**). Interestingly, the markers associated with these GO terms constituted only 0.1–3.3% of the total markers for Φ_{PSII} and 0.005–2.8% for PLA. This indicates that to improve predictive ability, genomic variance can be decomposed based on biologically meaningful sets of genes

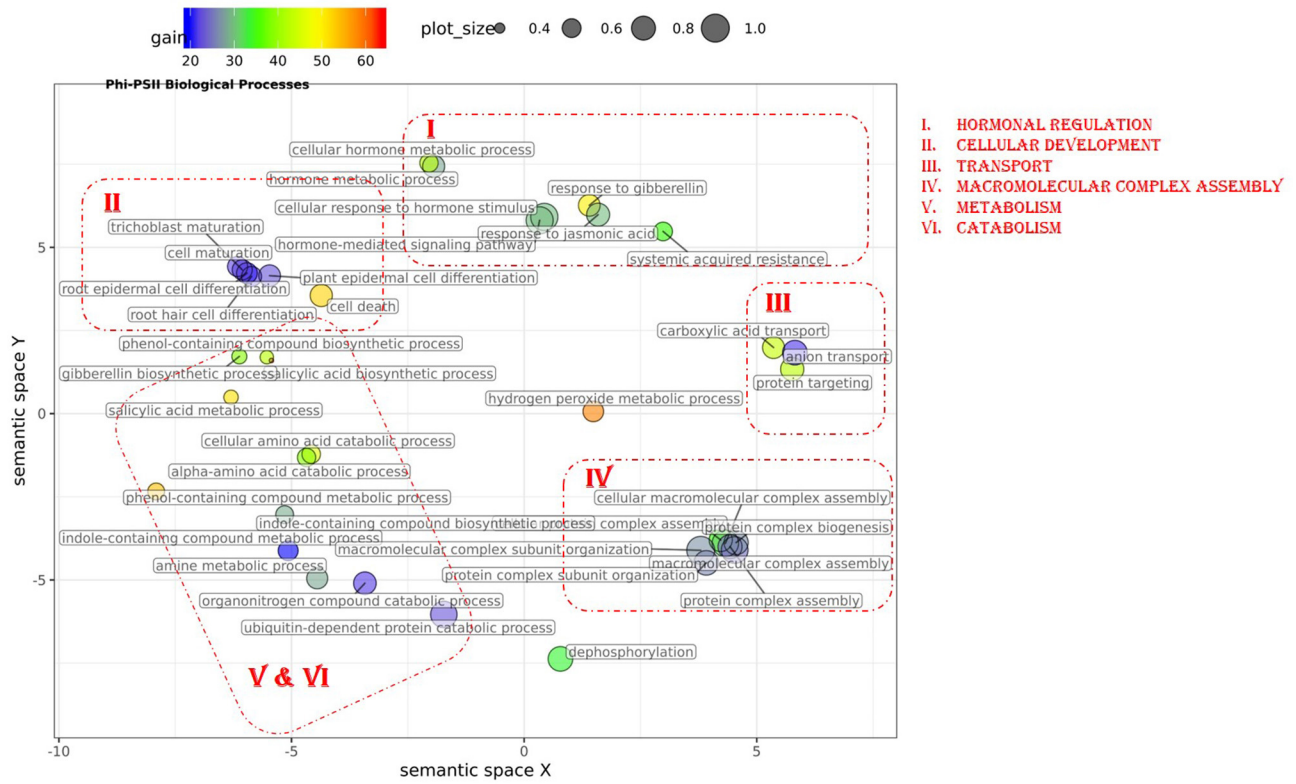

FIGURE 4 | Semantic clustering of GO informed prediction for Φ_{PSII}. Multidimensional scaling (MDS) plot of the representative subset (i.e., terms remaining after the redundancy reduction) of biological process GO terms informative for Φ_{PSII}. Semantically similar GO terms are clustered based on the *"SimRel"* semantic similarity measure using *Revigo*. Dot size is proportional to the number of genes annotated with the GO term, such that more general GO terms have larger circles. The *x* and *y* coordinates indicate relative cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the circle color.

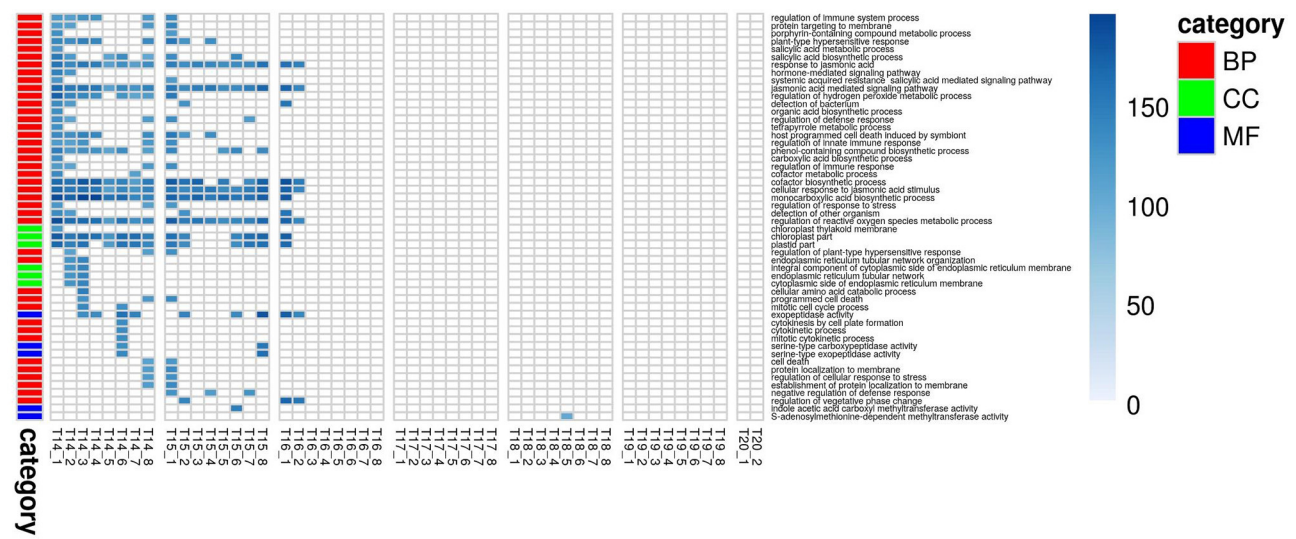

FIGURE 5 | Improvement in genomic predictive performance using GO for PLA. All GO terms that significantly improve GFBLUP models for PLA with %gain in accuracy (*r*) over GBLUP. Each GO term has a separate model for individual measurements indicated as T{day}_{Number of measurement}. The color bar identifies the GO terms as Biological Process (BP), Cellular Component (CC) and Molecular Function (MF).

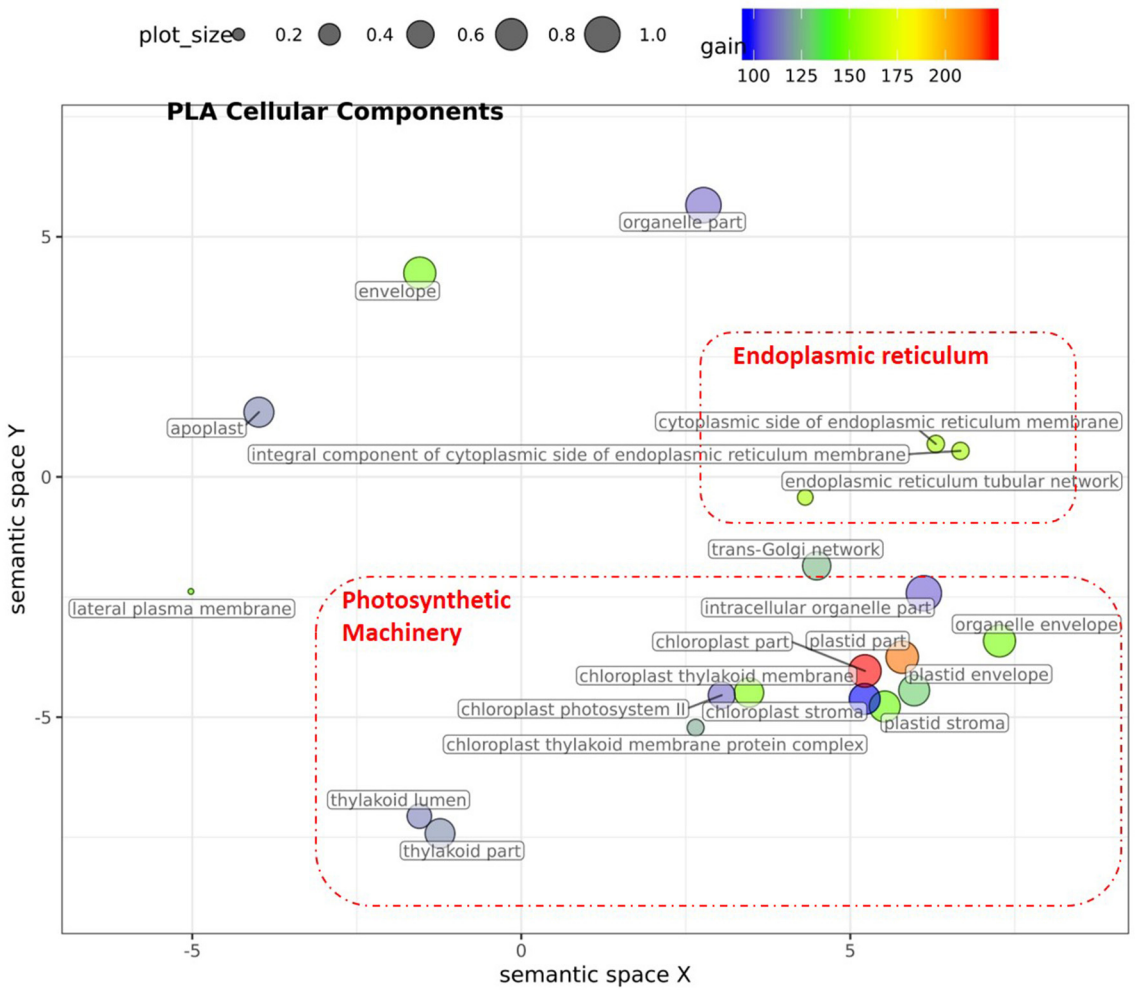

FIGURE 6 | Semantic clustering of GO informed prediction for PLA. Multidimensional scaling (MDS) plot of the representative subset (i.e., terms remaining after the redundancy reduction) of cellular component GO terms informative for PLA. Semantically similar GO terms are clustered based on the *"SimRel"* semantic similarity measure using *Revigo* (Supek et al., 2011). Dot size is proportional to the number of genes annotated with the GO term, such that more general GO terms have larger bubbles. The *x* and *y* coordinates indicate relative virtual cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the bubble color.

scattered over the genome rather than lie in adjacent regions such as in the MultiBLUP analysis above. Moreover, h_f^2 is positively correlated with GO gene group size ($r_{\Phi\mathrm{PSII}} = 0.87$, $r_{\mathrm{PLA}} = 0.77$) as well as with the likelihood ratio ($r_{\Phi\mathrm{PSII}} = 0.60$, $r_{\mathrm{PLA}} = 0.65$) of both trait models, indicating that incorporating meaningful prior subsets into the GFBLUP model improves goodness of fit.

From this we infer that GO-based prior knowledge can improve GP performance. The improvement is most prominent for traits with low heritability, where some of the GO terms appeared more frequently for PLA than Φ_{PSII} at multiple measurement times.

COEX Informed Prediction

Similar to genomic features based on GO, we made subsets of markers based on COEX clusters by selecting the markers within the ORFs of genes which were part of a given COEX cluster. Similar to GO based models, COEX models with zero and with 100% variance explained were discarded (**Supplementary Figure 5**). In general, more COEX models pass our model evaluation threshold (**Supplementary Figures 6B,D**) and they have a higher likelihood ratio than GO based models. This could be due to the genic overlap between groups and the enrichment of multiple related GO terms within a group.

For Φ_{PSII} we found 172 informative COEX gene groups potentially improving predictive ability, one of which was statistically significant ($p < 0.05$) after correcting for multiple testing using FDR (**Supplementary Figures 6B**, **8**). 355 COEX groups significantly improved predictive ability for PLA (**Figure 7**, **Supplementary Figure 6D**, **Supplementary Tables 3A,B**). The gain in accuracy was higher for PLA (80 to 243%) than for Φ_{PSII} (7 to 89%) and was negatively correlated with genomic heritability ($r_{\Phi\mathrm{PSII}}$

TABLE 1 | Known trait-specific GO terms improving genomic prediction performance for both traits.

GO ID	Ontology	Type	h^2_f	LR	*p*-value (unadj)	#gene	#marker	%gain	Cor(G_f,G_r)	h^2_{GBLUP}
Φ_{PSII}										
GO: 0009543	chloroplast thylakoid lumen	CC	0.07	10.53	1.48×10^{-2}	71	218	33	0.59	0.09
GO: 0031968	organelle outer membrane	CC	0.06	12.47	4.3×10^{-3}	72	345	40	0.61	0.08
GO:0044429	mitochondrial part	CC	0.14	47.05	2.3×10^{-3}	298	1069	38	0.81	0.09
GO:0005740	mitochondrial envelope	CC	0.13	8.43	2.7×10^{-2}	255	914	25	0.79	0.12
GO ID	**Ontology**	**Type**	h^2_f	**LR**	***p*-value (adj)**	**#gene**	**#marker**	**%gain**	**Cor(G_f,G_r)**	h^2_{GBLUP}
PLA										
GO:0044434	Chloroplast part	CC	0.32	101	5.26×10^{-5}	1211	5658	178	0.94	0.07
GO:0009535	chloroplast thylakoid membrane	CC	0.14	10	4.9×10^{-2}	322	1139	121	0.81	0.07
GO:0000911	cytokinesis by cell plate formation	BP	0.15	34	9.6×10^{-3}	204	1465	134	0.81	0.07
GO:0010090	trichome morphogenesis	BP	0.04	30	8.3×10^{-4}	31	65	154	0.40	0.06
GO:0010321	regulation of vegetative phase change	BP	0.14	18	4.9×10^{-3}	425	1512	106	0.84	0.07
GO:0048366	leaf development	BP	0.10	48	1.96×10^{-5}	99	487	187	0.62	0.06
GO:0090698	post-embryonic plant morphogenesis	BP	0.04	7	8.3×10^{-7}	4	11	207	0.20	0.06

The proportion of explained genomic heritability (h^2_f) by a GO term, likelihood ratio (LR) between GFBLUP and GBLUP models, Wilcoxon–Mann–Whitney test p-value, total number of genes and markers, %gain in accuracy (r), correlation between genomic relationship matrices based on GO term markers (G_f) and remaining markers (G_r) and total genomic heritability (h^2_{GBLUP}), for different trait specific GO terms that are common to both GO and COEX based analyses. For GO terms, the type is indicated—molecular function (MF), biological process (BP) and cellular component (CC).

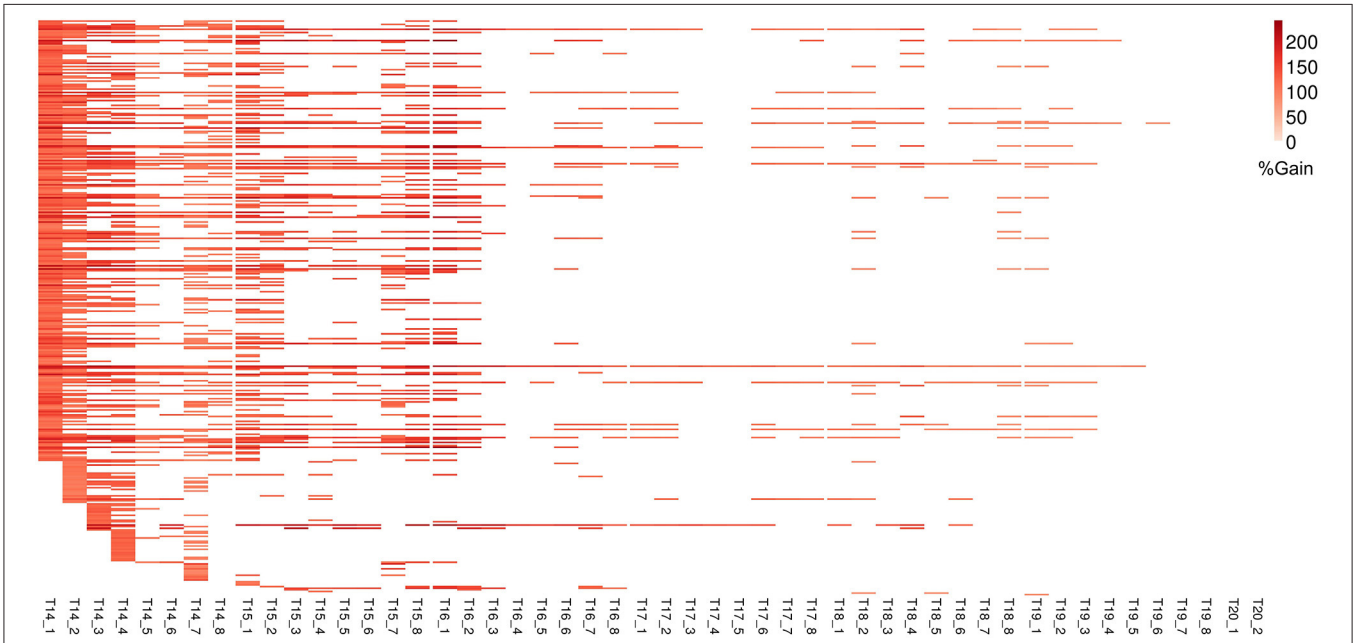

FIGURE 7 | Improvement in genomic prediction performance using co-expressed gene clusters for PLA. All COEX clusters that significantly improve GFBLUP models for PLA with %gain in accuracy (*r*) over GBLUP. Each COEX cluster has a separate model for individual measurements indicated as T{day}_{Number of measurement}. The clusters are ordered according to "cluster_sr_no" column in **Supplementary Table 3B**.

$= -0.86$, $r_{PLA} = -0.56$), like for GO informed prediction. This improvement was attributed to a maximum of only ∼5% of the total genomic markers in all groups. Interpretation of COEX gene groups is not as straightforward as of GO terms, which by nature carry an informative name. Interestingly, ∼90% of genes were common in the COEX groups for both traits, possibly due to the relatedness of the traits. To attach biological meaning to these groups we performed GO enrichment analysis on all groups together. We found 113 BP, 29 MF, and 24 CC most specific GO terms enriched in these clusters. The top 10 GO terms with highest fold enrichment include photosynthesis machinery, i.e., chloroplast stroma (GO:0009570), chloroplast

envelope (GO:0009941) cellular components; ATPase activity coupled with transmembrane ion transport (GO:0015662); and glucose metabolic process (**Supplementary Figure 11**, **Supplementary Table 5**). These results indicate that trait-specific co-expressed gene functional groups can also help improve prediction performance and that these groups capture biologically relevant functions.

Similar to GO informed prediction, ∼34% of COEX genes were common to the pre-selected photosynthesis related genes (*PSGENES*) for both traits, but here this is close to what we expect by chance. This indicates that, even though the COEX groups contain only a limited subset of all genes, they are not biased toward photosynthesis genes. The gain in predictive ability and explained genomic heritability (h_f^2) for Φ_{PSII} by the top COEX gene group was higher (89% resp. 14%) than those for the top GO feature (60% resp. 13%). Similarly, for PLA the top COEX gene group achieved a higher accuracy gain (242%) than the top GO group (197%), as shown in **Figure 2**. Notwithstanding these differences, we observed that many genes were common between GO and COEX based prediction for both traits (21 and 19% of all models passing the evaluation criteria for Φ_{PSII} and PLA resp.). These common genes in COEX based prediction were mainly enriched for many fundamental photosynthesis and growth related GO terms (**Supplementary Tables 7A,B**), e.g., light harvesting in photosystem I and photosynthetic electron transport in photosystem II (BP), chloroplast (CC), and ATP binding (MF).

The largest informative COEX groups for Φ_{PSII} and for PLA only differ slightly in sizes (3,176 and 2,840 genes, respectively), but on average, COEX groups were larger than the GO groups for both traits. The 95th percentile of genomic heritability explained individually by the COEX groups (h_f^2) was 70% for Φ_{PSII} and 39% for PLA, indicating that some Φ_{PSII} models could be over-estimated. Analogous to GO, h_f^2 was positively correlated with COEX gene group sizes ($r_{\Phi\mathrm{PSII}} = 0.88$, $r_{\mathrm{PLA}} = 0.40$) and likelihood ratio ($r_{\Phi\mathrm{PSII}} = 0.27$, $r_{\mathrm{PLA}} = 0.22$), indicating that incorporating meaningful prior subsets into the COEX model improved goodness of fit.

Together, our results illustrate that both of the meaningfully specific GO terms and more general COEX groups of genes with interrelated functions may improve GP predictive performance.

DISCUSSION

Predicting Photosynthesis

In this work, we aimed at improving GP performance by exploiting publicly available biological knowledge to group genes in three different ways: using our knowledge about the trait, using the Gene Ontology and using co-expression. Instead of developing new methodology, we focused on using existing BLUP methods, widely used in animal and plant breeding, to explore new sources of biological prior knowledge, e.g., clusters of co-expressed genes. The GFBLUP methodology was initially proposed for *Drosophila melanogaster* using Gene Ontology data as biological prior knowledge (Edwards et al., 2016). We also investigated to what extent different traits benefit from and the use of prior knowledge. Our results support a strong influence of different trait genetic architectures, since performance improvement was more evident for leaf area phenotypes than for Φ_{PSII}.

The approach can be generally applied to complex traits, but here we focused on photosynthesis and plant size. Besides serving as a case study, photosynthesis is also interesting in its own right, for two reasons. First, the genetic architecture of photosynthesis, though well-studied over the previous decades, is still poorly described in the quantitative genetic context (Van Rooijen et al., 2017). Secondly, it is an important target for improvement in crop breeding (Long et al., 2015). Modest improvements in photosynthesis efficiency by engineering photorespiratory pathways have demonstrated enormous yield gains (Kromdijk et al., 2016; South et al., 2019). The yield model of Monteith (Monteith, 1977) suggests that increased light use efficiency of photosystem II holds great potential to meet global food challenges by increasing the conversion efficiency of intercepted irradiance into biomass (ε_c) (Van Bezouw et al., 2019). Another determinant of plant growth rate is leaf area growth, involving precise regulation of photosynthesis machinery and growth hormones such as auxin (Zhang et al., 2017). Leaf area measurements from fluorescence based non-destructive optical phenotyping systems, can be efficiently used to screen plants at different growth stages with varying levels of photosynthetic rates (Weraduwage et al., 2015). Therefore, improved GP models for these traits could have impact in future crop breeding.

Following Edwards et al. (2016), we studied accuracy on internal test sets within the HapMap population. Further work is needed for data-driven selection of the most relevant terms for prediction on external test sets. For example, a possible strategy may be to select the feature with highest genomic variance explained, or with lowest p-value in the LRT we described. Our results indicate that biological priors driven GP models can be used to rank groups of genes potentially associated to the trait of interest along with improving prediction performance. The GWAS conducted on the same HapMap population for photosynthetic light use efficiency of photosystem II identified that the *A. thaliana* "Yellow Seedling 1" gene is involved in photosynthesis acclimation response (Van Rooijen et al., 2017). This *YS1* gene is annotated with GO Cellular Component terms chloroplast, intracellular membrane-bounded organelle and mitochondrion and GO Biological Process terms thylakoid membrane organization and photosystem II assembly. Our results using GO and COEX GP (**Table 1**) clearly demonstrate that these GO terms were most prevalent to improve the prediction and explain a large amount of genomic heritability. This indicates that genomic prediction and GWAS support each other as potentially useful tools for forward genetics.

The gain of predictive accuracy of the GP models compared to the base-model is trait-specific and negatively correlates with genomic heritability, which is promising for breeding at low h^2. This inverse relation may be due to the fact that we deal with highly polygenic, complex traits: many physiological and regulatory biological processes are involved in Φ_{PSII} under high light stress, e.g., PSII repair, ROX etc. Our models, testing groups

of genes individually, may not be able to improve performance for such cases. Another potential explanation lies in the ability of GFBLUP to capture small genetic variance at low h^2 in a separate random component, potentially including known causal genes, which is not possible in GBLUP.

Exploiting Biological Knowledge to Improve Genomic Prediction

With recent technological advances in both field and controlled environment high-throughput phenotyping systems, phenotypes can be measured at unprecedented scales. Phenotypes can vary in space and time due to genetics and environment alone, genotype-by-environment (GxE) interactions as well as stochastic and development effects. Component variances due to these factors can be calculated by precise modeling. If multiple measurements are available, GP models can be developed on individual measurements, treated as individual phenotypes, or on derived parameters, e.g., growth curves. We found that at each measurement timepoint, at least some GO (in particular cellular component terms) or COEX group could help to improve performance, and some were more frequent (**Figure 4**, **Supplementary Figure 7**). For example, for Φ_{PSII} no single GO or COEX gene group was capable of improving GP accuracy for all time points (either LL or HL separately), but a number of gene groups were able to improve PLA at multiple measurements (although not always meeting the threshold for significance). Phenotyping at an extended scale and GP modeling thus provides an opportunity to obtain biological insights. As an alternative to modeling at each timepoint separately, a whole time series or growth curve can be used instead. We did not pursue this here, as time series data is not generally available in most practical scenarios and we were interested to learn whether performance improvement was specific to growth stages and conditions e.g., models for Φ_{PSII} behaved differently under low and high light conditions.

Here, we mainly investigated two approaches to incorporate publicly available trait-specific biological information into GP, i.e., pre-selecting a list of genes and selecting sets or groups of genes based on predicted functional (i.e., GO) or expression (COEX) information. The approach using predicted functional information proved to be more useful in this context, but more approaches and sources of information can also be incorporated with a focus on prioritizing biologically related genomic regions. Moreover, knowledge from multiple heterogeneous sources can be combined to further pinpoint potential QTLs, termed as poly-omics GP models (Wheeler et al., 2014; Uzunangelov et al., 2020). These information sources may include (i) predicted variants effects, (ii) gene functions e.g., GO, COEX, (iii) networks of gene-gene and protein-protein interactions, stored in public resources like STRING (Mering et al., 2003), GeneMANIA (Warde-Farley et al., 2010); (iv) pathways, in which genes are grouped e.g., KEGG (Kanehisa and Goto, 2000); (v) previously generated GWAS and QTL results which indicate involvement of particular regions for specific traits e.g., AraGWAS (Togninalli et al., 2020), AraQTL (Nijveen et al., 2017), (vi) known connections to phenotypes and (vii) endophenotypes, usually measured using -omics data at different stages of genetic information flow toward phenotypes. The reliability of these sources of information is an important factor for credible analysis. Information describing the (un)certainty of annotations is generally available in the form of a score (e.g., for gene functions based on GO evidence scores or reliability scores generated by a prediction method). It remains an open question how to incorporate such scores in the process of using the biological knowledge for GP.

Our first approach, pre-selecting a gene list, seems to be naive but can be useful as a baseline for comparison with more complex statistical procedures. The group based approach is usually based on gene function, but this heavily depends on computational prediction, as for most of the genes in plants and animals, no experimental function annotation is available (Radivojac et al., 2013). Function prediction is often based on sequence similarity, which works well for predicting molecular functions but less so for biological processes. Using expression compendia based on multiple experiments poses an interesting alternative, since genes with similar expression patterns are more likely functionally related, hence more likely involved in the same biological process(es) (Kourmpetis et al., 2011). Alternatives are to define phenotype associated genomic regions based on differential gene expression levels (Fang et al., 2017) or metabolite levels and metabolic fluxes (Tong et al., 2020), or to construct haplotypes in genic regions based on their ontology information (Gao et al., 2018). The GP requiring genomics inferred relationship matrices (GRM), e.g., GBLUP and its variants, can make use of information derived from these sources to construct a population variance-covariance structure (Zhang et al., 2010, 2011; Fragomeni et al., 2017). A simple approach is to include multiple random effects for each knowledge source yielding its own variance-covariance structure for the population under study, in the mixed model equations (Guo et al., 2016). One way to combine multiple omics datasets is to prepare a Composite Relationship Matrix (CRM) as a linear combination of Genomic Relationship Matrices (GRMs), Expression Relationship Matrices (XRMs), Metabolome Relationship Matrices (MRMs), MicroRNA Relationship Matrices (miRMs) etc. (Wheeler et al., 2014).

Alternative Models for Genomic Prediction

Linear mixed model (LMM)-based genomic prediction, as used in this work, makes use of raw genotypes and parameter regularization to estimate thousands of SNP marker effects using only a few hundred observations ($p >> n$), employing different prior statistical assumptions on these parameters. This makes the approach fairly simple and interpretable; therefore, biological knowledge can be incorporated straightforwardly by employing these statistical assumptions. But with the increase in the ratio between markers and available phenotypes, serious overfitting problems may be encountered in these models (González-Recio et al., 2014), leading to a need to use prior knowledge in regularization. A more general set of statistical learning methods are Machine Learning (ML) methods for prediction and classification, capable of dealing with the dimensionality problem in a more flexible manner. In these methods, phenotypes

are regressed on nonlinear functions of genotypes rather than raw genotype values, compromising model interpretability but potentially improving prediction performance. Several studies have reported the use of Support Vector Machines (SVM), Reproducing Kernel Hilbert Spaces Regression (RKHS), Neural Networks (NN), Random Forests (RF), and boosting (De Los Campos et al., 2010; Ogutu et al., 2011) for genomic prediction. Still, low prediction accuracy remains a problem for complex traits. It will be interesting to further explore how biological knowledge can be incorporated into ML approaches for GP. One way could be to involve a knowledge driven regularization-based approach as demonstrated for disease prediction in human (Deng and Runger, 2013).

CONCLUSION

The wealth of publicly available transcriptomics and Gene Ontology based prior biological knowledge can be incorporated for genomic prediction of photosynthetic light use efficiency of photosystem II electron transport (Φ_{PSII}) and PLA. Significant improvement in prediction accuracy over the benchmark GBLUP model was obtained for several GO terms and COEX groups. This improvement is trait-specific and negatively correlates with genomic heritability; whereas, for projected leaf area we found more added value than for Φ_{PSII}. Many known photosynthesis-specific GO terms lead to improvements, providing evidence of the potential usefulness of this approach in future breeding practice. We foresee incorporation of heterogeneous prior biological information into machine learning algorithms as an active area of research in future.

MATERIALS AND METHODS

Datasets

Genotype Data

Genotype data of the 360 natural accessions in the core set of the *Arabidopsis thaliana* HapMap population, representing its global diversity, was obtained using Affymetrix 250k SNP array (Zhang and Borevitz, 2009; Baxter et al., 2010). The HapMap accessions were chosen as most accessions are more or less equally interrelated, so modeling is not heavily affected by population structure. Phenotypes of 344 accessions were available, so 16 accessions were removed from the analysis (CS76104, CS76112, CS76254, CS76257, CS76121, CS28051, CS28108, CS28808, CS28631, CS76086, CS76138, CS76212, CS76196, CS76110, CS76117, CS76118). Genotype data were subjected to quality control and all genotypes with a missing call in any accession were removed. Only 510 (0.24%) markers had minor allele frequency (MAF) <0.01 and 14,824 (6.9%) had MAF <0.05 (**Supplementary Figure 12**). To incorporate the effects of rare alleles along with common alleles in the GP model, the MAF filtering threshold was set at 0.01. Of subsequent markers in a window of 50bp with a Pearson correlation coefficient (r) < 0.999, one was removed, using PLINKv1.9 (Purcell et al., 2007). In total, 214,051 SNPs passed quality filtering, 213,541 remained after MAF filtering and 207,981 SNPs were available after LD pruning for the analyses. The resulting minimal distance between SNPs was found to be ∼550 bp.

Phenotype Data

The light use efficiency of Photosystem II electron transport (Φ_{PSII}) dataset was obtained from Van Rooijen et al. (2017), who measured it using chlorophyll fluorescence via NIR imaging at 790 nm. In this dataset, Φ_{PSII} was recorded three times a day; under 100 μmol m^{-2} s^{-1} (low light) for 2 days and for four continuous days after induction of high light stress at 550 μmol m^{-2} s^{-1} to study the photosynthetic acclimatory response. We measured PLA every 3 h starting from the afternoon of day 22 after sowing until early morning of day 29 using the *"Phenovator"* high-throughput automated phenotyping system (Flood et al., 2016), which results in total of 54 timepoints for this trait (**Supplementary Table 8**). Technical mis-match errors between the imaging system and the coordination of image analysis software were identified for some replicates at some time points for a small number of genotypes, but these were not found to influence overall results and the data was thus retained. Data of timepoints on day 22 was excluded from the analyses due to their relatively low coefficient of variation.

The *Phenovator* system has been designed to screen Arabidopsis plants for photosynthesis and growth on a larger temporal scale in a carefully controlled environment with minimal noise. The plants are grown over a table, spatially arranged into sowing blocks, imaged using a moveable monochrome camera recording 12 plants per image, and processed using an image processing software (available on demand from the authors). The system design allows spatial uniformity and temporal reproducibility by minimizing the design parameter variances. Therefore, we expected low variances of interactions between genotype and the design parameters; whereas, within image position and sowing position could have larger main effects and thus could be corrected for. Phenotypic values were taken as the average of one to four replicates of Best Linear Unbiased Estimators (BLUE) using the linear mixed model adjusted for experimental design factors (**Supplementary Table 9**) that were described in Flood et al. (2016). For this experiment, the important design factors are spatial row (x) and column (y) coordinate, the image position and the sowing block. Thus, the BLUE for phenotypic mean is calculated based on this equation, implemented in R with the *lmer* function (supplemental R script) using the lme4 package (Bates et al., 2007):

$$Y = Genotype + x + y + Image\_position + Sowing\_block + error \quad (1)$$

where *Genotype* is used as fixed effect and the other factors are defined as random effects.

Both traits, at all measurement times, showed approximately normal distributions (**Supplementary Figures 13**, **14**). The distributions are leptokurtic and left skewed for both traits (except for a few measurements for PLA on day 14 and day 15). The coefficients of variation under low light conditions for Φ_{PSII} ranged from 1.95 to 2.30% and 2.92 to 7.58% under high

light and 18.73 to 27.04% for PLA (**Supplementary Table 1**). Correlation between subsequent measurement times was high ($r > 0.9$) for both traits, except between measurements under low vs. high light conditions of Φ_{PSII}; therefore, these were analyzed separately.

Biological Priors

Co-expressed gene groups were obtained from the Arabidopsis expression compendium by Movahedi et al. (2011). GO data was retrieved using the R package "org.At.tair.db" (Carlson, 2019b) and genes were annotated using "GO.db" (Carlson, 2019a) irrespective of evidence codes. The set of genes in GO terms were up-propagated along the GO tree, such that each GO group in our analysis comprised of a set of all those genes attributed to itself or to all of its child terms. The up-propagated sets of genes were retrieved using the "GO2ALLTAIRS" method in the "org.At.tair.db" package. Markers in genes linked to a specific GO term or COEX cluster were used in the analyses.

Moreover, a set of 7,242 photosynthesis related genes was manually compiled (**Supplementary Table 6**) using four publicly available sources: KEGG (Kanehisa, 2001) pathways related to photosynthesis (i.e., ath00195, ath00197, ath00710); the Arabidopsis pathway database AraCyc for four photosynthesis pathways (i.e., Calvin cycle, photorespiration, oxygenic, light reaction); genes annotated with GO terms directly related to photosynthesis machinery; and all 51 priority genes selected for GWAS of photosynthesis acclamatory response identified by for this HapMap population.

Statistical Analysis

Linear Mixed Models

The Linear Mixed Model (LMM) with one random genomic component was used as baseline. This model (Equation 2), known as Genomic Best Linear Unbiased Prediction (GBLUP) (Habier et al., 2007; Vanraden, 2008) was used to predict marker effects, calculate genomic heritability (h^2_{GBLUP}) and the total additive genomic values, which is the sum of all marker effects:

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{g} + \boldsymbol{\varepsilon} \tag{2}$$

Here, $\tilde{\mathbf{y}}$ is an nx1 vector of adjusted phenotypes as described in section 5.1.2, $\boldsymbol{\mu}$ is the overall mean, $\mathbf{g}$ is an nx1 vector of genomic values captured by all genomic markers such that $\mathbf{g} = \hat{\mathbf{g}}$ and $\boldsymbol{\varepsilon}$ is an n-vector of residuals. The random genomic values $\mathbf{g}$ and residuals were assumed to be independent, normally distributed as $\mathbf{g} \sim N(0, \mathbf{G}\sigma^2_g)$, $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2_e)$. Here $\mathbf{G}$ is the genomic relationship matrix (GRM), providing variance-covariance structure of genotypes calculated from all genomic markers and $\mathbf{I}$ is the identity matrix.

Accordingly, for each GO and COEX gene groups, another linear mixed model similar to GBLUP but with two random genomic components (Equation 3), known as Genomic Feature Best Linear Unbiased Predictor (GFBLUP) (Edwards et al., 2016) was applied:

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{f} + \mathbf{r} + \boldsymbol{\varepsilon} \tag{3}$$

This model differs from GBLUP in that the total estimated genomic value ($\hat{\mathbf{g}} = \mathbf{f}+\hat{\mathbf{r}}$) is partitioned into genomic value captured by markers in a GO/COEX group ($\mathbf{f}$) and by the remaining markers ($\hat{\mathbf{r}}$), such that $\mathbf{f} \sim N(0, \mathbf{G_f}\sigma^2_f)$, $\mathbf{r} \sim N(0, \mathbf{G_r}\sigma^2_r)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2_e)$. For both GBLUP and GFBLUP, total genomic value $\hat{\mathbf{g}}$ of the test population was predicted conditional on observed phenotypes of the training population, using the approach mentioned by Edwards et al. (2016). The genomic relationship matrix $\mathbf{G}$ in the GBLUP model was constructed based on all genomic markers such that $\mathbf{G} = \frac{\mathbf{WW'}}{m}$, where $\mathbf{W}$ is an $n \times m$ genotype matrix (n genotypes and m markers), centered and scaled such that its i^{th} column $\mathbf{w_i} = \frac{(\mathbf{z_i} - 2p_i)}{\sqrt{2p_i(1-p_i)}}$, where $\mathbf{z_i}$ is the i^{th} column vector of $\mathbf{Z}$ having minor allele counts (0, 1, or 2) as entries and p_i is the MAF of the i^{th} marker. In our case, all genotypic locations were homozygous, so genotypes are coded as 0 or 2. For the GFBLUP model, the genomic relationship matrix $\mathbf{G_f}$ for each GO or COEX group was calculated from the markers linked to that group; $\mathbf{G_r}$ was constructed from the remaining markers.

The MultiBLUP model (Equation 4) was constructed according to the Adaptive MultiBLUP strategy proposed by (Speed and Balding, 2014). Briefly, the total genome was divided into adjacent but 50% overlapping regions of 10 kb. The genomic markers within these regions were tested as a group to estimate their association with the phenotype ($p < 10^{-5}$) and adjacent regions were merged if $p_{\text{Bonferroni}} < 0.05$. Subsequently, separate covariance matrices $\mathbf{K}_1$, $\mathbf{K}_2$,..., $\mathbf{K}_M$ were constructed for each region (M regions in total) based on its markers and genomic values $\mathbf{g}_1$, $\mathbf{g}_2$,..., $\mathbf{g}_M$ were estimated. The GRM based on all markers (equivalent to GBLUP) was used if no region was found significant. The total genomic value is $\hat{\mathbf{g}} = \sum_{m=1}^{M} \hat{\mathbf{g}}_m$ with i.i.d. $\mathbf{g}_m \sim N(0, \mathbf{K}_m\sigma^2_m)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2_e)$:

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \sum_{\boldsymbol{m}=1}^{\boldsymbol{M}} \mathbf{g}_{\boldsymbol{m}} + \boldsymbol{\varepsilon} \tag{4}$$

Variance components in all of these LMMs were estimated using the average information restricted maximum-likelihood (REML) procedure (Johnson and Thompson, 1995) implemented in the *greml* method of the R package *qgg* (Rohde et al., 2020) for GBLUP/GFBLUP, using a maximum of 100 iterations at a tolerance level of 10^{-5}; and LDAK v5.1 (http://dougspeed.com/) for MultiBLUP.

Total additive genomic value was predicted using 8-fold cross-validation. This involved training the model using 301 (78%) genotypes and using the remaining 43 for testing in each fold. The exact same accessions were used for both GBLUP and GFBLUP during each split to enable a fair comparison. Prediction accuracy of models was defined as Pearson correlation (r) between observed phenotypic values and predicted genomic values of the test population in each fold. The procedure was repeated 10 times, thus modeled predictive ability distributions consisted of 80 correlations or fewer if variances were over- or underestimated as described earlier by simulation studies (Kruijer et al., 2015). For comparison between models, the median of these correlations was used, and significance of the difference was tested using the non-parametric Wilcoxon–Mann–Whitney test

for assessing significant differences in median accuracy between GBLUP and GFBLUP. Subsequently, *p*-values were adjusted for multiple-testing correction by calculating False Discovery Rate (FDR) based on total number of GO/COEX groups multiplied by total number of time points (Edwards et al., 2016). For Φ_{PSII} we also analyzed results without FDR adjustment, which are referred as "informative" as opposed to "significant" throughout the text.

Model Performance Evaluation

GFBLUP models were compared to the benchmark GBLUP based on their goodness of fit, predictive ability and estimated genomic parameters. Using the likelihood ratio test (LRT) we tested the null-hypothesis $\sigma_f^2 = 0$. LRT *p*-values were based on the asymptotic distribution of the LRT-statistic, which is a mixture of a point mass at 0 and a χ^2-distribution with 1 degree of freedom (d.o.f.) (Edwards et al., 2015). The significantly improved GFBLUP models ($p_{\text{LRT}} < 0.05$) having predictive abilities greater than the benchmark GBLUP (i.e., *p-value* of Wilcoxon-Mann-Whitney tests < 0.05) were filtered for subsequent analysis. Genomic parameters were calculated from variance estimates of both models to analyze only models passing the abovementioned filtering criteria. This includes total genomic heritability explained ($h^2_{GBLUP} = \frac{\sigma_g^2}{(\sigma_g^2+\sigma_\varepsilon^2)}$) and proportion of genomic heritability explained by an individual GO/COEX group in GFBLUP models ($h^2_f = \frac{\sigma_f^2}{(\sigma_f^2+\sigma_r^2+\sigma_\varepsilon^2)}$). In order to check if we obtained a higher number of *PSGENES* in GO/COEX groups than expected by chance, we used the chi-square test with 1 d.o.f. to compare the observed vs. expected frequencies of *PSGENES* in these groups.

Semantic Clustering of GO Terms

Informative GO terms were clustered based on their semantic similarity using the *Revigo* (Supek et al., 2011) web server with *"SimRel"* semantic similarity metric equal to 0.7. The resulting GO clusters were plotted using a Multidimensional Scaling (MDS) plot in R, where maximum %gain in accuracy by each GO term was used to color the bubbles. GO terms enriched in COEX groups were found using the PANTHER classification system (Mi et al., 2019). Fisher's exact test was used for calculating enrichment *p*-values followed by multiple testing correction using the FDR, reporting enrichment at $p < 0.05$. These enriched GO terms were sorted in order of their GO hierarchical tree such that a child term was below its parent; thus, the most specific GO terms are the child GO terms in the bottom of that tree, were used for subsequent analysis.

DATA AVAILABILITY STATEMENT

All data and scripts have been uploaded to the Wageningen University & Research git server (https://git.wur.nl/faroo002/pub1).

AUTHOR CONTRIBUTIONS

MA and T-PN provided the genotype and phenotype datasets. MF performed the analyses. DR, AD, and HN were involved in designing the analyses and interpreting the results. WK helped with statistical analysis. MF wrote the manuscript with DR, AD, HN, and SM. All authors read the final manuscript.

FUNDING

MF was supported by the sandwich Ph.D. programme of Wageningen University and Research (WUR). The authors are grateful for the support of both WUR and NIBGE to conduct this study.

ACKNOWLEDGMENTS

We are thankful to Pádraic J Flood of Plant Breeding, Wageningen University and Research for reviewing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.609117/full#supplementary-material

Supplementary Figure 1 | Relation between genomic heritability and GBLUP predictive ability. GBLUP prediction accuracy is directly proportional to genomic heritability for both traits. **(A)** shows the relation between heritability and accuracy under low light (LL) and high light (HL) irradiance levels for Φ_{PSII}, **(B)** shows the same for PLA.

Supplementary Figure 2 | GBLUP accuracy (*r*) vs. genomic variance (h^2_{GBLUP}). Each dot corresponds to prediction accuracy (*r*) of GBLUP (*y*-axis) for each split of the data during cross-validation. The genomic variance explained by the model (*x*-axis) ranges from 0 to 1 and calculated as $h^2_{GBLUP} = \frac{\sigma_g^2}{(\sigma_g^2+\sigma_\varepsilon^2)}$. Models at different measurement times are colored differently. **(A)** represents GBLUP models for Φ_{PSII} and contains two separate clouds of dots, representing LL (left) and HL (right) models with different heritability ranges. **(B)** represents GBLUP models for PLA.

Supplementary Figure 3 | MultiBLUP predictive ability. The boxplots show the prediction accuracy (*r*) of MultiBLUP applied to 18 measurements of Φ_{PSII} and 50 measurements of PLA. The average accuracy is slightly lower than the average GBLUP accuracy (white star) for both traits. **(A)** shows the prediction accuracy under low light (LL) and high light (HL) irradiance levels for Φ_{PSII} whereas, **(B)** shows the same for PLA.

Supplementary Figure 4 | Number of genes and markers in GO and COEX features. Total number of genes and markers associated with those genes for both types of genomic features, i.e., GO (left) and COEX (right).

Supplementary Figure 5 | GFBLUP accuracy (*r*) vs. genomic variance (h^2_f) explained by a GO/COEX group. Each dot corresponds to prediction accuracy (*r*) of GFBLUP (y-axis) for each split of data during cross-validation for a particular GO **(A,C)** and COEX **(B,D)** group. The genomic variance explained by the particular GO/COEX (*x*-axis) ranges from 0 to 1. **(A,B)**: GFBLUP models for Φ_{PSII}; **(C,D)**: GFBLUP models for PLA.

Supplementary Figure 6 | GBLUP vs. GFBLUP predictive ability. Average prediction accuracy (*r*) of GBLUP vs. GFBLUP using GO terms **(A,C)** and COEX clusters **(B,D)** for Φ_{PSII} **(A,B)** and PLA **(C,D)**. The average was calculated over 80 splits of the data (8-fold cross-validation repeated 10 times), excluding models

where variance was undetermined). Red dots indicate models that passed our model evaluation criteria (see M&M).

Supplementary Figure 7 | Improvement in genomic prediction performance using informative GO terms for $\varphi_{PSII.}$ All informative GO terms with %gain in accuracy (*r*) of GFBLUP over GBLUP at multiple Φ_{PSII} measurement times, indicated by {Low|High light}{day}_{Number of measurement}. The color bar identifies GO terms as Biological Process (BP), Cellular Component (CC) or Molecular Function (MF).

Supplementary Figure 8 | Improvement in genomic prediction performance using informative COEX groups for φ_{PSII}. All informative COEX clusters with %gain in accuracy (*r*) of GFBLUP over GBLUP at multiple Φ_{PSII} measurement times, indicated by {Low|High light}{day}_{Number of measurement}.

Supplementary Figure 9 | Semantic clustering of GO informed prediction for Φ_{PSII}. Multidimensional scaling (MDS) plot of representative subset (i.e., terms remaining after the redundancy reduction) of informative GO terms molecular functions and cellular components, capable of improving predictive ability of GFBLUP models for Φ_{PSII}. Semantically similar GO terms are clustered based on the "SimRel" semantic similarity measure using *Revigo*. Dot size is proportional to the number of genes annotated with a GO term in the TAIR9 reference genome annotation. The *x* and *y* coordinates indicate relative cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the bubble color.

Supplementary Figure 10 | Semantic clustering of GO informed prediction for PLA. Multidimensional scaling (MDS) plot of representative subset (i.e., terms remaining after the redundancy reduction) of informative GO terms molecular functions and cellular components, capable of improving predictive ability of GFBLUP models for PLA. Semantically similar GO terms are clustered based on the "*SimRel*" semantic similarity measure using *Revigo*. Dot size is proportional to the number of genes annotated with a GO term in the TAIR9 reference genome annotation. The *x* and *y* coordinates indicate relative cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the bubble color.

Supplementary Figure 11 | Top 10 enriched GO terms in COEX clusters for Φ_{PSII} and PLA. Top 10 most specific GO terms enriched in 172 informative COEX clusters for the Φ_{PSII} and 355 for PLA traits. The horizontal axis measures the fold enrichment, i.e., the observed fraction of genes annotated with a particular GO term divided by the expected fraction in the reference genome of *Arabidopsis thaliana*. Enrichment *p*-values were found using Fisher's exact test with multiple testing correction using False Discovery Rate (FDR); only terms with $p_{FDR} < 0.05$ are shown.

Supplementary Figure 12 | Minor allele frequency spectrum (MAF). MAF distribution of all 214,051 chip markers. The orange bar represents all markers having MAF<5%, the red bar rare alleles with MAF<1%.

Supplementary Figure 13 | φ_{PSII} phenotypic data distributions using Best Linear Unbiased Estimates (BLUE). Distributions of genotypic means of BLUE values of genotypes in the dataset.

Supplementary Figure 14 | PLA phenotypic data distributions using Best Linear Unbiased Estimates (BLUE). Distributions of genotypic means of BLUE values of genotypes in the dataset.

Supplementary Table 1 | Best Linear Unbiased Estimated Phenotypic data statistics.

Supplementary Table 2a | Informative GO terms increasing GFBLUP prediction accuracy for Φ_{PSII}.

Supplementary Table 2b | GO terms significantly increasing GFBLUP prediction accuracy for PLA.

Supplementary Table 3a | Informative COEX improving GFBLUP prediction accuracy for Φ_{PSII}.

Supplementary Table 3b | COEX significantly improving GFBLUP prediction accuracy for PLA.

Supplementary Table 4 | Genomic features statistics.

Supplementary Table 5 | Enriched Go terms in Φ_{PSII} and PLA COEX analysis.

Supplementary Table 6 | List of genes used in GBLUP based on only photosynthesis genes markers.

Supplementary Table 7a | GO Enrichment of common genes between GO and COEX based analysis for Φ_{PSII}.

Supplementary Table 7b | GO Enrichment of common genes between GO and COEX based analysis for PLA.

Supplementary Table 8 | Raw measurements of Projected Leaf Area.

Supplementary Table 9 | Average best linear unbiased estimates (BLUE) of Projected Leaf Area.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* 25, 25–29. doi: 10.1038/75556

Azodi, C. B., Pardo, J., Vanburen, R., De Los Campos, G., and Shiu, S.-H. (2020). Transcriptome-based prediction of complex traits in maize. *Plant Cell* 32, 139–151. doi: 10.1105/tpc.19.00332

Bates, D., Sarkar, D., Bates, M. D., and Matrix, L. (2007). *The lme4 Package*. R package version 2, 74.

Baxter, I., Brazelton, J. N., Yu, D., Huang, Y. S., Lahner, B., Yakubova, E., et al. (2010). A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1; 1. *PLoS Genetics* 6:e1001193. doi: 10.1371/journal.pgen.1001193

Carlson, M. (2019a). *GO.db: A Set of Annotation Maps Describing the Entire Gene Ontology*. R package version 3.10.10.

Carlson, M. (2019b). *org.At.tair.db: Genome Wide Annotation for Arabidopsis*. R package version 3.10.10.

Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., and Poland, J. (2018). Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. *Plant Genome* 11:43. doi: 10.3835/plantgenome2017.05.0043

De Los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* 92, 295–308. doi: 10.1017/S0016672310000285

Deng, H., and Runger, G. (2013). Gene selection with guided regularized random forest. *Pattern Recogn.* 46, 3483–3489. doi: 10.1016/j.patcog.2013.05.018

Edwards, S. M., Sorensen, I. F., Sarup, P., Mackay, T. F., and Sorensen, P. (2016). Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. *Genetics* 203, 1871–1883. doi: 10.1534/genetics.116.187161

Edwards, S. M., Thomsen, B., Madsen, P., and Sørensen, P. (2015). Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. *Genet. Select. Evol.* 47:60. doi: 10.1186/s12711-015-0132-6

Ehsani, A., Janss, L., Pomp, D., and Sørensen, P. (2016). Decomposing genomic variance using information from GWA, GWE and eQTL analysis. *Anim. Genet.* 47, 165–173. doi: 10.1111/age.12396

Falconer, D., and Mackay, T. (1996). *Introduction to Quantitative Genetics*. Harlow: Longmans Green 3.

Fang, L., Sahana, G., Ma, P., Su, G., Yu, Y., Zhang, S., et al. (2017). Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet. Select. Evol.* 49:44. doi: 10.1186/s12711-017-0319-0

Flood, P. J., Kruijer, W., Schnabel, S. K., Van Der Schoor, R., Jalink, H., Snel, J. F. H., et al. (2016). Phenomics for photosynthesis, growth and reflectance in *Arabidopsis thaliana* reveals circadian and long-term fluctuations in heritability. *Plant Methods* 12:14. doi: 10.1186/s13007-016-0113-y

Fragomeni, B. O., Lourenco, D. A. L., Masuda, Y., Legarra, A., and Misztal, I. (2017). Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genet. Select. Evol.* 49:59. doi: 10.1186/s12711-017-0341-2

Gao, N., Teng, J., Ye, S., Yuan, X., Huang, S., Zhang, H., et al. (2018). Genomic prediction of complex phenotypes using genic similarity based relatedness matrix. *Front. Genet.* 9:364. doi: 10.3389/fgene.2018.00364

Gebreyesus, G., Bovenhuis, H., Lund, M. S., Poulsen, N. A., Sun, D., and Buitenhuis, B. (2019). Reliability of genomic prediction for milk fatty acid composition by using a multi-population reference and incorporating GWAS results. *Genet. Select. Evol.* 51:16. doi: 10.1186/s12711-019-0460-z

Gianola, D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753

González-Recio, O., Rosa, G. J., and Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Sci.* 166, 217–231. doi: 10.1016/j.livsci.2014.05.036

Guo, Z., Magwire, M. M., Basten, C. J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theor. Appl. Genet.* 129, 2413–2427. doi: 10.1007/s00122-016-2780-5

Habier, D., Fernando, R. L., and Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646

Jantzen, S. G., Sutherland, B. J. G., Minkley, D. R., and Koop, B. F. (2011). GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets. *BMC Res. Notes* 4:267. doi: 10.1186/1756-0500-4-267

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768. doi: 10.1534/genetics.115.177907

Johnson, D., and Thompson, R. (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.* 78, 449–456. doi: 10.3168/jds.S0022-0302(95)76654-1

Kanehisa, M. (2001). Prediction of higher order functional networks from genomic data. *Pharmacogenomics* 2, 373–385. doi: 10.1517/14622416.2.4.373

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Karaman, E., Cheng, H., Firat, M. Z., Garrick, D. J., and Fernando, R. L. (2016). An upper bound for accuracy of prediction using GBLUP. *PLoS ONE* 11:e161054. doi: 10.1371/journal.pone.0161054

Kourmpetis, Y. A., Van Dijk, A. D., Van Ham, R. C., and Ter Braak, C. J. (2011). Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources. *Plant Physiol.* 155, 271–281. doi: 10.1104/pp.110.162164

Kromdijk, J., Głowacka, K., Leonelli, L., Gabilly, S. T., Iwai, M., Niyogi, K. K., et al. (2016). Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. *Science* 354, 857–861. doi: 10.1126/science.aai8878

Kruijer, W., Boer, M. P., Malosetti, M., Flood, P. J., Engel, B., Kooke, R., et al. (2015). Marker-based estimation of heritability in immortal populations. *Genetics* 199, 379–398. doi: 10.1534/genetics.114.167916

Legarra, A., and Ducrocq, V. (2012). Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. *J. Dairy Sci.* 95, 4629–4645. doi: 10.3168/jds.2011-4982

Li, B., Zhang, N., Wang, Y. G., George, A. W., Reverter, A., and Li, Y. (2018). Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* 9:237. doi: 10.3389/fgene.2018.00237

Liu, P.-C., Peacock, W. J., Wang, L., Furbank, R., Larkum, A., and Dennis, E. S. (2020). Leaf growth in early development is key to biomass heterosis in Arabidopsis. *J. Exp. Botany* 71, 2439–2450. doi: 10.1093/jxb/eraa006

Long, S. P., Marshall-Colon, A., and Zhu, X.-G. (2015). Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. *Cell* 161, 56–66. doi: 10.1016/j.cell.2015.03.019

Lopez-Cruz, M., Olson, E., Rovere, G., Crossa, J., Dreisigacker, S., Mondal, S., et al. (2020). Regularized selection indices for breeding value prediction using hyper-spectral image data. *Sci. Rep.* 10:8195. doi: 10.1038/s41598-020-65011-2

Macleod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genom.* 17:144. doi: 10.1186/s12864-016-2443-6

Mering, C. V., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31, 258–261. doi: 10.1093/nar/gkg034

Meuwissen, T. H. E., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. Available online at: https://www.genetics.org/content/157/4/1819.long

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids Res.* 47, D419–D426. doi: 10.1093/nar/gky1038

Monteith, J. L. (1977). Climate and the efficiency of crop production in Britain. *Phil. Trans. R. Soc. London. Biol. Sci.* 281, 277–294. doi: 10.1098/rstb.1977.0140

Morgante, F. (2018). *Genetic Analysis and Prediction of Complex Traits in Drosophila melanogaster*. Ph.D. Thesis, North Carolina State University.

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* 11:e1004969. doi: 10.1371/journal.pgen.1004969

Movahedi, S., Van De Peer, Y., and Vandepoele, K. (2011). Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. *Plant Physiol.* 156, 1316–1330. doi: 10.1104/pp.111.177865

Nijveen, H., Ligterink, W., Keurentjes, J. J., Loudet, O., Long, J., Sterken, M. G., et al. (2017). Ara QTL-workbench and archive for systems genetics in *Arabidopsis thaliana*. *Plant J.* 89, 1225–1235. doi: 10.1111/tpj.13457

Ogutu, J. O., Piepho, H. P., and Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceed.* 5:S11. doi: 10.1186/1753-6561-5-S3-S11

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., et al. (2013). A large-scale evaluation of computational protein function prediction. *Nat. Methods* 10, 221–227. doi: 10.1038/nmeth.2340

Rimbert, H., Darrier, B., Navarro, J., Kitt, J., Choulet, F., Leveugle, M., et al. (2018). High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS ONE* 13:e0186329. doi: 10.1371/journal.pone.0186329

Rohde, P. D., Demontis, D., Børglum, A., and Sørensen, P. (2017). "Improved prediction of genetic predisposition to psychiatric disorders using genomic feature best linear unbiased prediction models," in *50th European Society of Human Genetics Conference: Posters* (Copenhagen).

Rohde, P. D., Fourie Sørensen, I., and Sørensen, P. (2020). qgg: an R package for large-scale quantitative genetic analyses. *Bioinformatics* 36, 2614–2615. doi: 10.1093/bioinformatics/btz955

Sarup, P., Jensen, J., Ostersen, T., Henryon, M., and Sorensen, P. (2016). Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genet.* 17:11. doi: 10.1186/s12863-015-0322-9

South, P. F., Cavanagh, A. P., Liu, H. W., and Ort, D. R. (2019). Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. *Science* 363:77. doi: 10.1126/science.aat9077

Speed, D., and Balding, D. J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* 24, 1550–1557. doi: 10.1101/gr.169375.113

Supek, F., and Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* 6:e21800. doi: 10.1371/journal.pone.0021800

Togninalli, M., Seren, Ü., Freudenthal, J. A., Monroe, J. G., Meng, D., Nordborg, M., et al. (2020). AraPheno and the AraGWAS Catalog 2020: a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana. *Nucleic Acids Res.* 48, D1063–D1068. doi: 10.1093/nar/gkz925

Tong, H., Küken, A., and Nikoloski, Z. (2020). Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth. *Nature Commun.* 11, 1–9. doi: 10.1038/s41467-020-16279-5

Uzunangelov, V., Wong, C. K., and Stuart, J. (2020). Highly accurate cancer phenotype prediction with AKLIMATE, a stacked kernel learner integrating multimodal genomic data and pathway knowledge. *bioRxiv [Preprint]*. doi: 10.1101/2020.07.15.205575

Van Bezouw, R. F. H. M., Keurentjes, J. J. B., Harbinson, J., and Aarts, M. G. M. (2019). Converging phenomics and genomics to study natural variation in plant photosynthetic efficiency. *Plant J.* 97, 112–133. doi: 10.1111/tpj.14190

Van Rooijen, R., Aarts, M. G. M., and Harbinson, J. (2015). Natural genetic variation for acclimation of photosynthetic light use efficiency to growth irradiance in Arabidopsis. *Plant Physiol.* 167, 1412–1429. doi: 10.1104/pp.114.252239

Van Rooijen, R., Kruijer, W., Boesten, R., Van Eeuwijk, F. A., Harbinson, J., and Aarts, M. G. M. (2017). Natural variation of YELLOW SEEDLING1 affects photosynthetic acclimation of *Arabidopsis thaliana*. *Nat. Commun.* 8:1421. doi: 10.1038/s41467-017-01576-3

Vanraden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Vanraden, P. M., Tooker, M. E., O'connell, J. R., Cole, J. B., and Bickhart, D. M. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Select. Evol.* 49:32. doi: 10.1186/s12711-017-0307-4

Wang, J., Zhou, Z., Zhang, Z., Li, H., Liu, D., Zhang, Q., et al. (2018). Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity*. 121, 648–662. doi: 10.1038/s41437-018-0075-0

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220. doi: 10.1093/nar/gkq537

Weraduwage, S. M., Chen, J., Anozie, F. C., Morales, A., Weise, S. E., and Sharkey, T. D. (2015). The relationship between leaf area growth and biomass accumulation in *Arabidopsis thaliana*. *Front. Plant Sci.* 6:167. doi: 10.3389/fpls.2015.00167

Wheeler, H. E., Aquino-Michaels, K., Gamazon, E. R., Trubetskoy, V. V., Dolan, M. E., Huang, R. S., et al. (2014). Poly-omic prediction of complex traits: OmicKriging. *Genetic Epidemiol.* 38, 402–415. doi: 10.1002/gepi.21808

Zhang, M., Hu, X. L., Zhu, M., Xu, M. Y., and Wang, L. (2017). Transcription factors NF-YA2 and NF-YA10 regulate leaf growth via auxin signaling in Arabidopsis. *Sci. Rep.* 7:1475. doi: 10.1038/s41598-017-01475-z

Zhang, X., and Borevitz, J. O. (2009). Global analysis of allele-specific expression in *Arabidopsis thaliana*. *Genetics* 182, 943–954. doi: 10.1534/genetics.109.103499

Zhang, Z., Ding, X., Liu, J., De Koning, D. J., and Zhang, Q. (2011). Genomic selection for QTL-MAS data using a trait-specific relationship matrix. *BMC Proceed.* 5:S15. doi: 10.1186/1753-6561-5-S3-S15

Zhang, Z., Liu, J., Ding, X., Bijma, P., De Koning, D. J., and Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE* 5:e12648. doi: 10.1371/journal.pone.0012648

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 January 2021
doi: 10.3389/fgene.2020.596258

Independent Validation of Genomic Prediction in Strawberry Over Multiple Cycles

Luis F. Osorio[1], Salvador A. Gezan[2†], Sujeet Verma[1] and Vance M. Whitaker[1]*

[1] Gulf Coast Research and Education Center, University of Florida, Wimauma, FL, United States, [2] School of Forest Resources and Conservation, University of Florida, Gainesville, FL, United States

The University of Florida strawberry (*Fragaria* × *ananassa*) breeding program has implemented genomic prediction (GP) as a tool for choosing outstanding parents for crosses over the last five seasons. This has allowed the use of some parents 1 year earlier than with traditional methods, thus reducing the duration of the breeding cycle. However, as the number of breeding cycles increases over time, greater knowledge is needed on how multiple cycles can be used in the practical implementation of GP in strawberry breeding. Advanced selections and cultivars totaling 1,558 unique individuals were tested in field trials for yield and fruit quality traits over five consecutive years and genotyped for 9,908 SNP markers. Prediction of breeding values was carried out using Bayes B models. Independent validation was carried out using separate trials/years as training (TRN) and testing (TST) populations. Single-trial predictive abilities for five polygenic traits averaged 0.35, which was reduced to 0.24 when individuals common across trials were excluded, emphasizing the importance of relatedness among training and testing populations. Training populations including up to four previous breeding cycles increased predictive abilities, likely due to increases in both training population size and relatedness. Predictive ability was also strongly influenced by heritability, but less so by changes in linkage disequilibrium and effective population size. Genotype by year interactions were minimal. A strategy for practical implementation of GP in strawberry breeding is outlined that uses multiple cycles to predict parental performance and accounts for traits not included in GP models when constructing crosses. Given the importance of relatedness to the success of GP in strawberry, future work could focus on the optimization of relatedness in the design of TRN and TST populations to increase predictive ability in the short-term without compromising long-term genetic gains.

Keywords: training population, *Fragaria*, breeding, Bayes B, genome-wide prediction, test population

OPEN ACCESS

Edited by:
Diego Jarquin, University of Nebraska–Lincoln, United States

Reviewed by:
Just Jensen, Aarhus University, Denmark
Alencar Xavier, Corteva Agriscience™, United States

****Correspondence:***
Vance M. Whitaker
vwhitaker@ufl.edu

†Present address:
Salvador A. Gezan, VSN International Ltd., Hemel Hempstead, United Kingdom

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 August 2020*
Accepted: *31 December 2020*
Published: *22 January 2021*

Citation:
Osorio LF, Gezan SA, Verma S and Whitaker VM (2021) Independent Validation of Genomic Prediction in Strawberry Over Multiple Cycles. Front. Genet. 11:596258. doi: 10.3389/fgene.2020.596258

INTRODUCTION

The development of high throughput genotyping and new methods for analyzing genome-wide molecular data are revolutionizing crop improvement. In particular, genomic prediction (GP) is helping to increase genetic gains for genetically complex traits in animal (Hayes et al., 2009), crop (Bernardo and Yu, 2007; Crossa et al., 2010; Gezan et al., 2017), and tree breeding programs (Kumar et al., 2012; Resende et al., 2012a). Genomic prediction relies on an available set of phenotypes and DNA marker data for a training population (TRN) that is used to fit a model to predict breeding

values (BV) based on DNA marker data alone for a testing population (TST). This methodology requires that the genome has been covered by a sufficiently dense panel of markers, that moderate to high linkage disequilibrium (LD) exists between marker loci and the underlying quantitative trait loci and that there is some degree of relatedness between the TRN and TST populations (Meuwissen et al., 2001).

As pointed out by Goddard (2009), LD constrains the number of markers to what is defined as "the number of chromosome segments" in a segregating population, which depends on the effective population size (N_e). If N_e decreases, it is expected that the individuals within the population will share larger chromosome segments, increasing prediction accuracy (Clark et al., 2012). Moreover, as N_e decreases, variability on which to select will decrease, but relatedness between individuals will increase leading to greater LD in the population (Albrecht et al., 2014). Therefore, GP methods will capture both LD and relatedness among individuals in the TRN and TST populations for predictions (Habier et al., 2007; Albrecht et al., 2014). Understanding the relative impacts of LD and relatedness in a breeding program may be helpful, since LD has greater potential to persist across populations and generations (Hayes et al., 2009).

Predictive ability (PA) is defined as the correlation between the observed phenotypic value and the BV: $[r(y, \hat{g})]$, and prediction accuracy is the correlation between the true BV and the estimated BV, $[r(g, \hat{g})]$ (Habier et al., 2007). Different empirical equations can be used to estimate prediction accuracy for GP in one population (Daetwyler et al., 2008; VanRaden, 2008), or multiple populations, traits and environments (Wientjes et al., 2015, 2016). However, there is a concern that after several consecutive breeding cycles using GP the prediction accuracy will decline due to changes in marker allele frequency (Habier et al., 2007; Goddard, 2009), and a gradual decay of LD. Therefore, it is suggested that GP models need to be periodically re-trained to sustain long-term genetic gains (Habier et al., 2007).

Assessment of GP is not trivial. Some published studies have been based on a single population with the use of cross-validation techniques (Crossa et al., 2010; Albrecht et al., 2011; Resende et al., 2012b). Cross-validation is a statistical technique used to evaluate models where an independent dataset is not available for validation. The most common approach, in the context of GP, is the k-fold cross-validation. Here, individual observations are randomly split into five or ten subsets, and all subsets except one are used as a training population with the remaining subset serving as a validation (or testing) population in a sequential approach. Because the same original population is both part of the TRN and TST populations, predictive ability and prediction accuracy from cross-validation are often upwardly biased (Amer and Banos, 2010; Michel et al., 2016), resulting in over-optimistic models. A better alternative is to independently validate the model with another separate trial (Amer and Banos, 2010; Hofheinz et al., 2012).

Some reports on independent validation and cross-validation across environments for multiple generations using a two-stage analysis have been published (Albrecht et al., 2014; Auinger et al., 2016; Michel et al., 2016, 2017). In these studies, higher predictive abilities have been reported for cross-validation, with a TRN population sampling individuals from multiple generations and validating with an independent trial, rather than predicting from a single generation and validating with an independent trial. However, in other studies, no significant differences in predictive ability or prediction accuracy were found by using independent validation from either TRN populations constituted as cross-validation from multiple years or from single years (Sallam et al., 2015; Đorđević et al., 2019). Nevertheless, as breeding programs progress in their use of GP, independent validations will become the reference to evaluate any model.

For training populations tested across multiple environments, genotype-by-environment (G × E) interactions may be important. Several GP studies using real data under different scenarios of locations and/or environments have modeled the effects of G × E or marker × E interactions (Burgueño et al., 2012; Jarquín et al., 2014, 2017). Previous studies on genotype by location interaction (Whitaker et al., 2012) and genotype by year interaction (Gezan et al., 2017) in the strawberry (*Fragaria* × *ananassa*) production area of Central Florida have indicated either very low or the absence of G × E interaction for the main strawberry commercial traits.

The strawberry breeding program at the University of Florida (UF) conducts genetic trials at the Institute of Food and Agricultural Sciences, Gulf Coast Research and Education Center (GCREC) in Balm, FL, United States. Each year a clonally replicated field trial of advanced breeding selections is phenotyped for several polygenic traits and genotyped via single-nucleotide polymorphism (SNP) arrays. These advanced selections arose from previous marker-assisted seedling selection for simply inherited disease resistance and fruit quality traits (Roach et al., 2016; Mangandi et al., 2017; Noh et al., 2017; Salinas et al., 2019) and subsequent visual field selection of the seedlings. Yearly advanced selection trials represent the elite parent pool of the breeding program and have been used to test GP methods (Gezan et al., 2017) and to apply GP for parent selection. These accumulated trials now allow further evaluation of models in strawberry over multiple breeding cycles.

The overall objective of the present study was to inform practical approaches for the use of GP in the breeding of horticultural crops by examining multiple cycles in the UF strawberry breeding program. Our specific objectives were to: (1) examine the effects on predictive ability of combining multiple cycles (or years) into TRN populations in the forward and backward directions; and (2) examine the effects of relatedness among the TRN and TST populations, LD and *Ne* on changes in predictive ability over time.

MATERIALS AND METHODS

Population and Field Testing

The elite population of the UF strawberry breeding program is treated as a single breeding pool from which the top-ranked parents of the previous year are used in a partial circular mating design to generate a large population of seedlings to be evaluated. This mating design is a modification of a partial diallel design with a reduced number of four to five crosses per parent, that

TABLE 1 | Incidence matrix for common genotypes tested among trials (above diagonal), full-sib families (diagonal, in bold) and common parents of full-sib families among trials (below diagonal).

Trials	T2	T4	T6	T8	T10	*N*
T2	**33**	37	29	28	30	217
T4	8	**30**	57	40	43	240
T6	2	7	**45**	88	69	237
T8	3	1	14	**43**	107	273
T10	2	3	10	13	**28**	266

N is the total number of tested phenotypes excluding common genotypes across trials. The T2–T10 nomenclature for the five trials conducted in five successive years is according to Gezan et al. (2017).

fall along an off-diagonal matrix of parental crosses (White et al., 2007). The best seedling selections are established the following year in an advanced-selection trial, the structure of which consists of a mixture of full-sib families, half-sib families, advanced selections, and cultivars. A representation of the structure of the population across cycles is presented in **Table 1**.

Replicated seedling and advanced-selection trials were previously established at two sites, the Gulf Coast Research and Education Center (GCREC) in Balm, FL (lat. 27° 45′ 37.98″ N, long. 82° 13′ 32.49″ W) and at the Florida Strawberry Growers Association in Dover, FL (lat. 28° 0′ 55.55″ N, long. 82° 14′ 5.24″ W), during the 2013–2014 and 2014–2015 seasons. Very low genotype by location interactions were observed for yield and quality traits (Whitaker et al., 2012). Consequently, these trials were subsequently carried out only at the GCREC.

The populations included in the present study were established at the GCREC site during five consecutive seasons from 2013–2014 to 2017–2018. The strawberry breeding program uses an overlapping generation breeding strategy in which all the main breeding activities, crossing, testing, and selection, take place each year (Borralho and Dutkowski, 1998), therefore each trial was considered a cycle in this sense and was given an even-numbered code starting with season 2013–2014 as T2 and ending with 2017–2018 as T10 according to the naming convention of Gezan et al. (2017). Several common genotypes were tested across years including cultivars and advanced selections chosen for further testing in the breeding process (**Table 1**). Therefore, these are essentially independent trials established under different yearly environmental conditions. Seedlings were clonally propagated by runners in a summer nursery near Monte Vista, Colorado (T2 and T4 trials) and at Crown Nursery in Malin, Oregon (T6, T8, and T10) and established in the fruiting field at GCREC in the first 2 weeks of October in each year. Site preparation, trial establishment and trial maintenance was carried out according to standard commercial practices for west-central Florida (Torres-Quezada et al., 2018). Pest control, fertilization and weed control varied among seasons according to environmental conditions. Bare-root clonal plants were arranged in a randomized complete block design with either five or six replications per trial and raised beds within replication. Each bed was subdivided into five to nine plots, each with a common control genotype to account for environmental variation along the bed. Genotypes were represented by a single runner plant in each plot (**Supplementary Table S1**).

Phenotyping and Genotyping

Five yield and fruit quality traits were assessed weekly from mid-November to mid-March in all five trials. At each harvest date, all ripe fruit per plant was removed. All marketable fruit (grams) by plant were considered as early marketable yield (EMY) if harvested before the first day of February. Total marketable yield (TMY) was calculated as the marketable fruit by plant collected until the first week of March. Average fruit weigh in grams, AWT, was estimated as the TMY divided by the number of marketable fruit. Total culls (TC), or unmarketable fruit, were counted and expressed as a proportion of the total number of fruits per plant (%). Soluble solids content (SSC) was measured five times during the season in each trial and was calculated as the mean of all measurements. One ripe fruit from each plant was squeezed by hand onto a handheld digital refractometer.

There were a total of 1,715 entries planted in these five trials that were phenotyped and genotyped using the Affymetrix Axiom® IStraw90 (Bassil et al., 2015) and IStraw35 (Verma et al., 2017) SNP arrays. Quality control was performed on a total of 14,332 segregating SNP markers in which SNPs with MAF < 0.05, and missing marker data > 0.05 were eliminated, yielding a total of 9,908 markers for the analyses. Missing values for each of the markers were imputed based on average allele frequency. The 1,715 phenotypes represented 1,558 unique individuals including advanced selections and varieties that were repeated across trials.

Genomic Prediction Model Analyses

The GP approach implemented was based on best linear unbiased estimates (BLUE) following one-stage analysis of tested phenotypes adjusted for the experimental factors in each trial. In most years, row and column location of each plant in the trial was recorded and the general linear mixed model was modified by adding spatial factors (row, col) and correlated residuals (autoregressive of order 1 for row and column), or independent residual units. Hence, multiple linear mixed models were tested for each trait and evaluated based on the Akaike and Bayesian information criteria (AIC and BIC, respectively) as well as their numbers of parameters (Isik et al., 2017).

Genomic Best Linear Unbiased Prediction, GBLUP (VanRaden, 2008) allowed the testing of complex models and was used only to assess genotype by year interactions (G × Y) between pairs of years and calculate heritabilities. The multi-year model assumed the genotypes among years were correlated such that genetic correlations could be estimated among years, using a factor analytic variance-covariance structure with two unknown factors (as fully described by Smith et al., 2001). Factor analytic models have been used to a large degree in plant breeding programs to model G × E interaction with heterogeneous variances between environments, and have shown to work well for crop species in multi-environment tests (for example, Burgueño et al., 2007, 2012; Crossa et al., 2006; Oakey et al., 2016; Dias et al., 2018). We used a multivariate model with a factor analytic variance-covariance structure with two (K) unknown factor loadings. When the factor analytic model is applied to the matrix of genotypic effects in each year (u_g), the model can be written as: $u_g = (\Gamma \otimes I_m) f + \delta$, where Γ is the matrix of

K vector loadings, f is a vector of genotypic scores; I_m is the vector of genotypes in each year and δ is the vector of genetic regression residuals. The variance of the genotype effects by year takes the form: $var\ \left(u_g\right) = (\Gamma\ \Gamma^{'} + \ \psi) \otimes\ I_m$ where ψ is a diagonal matrix with ψ_i as the specific variance for the i^{th} year, and the matrix across years is $G = (\Gamma\ \Gamma^{'} + \ \psi)$.

In this analysis, a genomic relationship matrix G was generated using all 9,908 markers and following the methodology described by Yang et al. (2010). The G matrix and its inverse were performed with the software GenoMatrix (Nazarian and Gezan, 2016), and model fitting was carried out with ASReml-R version 4.0 (Butler et al., 2017) R version 3.5.1 (R Core Team, 2018).

Genomic prediction models, for this study, were obtained by Bayes B and GBLUP, however, Bayes B has been shown to capture both marker-quantitative trait loci association effects and genetic relationship effects better than BLUP methods (Zhong et al., 2009). Even though, GBLUP has indicated to have a good performance for real data application (de los Campos et al., 2013), in a previous strawberry prediction study (Gezan et al., 2017). Bayes B performed slightly better for low-heritability traits and was therefore the main focus in our estimation of predictive ability for each TST population. In Bayes B, the analysis of each trait within each year was performed according to the following mixed model: $y = 1\mu + Z\beta + e$, where y is the response vector of BLUES, μ is the intercept, β is a vector of random marker effects (coded 0, 1, 2) associated with the incidence matrix Z and e is the vector of residual effects. Bayes B is a variable selection and shrinkage method, which assumes that some SNP effects are non-zero with probability 1-π while others have zero effects with probability π, following a mixture of two different prior densities with a point of mass at zero and a slab with a scaled-t density (de los Campos et al., 2013). In this study, we defined the priors according to the default hyper-parameters recommended by Pérez and de los Campos (2014).

We estimated predictive abilities by fitting the model for each trait with data from each individual trial as a training set (e.g., T2) and predicting to other trials (or years), as testing sets (e.g., T4, T6). Therefore, when we used T2 as TRN population we made a prediction for all T4 to T10 trials, by employing a single matrix of marker effects. The genotypes in these trials are genetically related to various degrees, but they are statistically independent in the process of fitting and evaluating the genetic model. After the single predictions were performed, we increasingly averaged successive predictions from previous years to the latest cycle (T10) and evaluated their effect on predictive ability in both forward (T2, T24,...) and backward (T8642, T864,...) directions. Each of these combinations was evaluated including or excluding common genotypes trialed across years. The Bayes B model was fitted in R (R Core Team, 2018) using the R package BGLR (Pérez and de los Campos, 2014) implementing a Markov Chain Monte Carlo method with 50,000 iterations where the first 10,000 were used as a burn-in. Each trait in each year was run five times and the predictive ability (PA) was estimated as the average of all runs, and trace plots of the residual variance were checked. The heritability of adjusted clonal mean phenotypes was estimated using GBLUP, with and without common genotypes, as $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$, where σ_a^2 is the additive variance and σ_e^2 is the estimated residual variance. Even though there was a moderate number of full-sib families in each trial (**Table 1**), we did not estimate within-family predictive ability for each cycle because of the unbalanced and small number of seedlings per family, mostly varying between 3 and 10. Within-family predictive ability is estimated in a different study (in preparation) established for three consecutive years with few biparental crosses and a large number of seedlings per family (60–75).

Linkage Disequilibrium and Effective Population Size

The previously mentioned set of 9,908 SNP markers was used to estimate effective population size, *Ne*. This set of markers was selected out of 14,332 markers in season 2015–2016 using the GenoMatrix software (Nazarian and Gezan, 2016) and was used for all other GP analyses. A closely related set of 9,622 genetically mapped SNP markers from Axiom IStraw35 SNP array (Verma et al., 2017) were used to estimate linkage disequilibrium (LD) for the five trials – T2, T4, T6, T8, and T10. These markers were distributed among 28 linkage groups (LGs) with a minimum number of 15 markers and maximum number of 720 markers per LG (**Supplementary Table S2**). The multi-year dataset comprising all cycles was divided into five different subsets based on crossing year. The purpose of dividing datasets this way was to estimate the distribution of LD structure and *Ne* of each trial without the genetic background influence of parents and common genotypes among trials. All individuals from T2 were included: parents, selections, and ancestors connected to the rest of the trials. Datasets for subsequent cycles T4, T6, T8, and T10 for the purposes of LD and *Ne* estimation included no founders or check cultivars, as the inclusion of common individuals across trials might influence haploblock structure estimation.

The R packages synbreed (Wimmer et al., 2012) and LDcorSV (Desrousseaux et al., 2017) were used to estimate LD based on population relatedness (r^2) and without relatedness ($r^2{}_v$), respectively (Mangin et al., 2012). The LD decay in genetic distance (Mb) was fitted with a non-linear regression model within the synbreed package. *Ne* was estimated using an LD-based approach and allele frequency threshold of 0.05 (Waples, 2006) via NeEstimator v2.1 software (Do et al., 2014). NeEstimator V2.1 (2017) is a tool for estimating contemporary effective population size (Ne) using multi-locus diploid genotypes from population samples. Unlike V1, NeEstimator V2.1 does not include third-party programs; all methods are implemented by NeEstimator V2.1 code and also implements a bias-corrected version of the method based on linkage disequilibrium (LD).

RESULTS

Training GP Models With Multiple Cycles

The effect of using a GP model over multiple breeding cycles without retraining can be seen when using T2 as a training population for all successive cycles (**Figure 1**). For all traits

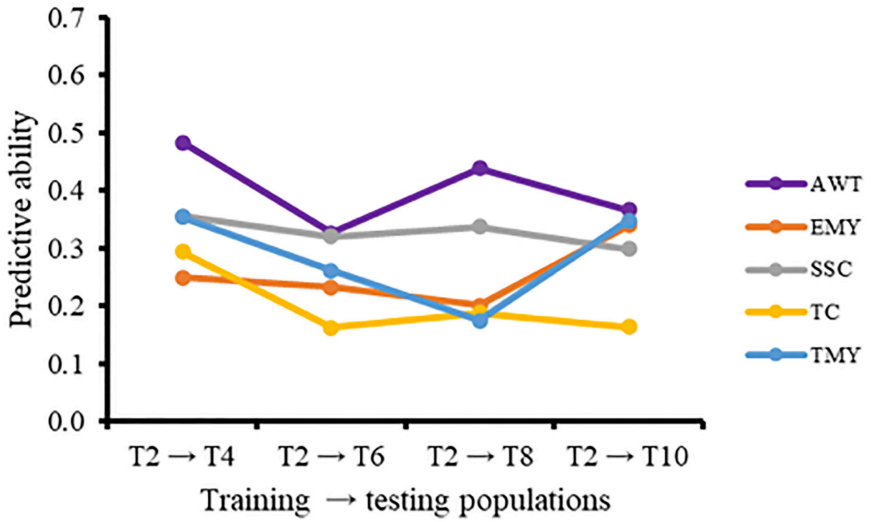

FIGURE 1 | Predictive ability (PA), without common genotypes and varieties, using T2 as an independent training population to predict later cycles for five traits. AWT, average fruit weight (g); EMY, early marketable yield (g per plant); SSC, soluble solids content (°Brix); TC, proportion of total culls (%); TMY, total marketable yield (g per plant).

except EMY there was a negative trend in predictive ability over time. The increase in predictive ability of EMY and TMY from cycle 2 to cycle 3 seems to be associated with an increase in heritability, from the TRN to the TST population, that was not present in other traits. The inclusion of additional cycles to the training population in the forward direction for prediction of trial T10 resulted in increased predictive abilities (**Figures 2A,B**). Predictive abilities for AWT and TMY tended to increase continuously, whether common genotypes across trials were included or not, while the trends for the other traits were more variable, but still showing an overall positive trend.

Predictive abilities were noticeably higher when common genotypes were included across cycles (**Figure 2**), and in this scenario backward predictions had on average higher predictive abilities for all traits than forward predictions. When common genotypes were included in the analyses, adding additional cycles to the training population in the backward direction gave little improvement. For example, there seemed to be no improvement in predictive ability when trial T2 was added to a training population consisting of trials T8, T6, and T4. However, when common genotypes were excluded, the addition of cycles to the training population in the backward direction noticeably improved predictive abilities for most traits.

Genetic Relationships

Single-cycle predictive abilities based on Bayes B are depicted in **Table 2**. The scenario in which all common genotypes between TRN and TST populations were included had a higher average predictive ability (0.35) than for the scenario excluding common genotypes (0.24), as expected. The trait AWT, when common individuals were included, had the highest average PA (0.43) of all traits across cycles, with a range from 0.38 to 0.53, followed by SSC (0.38), TMY (0.35), EMY (0.30), and TC (0.28). A similar pattern was noted when excluding common individuals, where AWT had the highest average PA (0.33) varying from 0.15 to 0.48, followed by SSC (0.26), TMY (0.24), EMY (0.18), and TC (0.18). The predictive abilities estimated by Bayes B and GBLUP were very similar (**Table 2** and **Supplementary Table S4**).

Heritabilities and G × E Interaction

Genomic heritability estimates are presented in **Figure 3**. Heritability estimates excluding common genotypes among trials between TRN and TST were lower than those estimates including common individuals across trials in 80% of the cases. However, the range of heritabilities in both scenarios was wide and similar, whether excluding or including common individuals, mostly varying from 0.15 to 0.65, except for the wider range for TC (0.0–0.81). Overall, average additive genetic correlations across trials were very high, indicating very little if any G × Y interaction (**Table 3**). Though a few values in some cycles showed moderate correlations, such as for EMY (0.70) and TC (0.72), all remaining values were higher than 0.79 (**Supplementary Table S3**).

Linkage Disequilibrium and Effective Population Size

A set of 9,622 markers were mapped to 40 linkage groups, the number of markers per LG varying from 15 to 720. We plotted r^2 and r^2v (r^2 with no relatedness bias) for T2 and T10 against genomic distances in Mb for T2 and T10 (**Figure 4**). We also compared the decay of LD between T2 and T10. Maximum r^2 was 0.4 in T2 and 0.47 in T10. In T2, r^2 decreased to 0.2 at 3.5 Mb (**Figure 4A**), compared to an r^2 of 0.2 at 4.2 Mb for T10 (**Figure 4C**). Similar trends were observed for r^2v, with a slower decay of LD in T10 compared to T2 (**Figures 4B,D**). Much higher values overall for r^2 compared to r^2v indicates that a substantial portion of apparent LD was due to relatedness

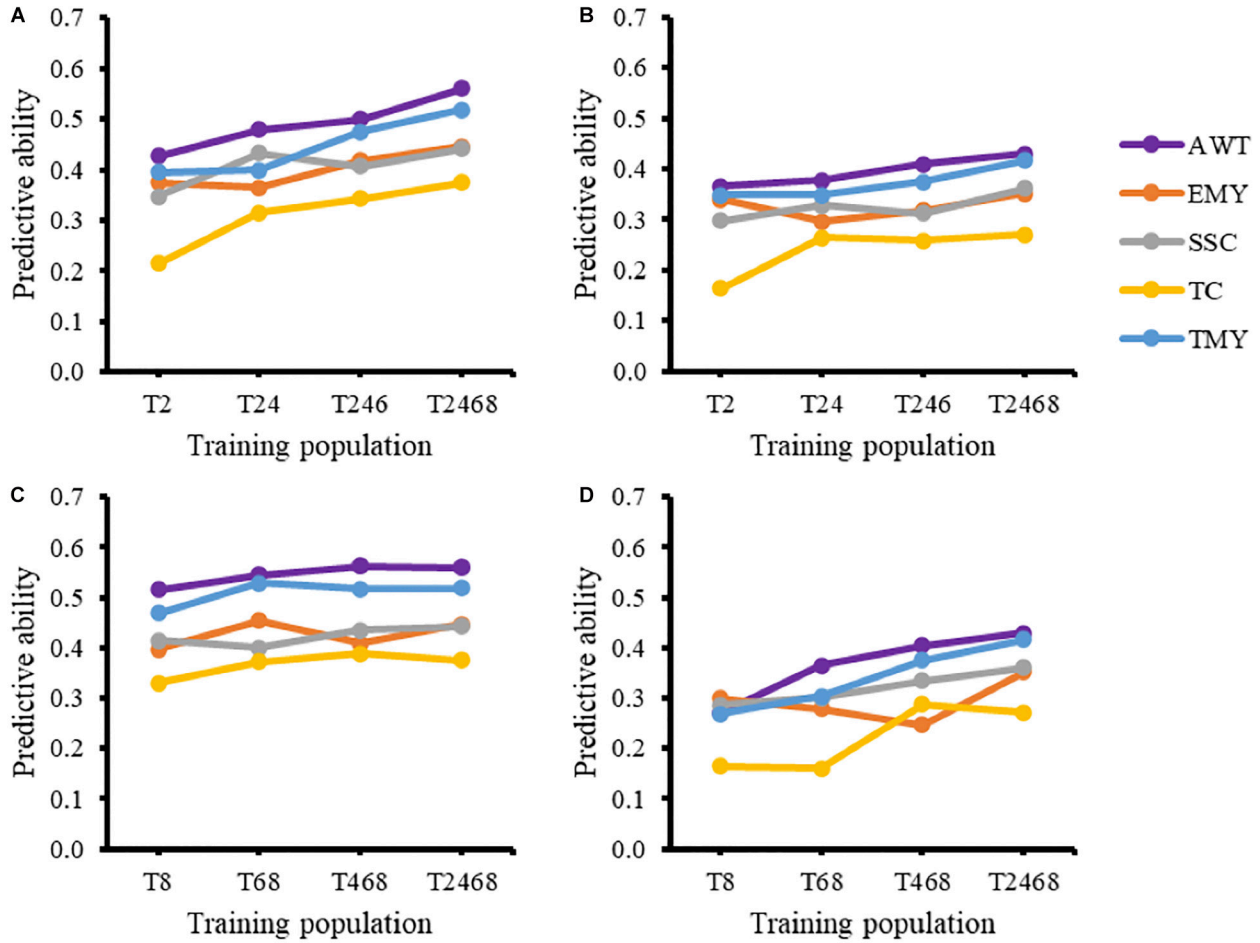

FIGURE 2 | Forward **(A,B)** and backward predictions **(C,D)** of trial T10 (testing population) and the effect of model averaging the training population on predictive ability under two scenarios: including common genotypes between the training and testing populations **(A,C)** and excluding them **(B,D)**. AWT, average fruit weight (g); EMY, early marketable yield (g per plant); SSC, soluble solids content (°Brix); TC, proportion of total culls (%); TMY, total marketable yield (g per plant).

(**Supplementary Table S2**). The effective population sizes, *Ne*, for each of the cycles were 25, 17, 23, 23, and 20 for T2, T4, T6, T8, and T10, respectively, possibly indicating a slight decrease over time.

DISCUSSION

Independent validation with TRN populations from five breeding cycles was utilized to evaluate GP methods and inform practical approaches for its implementation in the strawberry breeding program at UF. The impact of averaging multiple single predictions, genetic relationships among the cycles, heritabilities, G × Y interactions, LD and *Ne* were explored separately. The estimation of trait additive correlations across years, G × Y, using multivariate analyses is complex due to the heterogenous variances-covariances among environments and the environmental effects to be fitted. When the number of traits is high using a parsimonious FA matrix in modeling the G × Y interaction has advantages in convergence compared to models using an unstructured variance-covariance matrix. Previous results showed that increasing the number of components of FA models would give better estimates of variance-covariance estimates; however, these models may or may not increase predictive ability, and it is questionable whether it would improve the model fit (Burgueño et al., 2011). Though our estimates of additive correlations across years (**Table 3**) might be upwardly biased, they reflect the low G × Y interactions present for the traits evaluated.

Our focus on the estimation of predictive abilities was due to the primary emphasis in this study on practical outcomes and applications; however, it is possible to use deterministic formulae to calculate prediction accuracies between different cycles, which we would expect to provide very similar trends (Wientjes et al., 2015). Prediction accuracy and the reliability of predictions has been shown to decline across generations due to a decrease in genetic relationships between the TRN and TST populations (Habier et al., 2007; Pszczola et al., 2012) as well as the break-up of LD and consequent reduction of genetic variance explained by the markers (Goddard, 2009). Therefore,

TABLE 2 | Forward predictive ability (PA) for five traits estimated using Bayes B, for pairs of trials using: (A) all individuals including varieties and advanced selections in common among each pair of trials, and (B) excluding common individuals.

		(A)					(B)				
Trait	**Trial**	**T2**	**T4**	**T6**	**T8**	**T10**	**T2**	**T4**	**T6**	**T8**	**T10**
AWT	T2		[a]0.53	0.39	0.45	0.43		0.48	0.33	0.44	0.37
	T4			0.38	0.41	0.44			0.31	0.29	0.30
	T6				0.42	0.42				0.15	0.32
	T8					0.51					0.27
	T10										
EMY	T2		0.30	0.32	0.22	0.37		0.25	0.23	0.20	0.34
	T4			0.18	0.23	0.26			0.06	0.13	0.12
	T6				0.28	0.40				0.09	0.13
	T8					0.40					0.30
	T10										
SSC	T2		0.42	0.38	0.41	0.35		0.36	0.32	0.34	0.30
	T4			0.40	0.39	0.40			0.27	0.23	0.25
	T6				0.35	0.27				0.11	0.15
	T8					0.41					0.29
	T10										
TC	T2		0.32	0.24	0.19	0.22		0.29	0.16	0.19	0.16
	T4			0.36	0.29	0.33			0.10	0.24	0.28
	T6				0.20	0.29				0.15	0.06
	T8					0.33					0.17
	T10										
TMY	T2		0.40	0.36	0.24	0.39		0.35	0.26	0.17	0.35
	T4			0.29	0.24	0.33			0.16	0.09	0.25
	T6				0.32	0.46				0.22	0.24
	T8					0.47					0.27
	T10										

AWT, average fruit weight (g); EMY, early marketable yield (g per plant); SSC, soluble solids content (%); TC, proportion of total culls (%); TMY, total marketable yield (g per plant). [a]*Predictive ability ranges from low (light color) to high (dark color).*

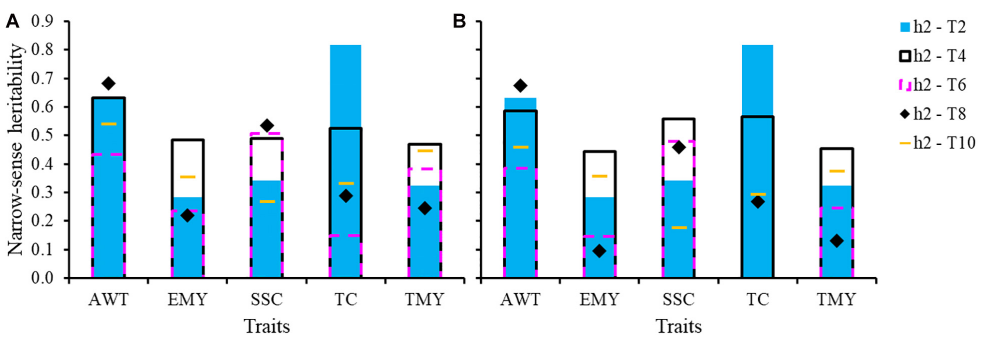

FIGURE 3 | Genomic narrow-sense heritabilities for five traits for each trial with: **(A)** all genotypes including varieties and advanced selections in common among pairs of trials, and **(B)** excluding common genotypes. AWT, average fruit weight (g); EMY, early marketable yield (g per plant); SSC, soluble solids content (°Brix); TC, proportion of total culls (%); TMY, total marketable yield (g per plant).

retraining models for GP is recommended every generation (Wolc et al., 2011; Pszczola and Calus, 2016). Currently, in the UF strawberry breeding program the decay of predictive ability over successive cycles without including common individuals (**Figure 1**) is offset by updating the GP model every year with phenotypic and marker data from the latest field trial. Besides, significant decreases in selection accuracy over generations are not expected if marker density is sufficiently high (Solberg et al., 2008). The number of markers used in this set of trials (∼10,000) might be considered small when compared with some other breeding programs, particularly for animals. However, the most complete strawberry genetic map developed for UF germplasm (unpublished) has a total length of 1729.5 cM, meaning that on average more than five markers per cM were

utilized in this study, which should be more than enough to account for genome-wide allelic diversity in an elite strawberry breeding population.

The results obtained by comparing predictive abilities estimated by Bayes B, as well as a previous report using different methods of predictions (Gezan et al., 2017), indicate that, for the commercial traits reported, Bayes B may produce slightly greater predictive abilities than GBLUP. Therefore, we are using Bayes B operationally in the breeding program and have focused on the use of Bayes B for this report. Overall, predictive abilities using single cycles (or trials) as training populations (**Table 2**) were in the general range of estimates reported from other crops and environments (Sallam et al., 2015; Đorđević et al., 2019). Using multiple cycles by averaging predictions across cycles noticeably increased predictive ability, whether individuals common to multiple trials were included in the analyses or not. Thus, the size of the training population, which is known to be important for the success of GP, was increased, not in the traditional sense (Asoro et al., 2011; Zhang et al., 2017), but with the addition of independent training populations from each cycle. Improvements in the estimation of PAs by adding multiple cycles of training populations could also come from averaging G × Y interaction effects, though we have shown these to be quite low (**Table 3**).

The presence of population structure across the breeding cycles has important effects on GP (Asoro et al., 2011). Genetic relationships in the strawberry breeding populations studied arise from two primary sources: the first is the continued testing across years of promising advanced selections and check cultivars during the process of variety development, and the second is the use of common parents across years which increases relatedness at the half-sib family level (**Table 1**). The impacts of genetic relationships and cosegregation can be seen by comparing the structure of the TRN populations in **Table 1** with the predictive abilities in **Table 2** when including common individuals and when excluding them. As shown in **Table 1**, the average number of common genotypes among T2 or T4 with the other trials is 31 and 44 genotypes, respectively. Among the T6, T8, and T10 trials the average number of common individuals with others is 61, 66, and 62, respectively, partly reflecting the larger number of genotypes included in these later trials. This helps to explain the increasing average differences in predictive ability across traits over time between scenarios where common individuals are included versus excluded: T2 (0.05), T4 (0.12), T6 (0.18), and T8 (0.17). Common parents as a source of relatedness is highlighted by the fact that the average number of parents shared among individuals for either T2 or T4 with the other trials is four and five, respectively, but for T6, T8, and T10 trials the average number of shared parents is eight, eight and seven, respectively. In other words, the increase in genetic relationships across cycles over time is clearly one of the factors favoring predictive ability in this breeding program.

The strength of family relationships within and across populations has been shown to influence the reliability and the accuracy of genomic predictions in several studies. In Pszczola et al. (2012) the effect of four TRN populations with increasing numbers of half-sib families (5, 20, 40) for a fixed number of offspring and a random population with the same number of individuals was simulated. Based on their results and other studies (Calus, 2010), the authors concluded that highly related TRN populations that have a small number of families with large number of offspring per family yield lower accuracy of prediction compared to TRN populations with more half-sib families or random populations. In the UF strawberry breeding program the composition of the TRN population is largely determined by the field performance of seedlings selected in the previous year. Different numbers of seedlings are selected from each full-sib family based on performance, while also aiming to have, if possible, all families represented to maintain genetic diversity. This resulted in small and unbalanced numbers of individuals representing each full-sib family, which is why within-family predictions were not performed in this study. Ultimately, optimizing the design of the TRN population at the family level is achievable, but constraining the number of selections in the best families may negatively affect genetic gains, at least in the short-term. The increase from two common parents between T2 and T10 to 13 common parents between T8 and T10 might have had a positive effect on predictive ability. Yet this is not obvious, since in the scenario of excluding common individuals the predictive ability for all traits from T8 to T10 (**Figure 2D**) was lower than the predictive ability from T2 to T10 (**Figure 2B**), indicating the low impact of the number of half-sibs in this scenario. When including common individuals, the situation is reversed, with T8 having greater ability than T2 to predict T10. It is also important to note that backward predictions when common individuals are included quickly reach a plateau, with the addition of T6 to T8 giving a very small increase in PA and the addition of T4 and T2 giving no improvement (**Figure 2C**). Together these results highlight the importance of relatedness to predictive ability, particularly in the case of common individuals.

TABLE 3 | Average additive genetic correlations for five traits across trials, including common individuals among trials, using GBLUP and a factor analytic of order 2 (FA2) variance-covariance matrix, together with the proportion of the total genetic variance explained (VE%) by FA2.

r_a	AWT	EMY	SSC	TC	TMY
Mean	0.96	0.95	0.9	0.9	0.95
Range	0.94–1.00	0.69–1.00	0.87–1.00	0.72–1.00	0.86–1.00
VE%	97.9	100.0	97.3	98.9	100.0

AWT, average fruit weight (g); EMY, early marketable yield (g per plant); SSC, soluble solids content (%); TC, proportion of total culls (%); TMY, total marketable yield (g per plant).

Marker-based genomic heritability estimates from this study are higher than the previously reported pedigree-based estimates for T2 and T4 (Gezan et al., 2017). This is not surprising, as marker-based relationships are more precise. Many studies have shown positive correlations between predictive ability and narrow sense heritability, consistent with the present study (Calus et al., 2008; Daetwyler et al., 2008). The presence of G × Y interactions may cause rank changes across years, when pairwise genetic correlations among years are below r_a = 0.8 (White et al., 2007; Goddard and Hayes, 2007). In this study,

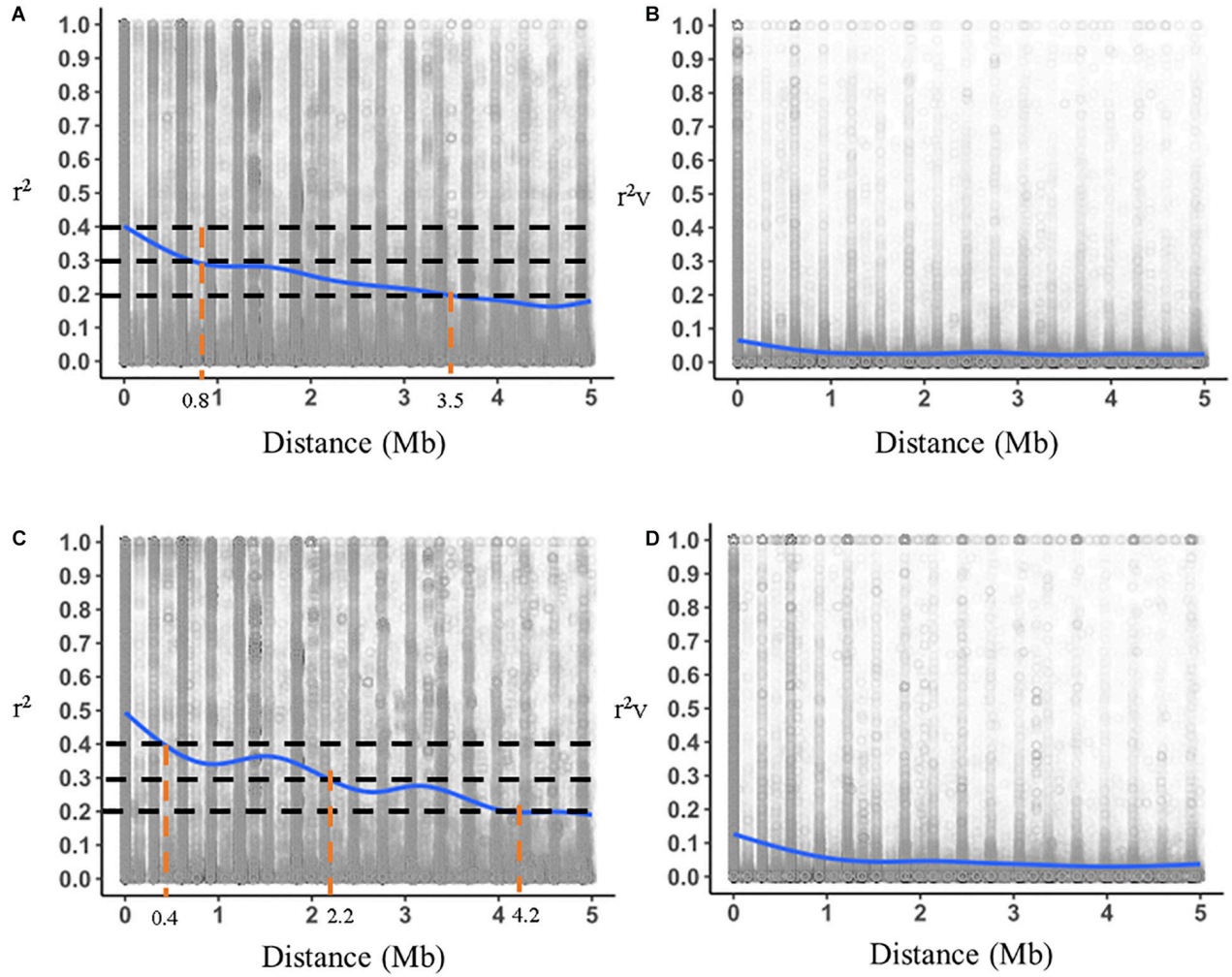

FIGURE 4 | Comparison of genome-wide decay of r^2 (linkage disequilibrium with relatedness bias) **(A,C)** and r^2v (linkage disequilibrium without relatedness bias) **(B,D)** for the T2 (Cycle 0) and T10 (Cycle 4) trials. Horizontal dotted lines represent thresholds of r^2 for comparison and vertical dotted lines represent genomic distances where LD intersects with thresholds.

almost all additive correlations were above 0.8, suggesting low G × Y interactions that will have little effect on PA. Most of the strawberry production in Florida is concentrated within a 30-mile radius of Plant City, and genotype by location interaction is minimal within this region. On the other hand, G × Y is more unpredictable and should be monitored closely over time. Modeling G × Y could allow trials to be pooled into a single training population, as opposed to averaging predictions across cycles, possibly improving PA.

Estimates of intra-linkage group regular pairwise LD (r^2) and LD corrected for relatedness (r^2_v) for T2 were slightly lower than our previous estimates of $r^2 = 0.26$ and $r^2_v = 0.04$ (Gezan et al., 2017). One possible reason is that the original study utilized 17,479 markers from the IStraw90 SNP array, while the present analysis was based on 9,622 markers from the IStraw35 array (Verma et al., 2017) which also provides the same quality of data but at a reduced cost. Simulation studies have shown that overestimation of LD (r^2) comes first from multiples copies of the same genotype and second from the progeny of full-sib families (Mangin et al., 2015). In our analysis, we estimated r^2 based on a single copy of each phenotype (common individuals removed), but there were multiple full-sib families with different numbers of offspring in each cycle; therefore, the bias of the r^2 estimate should only be due to this second factor. The presence of LD corrected for relatedness is the driving force for the long-term success of GP in the breeding population, as r^2_v represents the prediction accuracy that will tend to persist over multiple cycles without the need for retraining (Mangin et al., 2012; Habier et al., 2013). The dramatic decrease in LD when removing relatedness bias once again emphasizes the importance of relatedness in this population as it relates to the success of GP models.

The impact of *Ne* on prediction accuracy has been reported in animals, forest trees and tree fruit species (Kumar et al., 2012; Daetwyler et al., 2013; Bartholome et al., 2016). In long generation tree species, the use of elite populations with *Ne* ranging from 10 to 50 is a common practice to increase genetic gains. In this

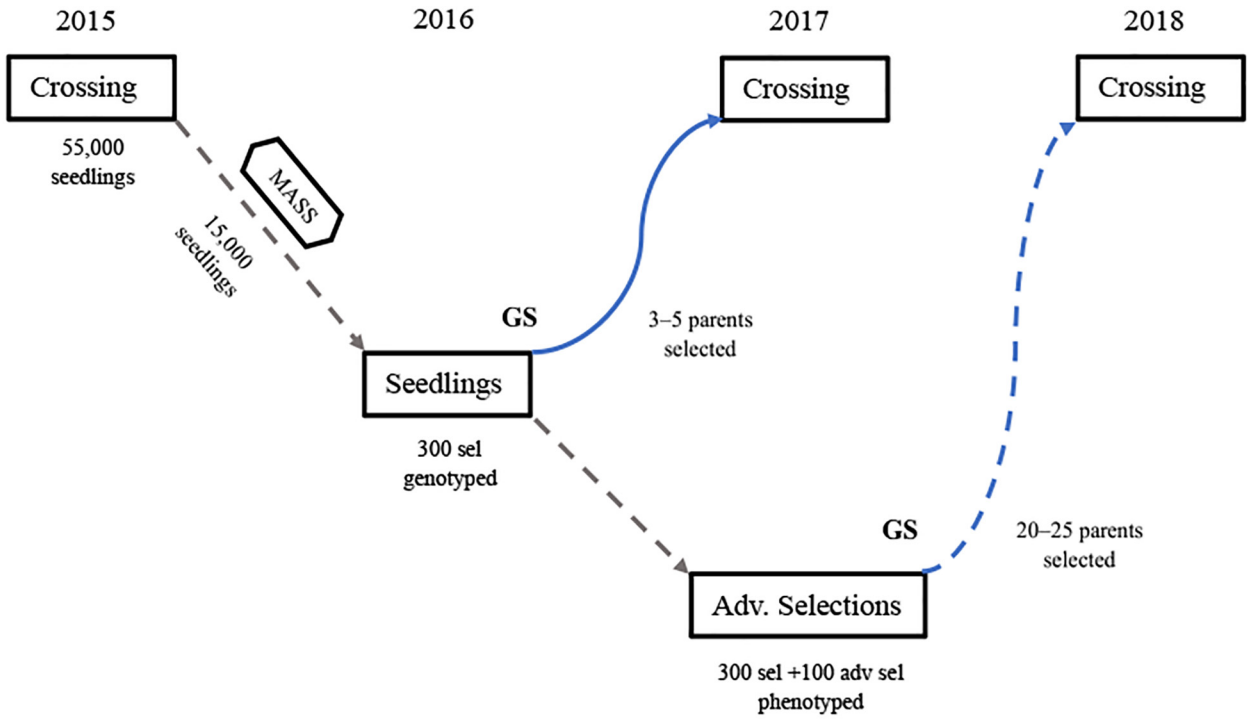

FIGURE 5 | A single UF strawberry breeding cycle (overlapping cycles not shown) combining marker-assisted seedling selection (MASS) for disease resistance and other simply controlled traits, combined with the early implementation of GP for using untested genotypes in crosses 1 year early (blue line), prior their phenotyping in a replicated field trial of elite clones (GP trial). Once all clones are phenotyped, GP models are updated with that information for the estimation of BVs to guide future crosses (dashed line). For ease of visualization, this figure shows only the cycle beginning in 2015. However, due to yearly overlapping cycles, all breeding program activities including crossing, MASS, seedling trials, GP trials, etc., are carried out every year.

study, effective population size appears to have decreased slightly from T2 (Ne = 25) to T10 (Ne = 20). In the present study this apparent slight reduction in Ne and the corresponding increase in the extent of LD from T2 to T10 are likely contributing to increased predictive ability with the addition of later cycles. In the long-term it is important to recognize that intensive recurrent selection increases inbreeding. Therefore, to maintain long-term breeding progress, it will be important to continue to introgress diversity into the elite breeding population.

The last 5 years of implementation of GP in the UF strawberry breeding program has allowed the use of some parents earlier in the breeding cycle and has increased the accuracy of estimation of breeding values. This study makes clear that the use of average predictions from multiple cycles in training GP models is very beneficial, at least up to four cycles when common individuals are included across trials. Based on these results, the following steps are currently used for the application of GP in the UF strawberry breeding program (**Figure 5**):

(1) In the summer prior to each winter fruiting/crossing season, which in Florida extends roughly from mid-November through March, phenotypic and marker data from up to four previous cycles, including common individuals across trials, are used to train Bayes B models predicting the BVs of the most recent advanced selections. These selections were seedlings in the previous cycle and are genotyped over the summer but are not yet phenotyped for the five measured commercial traits AWT, EMY, SSC, TC, and TMY.
(2) Breeding values for these five traits are combined in a selection index using economic weights for each trait to rank the advanced selections for their overall potential as parents.
(3) In November and December, early-season field observations are made for these advanced selections for all visually evaluated traits, including: fruit shape, color, and flavor, disease resistance, plant architecture, etc.
(4) Three to five advanced selections (out of approximately 25–40 total parents) that are noted for early-season field traits and ranked highly in the BV selection index are selected for use as parents in controlled crosses as males. These males are crossed to one or more elite females that have been field evaluated for multiple seasons and have complementary traits to the males chosen by GP. In this way, approximately 10% of crosses have a male parent chosen via GP methods that is being used in crossing at least 1 year earlier in the breeding cycle than normal.

As this study suggests, increasing the size of the training population will increase prediction accuracy, but at some point, increasing size will not further improve GP models. This appears to have occurred for the UF strawberry breeding program at the

fourth cycle. Given the demonstrated importance of relatedness in this study, future work on the optimal design of the relatedness within and among TRN and TST populations (choosing which genotypes to establish in each trial) could possibly increase predictive ability in the short term without compromising the potential of future genetic gains. It will also be important to monitor the performance of crosses chosen via GP versus those designed in the traditional manner to empirically test whether the implementation of GP in the breeding program is achieving the desired results.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in Data Dryad via the following link: https://datadryad.org/stash/dataset/doi:10.5061/dryad.b5mkkwhc7.

AUTHOR CONTRIBUTIONS

VW, SG, and LO conceived and designed the study. SV prepared the SNP data for the GP analyses and carried out the analyses of LD and Ne. SG and LO performed the GP analyses. LO wrote the initial draft and VW, SG, and SV corrected it and improved it. All authors read and approved the manuscript.

FUNDING

Financial support was provided through the Florida Agricultural Experiment Station, the Florida Strawberry Growers Association, and two USDA/NIFA Specialty Crop Research Initiative projects: "RosBREED: Combining disease resistance with horticultural quality in new rosaceous cultivars" under award number 2014-51181-22378 and "Next-Generation Disease Resistance Breeding and Management Solutions for Strawberry" under Award Number 2017-51181-26833.

ACKNOWLEDGMENTS

The authors acknowledge the continuous efforts of the UF strawberry breeding staff in establishing, maintaining, collecting, and providing the phenotypic data from all the genetic trials used in these analyses. The authors also thank Dr. Marcio Resende for his comments on an earlier version of this manuscript and the reviewers for improving the initial version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.596258/full#supplementary-material

REFERENCES

Albrecht, T., Auinger, H. J., Wimmer, V., Ogutu, J. O., Knaak, C., Ouzunova, M., et al. (2014). Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theor. Appl. Genet.* 127, 1375–1386. doi: 10.1007/s00122-014-2305-z

Albrecht, T., Wimmer, V., Auinger, H. J., Erbe, M., Knaak, C., Ouzunova, M., et al. (2011). Genome-based prediction of testcross values in maize. *Theor. Appl. Genet.* 123, 339–350. doi: 10.1007/s00122-011-1587-7

Amer, P. R., and Banos, G. (2010). Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit. *J. Dairy Sci.* 93, 3320–3330. doi: 10.3168/jds.2009-2845

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., and Jannink, J.-L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome* 4, 132–144. doi: 10.3835/plantgenome2011.02.0007

Auinger, H. J., Schönleben, M., Lehermeier, C., Schmidt, M., Korzun, V., Gelger, H. H., et al. (2016). Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). *Theor. Appl. Genet.* 129, 2043–2053. doi: 10.1007/s00122-016-2756-5

Bartholome, J., Heerwaarden, J. V., Isik, F., Boury, C., Vidal, M., Plomion, C., et al. (2016). Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics* 17:604. doi: 10.1186/s12864-016-2879-8

Bassil, N. V., Davis, T. M., Zhang, H., Ficklin, S., Mittmann, M., Webster, T., et al. (2015). Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. *BMC Genomics* 16:155. doi: 10.1186/s12864-015-1310-1

Bernardo, R., and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47, 1082–1090. doi: 10.2135/cropsci2006.11.0690

Borralho, N. M. G., and Dutkowski, G. W. (1998). Comparison of rolling front and discrete generation breeding strategies for trees. *Can. J. For. Res.* 28, 987–993.

Burgueño, J., Crossa, J., Cornelius, P. L., Trethowan, R., McLaren, G., and Krishnamachari, A. (2007). Modeling additive x environment and additive x additive x environment using genetic covariances of relatives of wheat genotypes. *Crop Sci.* 47, 311–320. doi: 10.2135/cropsci2005.11-0427

Burgueño, J., Crossa, J., Cotes, J. M., San-Vicente, F., and Das, B. (2011). Prediction assessment of linear mixed models for multienvironment trials. *Crop Sci.* 51, 944–954. doi: 10.2135/cropsci2010.07.0403

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52, 707–719. doi: 10.2135/cropsci2011.06.0299

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. G., and Thompson, R. (2017). *ASReml-R Reference Manual Version 4*. Hemel Hempstead: VSN International Ltd.

Calus, M. P. L. (2010). Genomic breeding value prediction: methods and procedures. *Animal* 4, 157–164. doi: 10.1017/S1751731109991352

Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W., and Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 128, 553–561. doi: 10.1534/genetics.107.080838

Clark, S. A., Hickey, J. M., Daetwyler, H. D., and van der Werf, J. H. J. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of references data sets in livestock breeding schemes. *Genet. Sel. Evol.* 44:4. doi: 10.1186/1297-9686-44-4

Crossa, J., Burgueño, J., Cornelius, P. L., McLaren, G., Trethowan, R., Krishnamachari, A., (2006). Modeling genotype x environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Sci.* 46, 1722–1733. doi: 10.2135/cropsci2005.11-0427

Crossa, J., de los Campos, G., Perez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Daetwyler, H. S., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi: 10.1534/genetics.112.147983

Daetwyler, H. S., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3:e3395. doi: 10.1371/journal.pone.0003395

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

Desrousseaux, D., Sandron, F., Siberchicot, A., Cierco-Ayrolles, C., and Mangin, B. (2017). *LDcorSV: Linkage Disequilibrium Corrected by the Structure and the Relatedness. R Package Version 1.3.2*.

Dias, K. O. D. G., Gezan, S. A., Guimarães, C. T., Nazarian, A., da Costa-Silva, L., Parentoni, S. N., et al. (2018). Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. *Heredity* 121, 24–37. doi: 10.1038/s41437-018-0053-6

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillet, B. J., and Ovenden, J. R. (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Mol. Ecol. Resour.* 14, 209–214. doi: 10.1111/1755-0998.12157

Đorđević, V., Ćeran, M., Miladinović, J., Balešević-Tubić, S., Petrović, K., Miladinov, Z., et al. (2019). Exploring the performance of genomic prediction models for soybean yield using different validation approaches. *Mol. Breed.* 39:74. doi: 10.1007/s11032-019-0983-6

Gezan, S. A., Osorio, L. F., Verma, S., and Whitaker, V. M. (2017). An experimental validation of genomic selection in octoploid strawberry. *Hort. Res.* 4:16070. doi: 10.1038/hortres.2016.70

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximization of long-term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Goddard, M. E., and Hayes, B. J. (2007). Genomic selection. *J. Anim. Breed. Genet.* 124, 323–330. doi: 10.1111/j.1439-0388.2007.00702.x

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics* 194, 597–607. doi: 10.1534/genetics.113.152207

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646

Hofheinz, N., Borchardt, D., Weissleder, K., and Frisch, M. (2012). Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theor. Appl. Genet.* 125, 1639–1645. doi: 10.1007/s00122-012-1940-5

Isik, F., Holland, J., and Maltecca, C. (2017). *Genetic Data Analysis for Plant and Animal Breeding*. Berlin: Springer.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607. doi: 10.1007/s00122-013-2243-1

Jarquín, D., da Silva, C. L., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., et al. (2017). Increasing genomic prediction accuracy by modeling G x environment interactions in Kansas wheat. *Plant Genome* 10, 1–15.

Kumar, S., Bink, M. C. A. M., Volz, R. K., Bus, V. G. M., and Chagné, D. (2012). Towards genomic selection in Apple (*Malus* x domestica Borkh.) breeding programmes: prospects, challenges, and strategies. *Tree Gen. Genomes* 8, 1–14. doi: 10.1007/s11295-011-0425-z

Mangandi, J., Verma, S., Osorio, L., Peres, N., van de Weg, E., and Whitaker, V. M. (2017). Pedigree-based analysis in a multiparental population of octoploid strawberry reveals QTL alleles conferring resistance to *Phytophthora cactorum*. *G3* 7, 1707–1719. doi: 10.1534/g3.117.042119

Mangin, B., Sandron, F., Henry, K., Devaux, B., Willems, G., Devaux, P., et al. (2015). Breeding patterns and cultivated beets origins by genetic diversity and linkage disequilibrium analyses. *Theor. Appl. Genet.* 128, 2255–2271. doi: 10.1007/s00122-015-2582-1

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2012). Novel measures of linking disequilibrium that correct the bias due to population structure and relatedness. *Heredity* 108, 285–291. doi: 10.1038/hdy.2011.73

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

Michel, S., Ametz, C., Gungor, H., Akgfil, B., Epure, D., Grausguber, H., et al. (2017). Genomic assisted selection for enhancing line breeding: merging genomic and phenotypic selection in winter wheat breeding programs with preliminary yield trials. *Theor. Appl. Genet.* 130, 363–376. doi: 10.1007/s00122-016-2818-8

Michel, S., Ametz, C., Gungor, H., Epure, D., Grausguber, H., Lfinscherberger, F., et al. (2016). Genomic selection across multiple breeding cycles. *Theor. Appl. Genet.* 129, 1179–1189. doi: 10.1007/s00122-016-2694-2

Nazarian, A., and Gezan, S. A. (2016). GenoMatrix: a software package for pedigree-based and genomic prediction analyses on complex traits. *J. Hered.* 107, 372–379. doi: 10.1093/jhered/esw020

Noh, Y. H., Lee, S., Whitaker, V. M., Cearley, K. R., and Cha, J. S. (2017). A high-throughput marker-assisted selection system combining rapid DNA extraction and high-resolution melting and simple sequence repeat analysis: strawberry as a model for crops. *J. Berry Res.* 7, 23–31. doi: 10.3233/JBR-160145

Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C., and Waugh, R. (2016). Genomic selection in multi-environment crop trials. *G3* 6, 1313–1326. doi: 10.1534/g3.116.027524

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

Pszczola, M., and Calus, M. P. L. (2016). Updating the reference population to achieve constant predictive reliability across generations. *Animal* 10, 1018–1024. doi: 10.1017/S1751731115002785

Pszczola, M., Strabel, T., Mulder, A., and Calus, M. P. L. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95, 389–400. doi: 10.3168/jds.2011-4338

R Core Team, (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Resende, M. F. R. J., Muñoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., et al. (2012a). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol.* 193, 617–624.

Resende, M. F. R. J., Muñoz, P., Resende, M. D. V., Garrick, D. G., Fernando, R. L., Davis, J. M., et al. (2012b). Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda l.). *Genetics* 190, 1503–1510. doi: 10.1534/genetics.111.137026

Roach, J., Verma, S., Peres, N., Jamieson, A., van de Weg, E., Bink, M. C. A. M., et al. (2016). FaRXf1: a locus conferring resistance to angular leaf spot caused by Xanthomonas fragariae in octoploid strawberry. *Theor. Appl. Genet.* 129, 1191–1201. doi: 10.1007/s00122-016-2695-1

Salinas, N. R., Verma, S., Peres, N., and Withaker, V. M. (2019). FaRCa1: a major subgenome-specific locus conferring resistance to Colletotrichum acutatum in strawberry. *Theor. Appl. Genet.* 132, 1109–1120. doi: 10.1007/s00122-018-3263-7

Sallam, A. H., Endelman, J. B., Jannink, J.-L., and Smith, K. P. (2015). Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *Plant Genome* 8, 1–15. doi: 10.3835/plantgenome2014.05.0020

Smith, A., Cullis, B., Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57, 1138–1147. doi: 10.1111/j.0006-341x.2001.01138.x

Solberg, T., Sonesson, R. A. K., Woolliams, J. A., and Meuwissen, T. H. E. (2008). Genomic selection using different marker types and densities. *J. Anim. Sci.* 86, 2447–2454. doi: 10.2527/jas.2007-0010

Torres-Quezada, E. A., Zotarelli, L., Whitaker, V. M., Darnell, R. L., Santos, B. M., and Morgan, K. (2018). Planting dates and transplant establishment methods on early-yield strawberry in west-central Florida. *Hortech* 28, 615–623. doi: 10.21273/HORTTECH04079-18

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy. Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Verma, S., Bassil, N. V., van de Weg, E., Harrison, R. J., Monfort, A., Hidalgo, J. M., et al. (2017). Development and evaluation of the Axiom® IStraw35 384HT array for the allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. *Acta Hortic.* 1156, 75–82. doi: 10.17660/ActaHortic.2017.1156.10

Waples, R. S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conserva. Genet.* 7, 167–184. doi: 10.1007/s10592-005-9100-y

Whitaker, V. M., Osorio, L. F., Hasing, T., and Gezan, S. (2012). Estimation of genetic parameters for 12 fruit and vegetative traits in the University of

Florida strawberry breeding population. *J. Amer. Soc. Hort. Sci.* 137, 316–324. doi: 10.21273/JASHS.137.5.316

White, T., Adams, W. T., and Neale, D. B. (2007). *Forest Genetics*. Cambridge, MA: CABI Publishing.

Wientjes, Y. C. J., Bijma, P., Veerkamp, R. F., and Calus, M. P. L. (2016). An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. *Genetics* 202, 799–823. doi: 10.1534/genetics.115.183269

Wientjes, Y. C. J., Veerkamp, R. F., Bovenhuis, H., Schrooten, C., Bijma, P., and Calus, M. P. L. (2015). Empirical and deterministic accuracies of across population genomic prediction. *Genet. Sel. Evol.* 47:5. doi: 10.1186/s12711-014-0086-0

Wimmer, V., Albrecht, T., Auinger, H. J., and Schfin, C. C. (2012). Synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* 28, 2086–2087. doi: 10.1093/bioinformatics/bts335

Wolc, A., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., et al. (2011). Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Gen. Sel. Evol.* 43:23.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., et al. (2017). Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. *Front. Plant Sci.* 8:1916. doi: 10.3389/fpls.2017.01916

Zhong, S., Dekkers, J. C. M., Fernando, R. L., and Jannink, J.-L. (2009). Factors affecting accuracy from genomic selections in populations derived from multiple inbreed lines: a barley case study. *Genetics* 182, 355–364. doi: 10.1534/genetics.108.098277

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 24 February 2021
doi: 10.3389/fgene.2021.643761

Heterosis and Hybrid Crop Breeding: A Multidisciplinary Review

*Marlee R. Labroo, Anthony J. Studer and Jessica E. Rutkoski**

Department of Crop Sciences, University of Illinois at Urbana–Champaign, Urbana, IL, United States

Although hybrid crop varieties are among the most popular agricultural innovations, the rationale for hybrid crop breeding is sometimes misunderstood. Hybrid breeding is slower and more resource-intensive than inbred breeding, but it allows systematic improvement of a population by recurrent selection and exploitation of heterosis simultaneously. Inbred parental lines can identically reproduce both themselves and their F_1 progeny indefinitely, whereas outbred lines cannot, so uniform outbred lines must be bred indirectly through their inbred parents to harness heterosis. Heterosis is an expected consequence of whole-genome non-additive effects at the population level over evolutionary time. Understanding heterosis from the perspective of molecular genetic mechanisms alone may be elusive, because heterosis is likely an emergent property of populations. Hybrid breeding is a process of recurrent population improvement to maximize hybrid performance. Hybrid breeding is not maximization of heterosis *per se*, nor testing random combinations of individuals to find an exceptional hybrid, nor using heterosis in place of population improvement. Though there are methods to harness heterosis other than hybrid breeding, such as use of open-pollinated varieties or clonal propagation, they are not currently suitable for all crops or production environments. The use of genomic selection can decrease cycle time and costs in hybrid breeding, particularly by rapidly establishing heterotic pools, reducing testcrossing, and limiting the loss of genetic variance. Open questions in optimal use of genomic selection in hybrid crop breeding programs remain, such as how to choose founders of heterotic pools, the importance of dominance effects in genomic prediction, the necessary frequency of updating the training set with phenotypic information, and how to maintain genetic variance and prevent fixation of deleterious alleles.

Keywords: heterosis, inbreeding depression, genomic selection, reciprocal recurrent genomic selection, dominance, autogamous

OPEN ACCESS

Edited by:
Diego Jarquin, University of Nebraska–Lincoln, United States

Reviewed by:
Jinliang Yang, University of Nebraska–Lincoln, United States
Tian Qing Zheng, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, China

***Correspondence:**
Jessica E. Rutkoski
jrut@illinois.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 December 2020*
Accepted: *08 February 2021*
Published: *24 February 2021*

Citation:
Labroo MR, Studer AJ and Rutkoski JE (2021) Heterosis and Hybrid Crop Breeding: A Multidisciplinary Review. Front. Genet. 12:643761. doi: 10.3389/fgene.2021.643761

INTRODUCTION

Hybrid crop varieties vastly outperform their inbred progenitors in economically important species ranging from maize (*Zea mays*) to oil palm (*Elaeis guineensis*; Duvick, 2005; Fu et al., 2014; Cros et al., 2015). However, hybrid breeding requires more time and resources than inbred breeding (Troyer and Wellin, 2009; Longin et al., 2014; Cros et al., 2018). The effectiveness of hybrid breeding can be improved by genomic selection, in which marker information partially replaces phenotypes in estimation of breeding values (Heffner et al., 2009). Genomic selection can shorten the breeding cycle, reduce the costs of phenotyping, and improve selection accuracies (Lorenz et al., 2011; Heslot et al., 2015; Zhao et al., 2015b; Schulthess et al., 2017; Kadam and Lorenz, 2018). Genomic selection also opens new opportunities to establish hybrid breeding programs

in crops which are widely cultivated as inbreds, such as wheat (*Triticum aestivum*; Zhao et al., 2015b). Here, we compare and contrast genomic selection with conventional selection in hybrid crop breeding. We summarize the quantitative genetic model of phenotype, and we synthesize quantitative, evolutionary, phenotypic, and molecular genetic perspectives to explain the bases of heterosis and its role in breeding hybrids. Then, we cover the fundamentals of genomic prediction and its uses in genomic selection at all stages of the hybrid breeding cycle, including selection strategies for long-term gain. In closing, we outline factors which influence the success of hybrid breeding programs relative to inbred breeding programs.

QUANTITATIVE GENETIC MODEL OF PHENOTYPE

To consider genomic selection for hybrid performance and heterosis, it is necessary to understand the statistical model of phenotype used in quantitative genetics. The observed performances of individuals in a population are their phenotypic values (Falconer and Mackay, 1996). The variance of individuals' phenotypic values is due to genetic and non-genetic variance components and their interactions (**Supplementary Table 1**; Eq. 1; Falconer and Mackay, 1996). If non-genetic variance were absent, then phenotypic variance would be equal to genetic variance. Detecting genetic variance does not require demonstrating molecular modes of gene action, and genetic effects are indirectly observed as differences in phenotypes (Falconer and Mackay, 1996). For example, if there are no differences in individuals' phenotypes and thus no phenotypic variance, then genetic effects and genetic variance are zero. Even though at the molecular genetic level cellular machinery dynamically generates and maintains identical phenotypes, these are not genetic effects or genetic variance in the quantitative genetic sense. Similarly, the amount of genetic variation, genetic diversity, or nucleotide diversity cannot be inferred from the magnitude of genetic variance even though genetic variation underlies genetic variance. If the most genetically diverse lines of a population are sampled and their phenotypes are identical, then genetic variance is nonetheless zero, assuming no non-genetic variance. If the phenotype is also measured in closely related lines but varies greatly, then genetic variance is large, even if the lines have nucleotide polymorphisms in just one gene.

Total genetic variance can be further partitioned into additive, dominance, and epistatic variance (**Supplementary Table 1**; Eq. 2; Falconer and Mackay, 1996). Intuitively, individuals share alleles to the degree that they are related (Falconer and Mackay, 1996; Fisher, 1918). Under the infinitesimal model, an impossibly large number of alleles additively affect quantitative trait phenotypes, so the proportion of shared alleles among relatives is expected to produce concomitant phenotypic resemblance (Fisher, 1918). The more that relatives phenotypically resemble each other in proportion to their degree of relatedness, the greater the proportion of phenotypic variance that can be explained by additive genetic variance, assuming zero non-genetic variance (Fisher, 1918). If dominance and epistatic variance is present, relatives may resemble each other more than expected by a strictly additive model (Lynch and Walsh, 1998).

Genetic variance is also viewed as statistical effects of alleles at individual loci in a population (Falconer and Mackay, 1996). Alleles can have additive and dominance effects on genetic value. At a given locus, the additive effect of an allele, a, is the average genetic value of all individuals which are homozygous for the allele (Falconer and Mackay, 1996). The dominance effect of the allele, d, is the average genetic value of all individuals which are heterozygous for the allele (Falconer and Mackay, 1996). Since epistasis requires multiple alleles, single alleles do not have epistatic effects.

At the population level, the average effect of substituting one allele for another at a given locus on the genetic mean of the population depends not only on the additive and dominance effects of the allele, but also its frequency (**Supplementary Table 1**, Eq. 3—5; Falconer and Mackay, 1996). The average effect of an allele, α, is its coefficient in regression of genetic value on the number of copies of the allele in each genotype at the locus (**Supplementary Table 1**; Eq. 6; Falconer and Mackay, 1996). If dominance occurs, then observed genetic values do not fall exactly on the regression line of genetic value on allele copy number per genotype (Falconer and Mackay, 1996). The deviation of the heterozygote genetic value from the regression line is the dominance deviation of the allele, δ (**Supplementary Table 1**; Eq. 6; Falconer and Mackay, 1996). If more than one locus affects phenotype, then epistatic interactions between and/or among allelic effects across loci can also contribute to genetic value (**Supplementary Table 1**; Eq. 7; Falconer and Mackay, 1996). The statistical effects of alleles can be used directly to calculate respective genetic variances, but realistically it is almost always unknown which alleles affect phenotype or which individuals carry which alleles (Falconer and Mackay, 1996). Therefore, in practice, genetic variances are estimated from resemblance among relatives, not *a priori* from allelic effects (Falconer and Mackay, 1996).

Statistical genetic average, dominance, and epistatic effects do not represent underlying biological modes of gene action in most experimental and breeding settings, and modes of gene action cannot be inferred from the relative contribution of each source of statistical genetic effects to the genetic value (Cordell, 2002; Crow, 2010; de los Campos et al., 2015; Huang and Mackay, 2016; Manfredi et al., 2017). By definition, biologically dominant or epistatic gene action is largely captured by statistical average effects, because statistical dominance and epistasis are residual deviations from average effects (Cheverud and Routman, 1995). Average, dominance, and epistatic effects refer to their statistical formulations throughout this review unless specified as biological.

QUANTITATIVE GENETIC MODEL OF HETEROSIS

A rationale for hybrid breeding is the systematic exploitation of heterosis (Schulthess et al., 2017). Mid-parent heterosis is

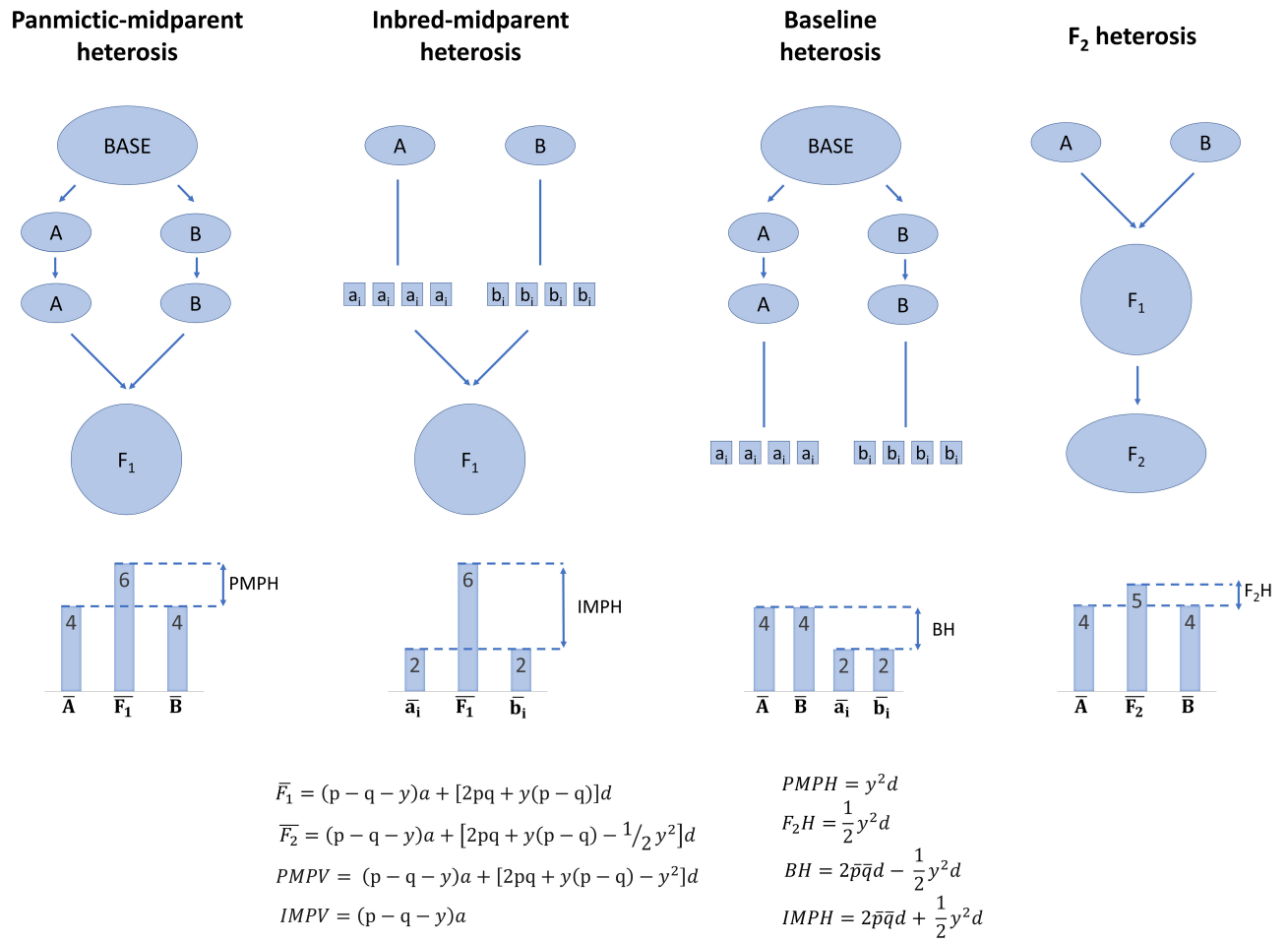

FIGURE 1 | Illustration of the partitions of population-level heterosis by Lamkey and Edwards (1999). Arrows indicate random mating or random crossing, and lines indicate selfing to homozygosity. Note that only dominance, additive × dominance, and dominance × dominance effects can contribute to baseline heterosis, but dominance, additive × dominance, dominance × dominance, and additive × additive effects can contribute to panmictic-midparent heterosis, inbred-midparent heterosis, and F_2 heterosis. Equations are further described in **Supplementary Table 1**.

the difference between a progeny genetic value and its mid-parent value, or the average of its parents' genetic values (**Supplementary Table 1**.; Eq. 8; Lynch and Walsh, 1998). Under the additive model, the genetic value of a progeny is expected to be equal to the average genetic value of its parents. Thus, mid-parent heterosis results from dominance and epistatic deviation. However, mid-parent heterosis of a single cross is neither a measure of dominance or epistatic effects nor a measure of heterosis in a population.

It is important to define heterosis further at the population level, because (a) heterosis emerges at the population level, even if it partially can be observed in single crosses, and (b) breeding involves populations rather than individuals alone (**Figure 1**; Lamkey and Edwards, 1999). In a group of individuals which can potentially intermate, such as a species, random mating may not occur among all individuals. Non-random mating of individuals—or any factor which leads to Hardy–Weinberg disequilibrium, such as migration—can cause distinct subpopulations form within the overall population, termed population structure. Within subpopulations, mating is random and Hardy–Weinberg equilibrium is reached, but among subpopulations mating is non-random. The subpopulations are inbred relative to the population that would result if random mating had occurred among all individuals in the overall population, and allele frequencies may come to differ among subpopulations.

What results if two of these subpopulations are randomly mated to each other? The mean genetic value of their F_1 may differ from the mean of the average genetic values within each subpopulation, and this difference is termed panmictic-midparent heterosis (**Supplementary Table 1**; Eq. 9—11; Lamkey and Edwards, 1999). Panmictic-midparent heterosis is thought to result from (a) dominance, as allele frequencies differ between the parent populations, and dominant genotypes that do not occur in the parents are observed in their F_1, and/or (b) additive × additive epistasis, as new interactions among alleles are possible in the F_1 compared to the parents. The portion of panmictic-midparent heterosis due to dominance, if

present, can be thought of as recovery from inbreeding due to population structure, because subpopulations are by definition inbred relative to a population in which structure had never occurred. Although this base population in which structure never occurred is hypothetical and cannot be observed, it is possible to form an analogous population in Hardy–Weinberg equilibrium by randomly intermating the two subpopulations to form an F_1, then randomly mating the F_1 to form an F_2. Heterosis in the F_2, or F_2 heterosis, is reduced by half compared to the panmictic-midparent heterosis in the F_1 (**Supplementary Table 1**; Eq. 12—13; Lamkey and Edwards, 1999).

Panmictic-midparent heterosis can be positive or negative. If panmictic-midparent heterosis is negative, it is sometimes referred to as outbreeding depression (Waser and Price, 1994; Lynch and Walsh, 1998; Grindeland, 2008; Oakley et al., 2015). Outbreeding depression is thought to primarily result not from dominance, but rather from loss of favorable additive × additive epistases as co-selected, genomically compatible combinations of alleles are separated in the F_1 of two random-mating populations (Dobzhansky, 1941; Welch, 2004). These losses of favorable biological epistases are termed Dobzhansky–Muller incompatibilities (Dobzhansky, 1941).

What results if, within a subpopulation, inbreeding occurs rather than random mating? Inbred lines are often observed to have a lower mean genetic value than the mean of the less inbred subpopulation, a phenomenon referred to as inbreeding depression (Charlesworth and Charlesworth, 1987; Falconer and Mackay, 1996; Charlesworth and Willis, 2009). Inbreeding depression is thought to result biologically from (a) deleterious recessive alleles driven to homozygosity, (b) homozygosity at overdominant loci at which the heterozygous state outperforms either homozygote, and/or (c) a loss of favorable epistatic interactions between heterozygous genotypes (Davenport, 1908; East, 1908; Shull, 1908; Falconer and Mackay, 1996). If it were possible to randomly mate the inbred lines without selection to form an F_1, the original subpopulation would be reconstituted and its mean restored to its original state if inbreeding depression were due to dominance and/or epistasis (Falconer and Mackay, 1996). Interestingly, there is some evidence of inbreeding depression due to epigenetic changes which may not be reversible by random mating, termed hybrid decay or hybrid dysgenesis (de la Luz Gutiérrez-Nava et al., 1998; Xue et al., 2019). Though hybrid decay is not thought to be a universal cause of inbreeding depression and has not prevented production of hybrids from inbreds in commercial programs, it is unknown how widespread this occurrence is.

If two subpopulations are again considered, an F_1 of the subpopulations can be produced in two ways: (a) the randomly mating subpopulations can be randomly intermated, or (b) each subpopulation can first be selfed to produce fully inbred lines, and the fully inbred lines can be randomly intermated (Lamkey and Edwards, 1999). The average genetic value of the F_1 resulting from either of these processes is equal (**Supplementary Table 1**; Eq. 9; Lamkey and Edwards, 1999). Some of the mean genetic value of the F_1 is due to baseline heterosis, or the restoration of what was lost due to inbreeding depression during selfing of both the parent subpopulations (**Supplementary Table 1**; Eq. 14; Lamkey and Edwards, 1999). However, the panmictic-midparent heterosis that results from crossing two randomly mating subpopulations also contributes to the genetic value of these F_1. Therefore, inbred-midparent heterosis is defined as the sum of baseline heterosis and panmictic-midparent heterosis, which is equivalent to the difference of the mean genetic value of the F_1 and the mean genetic value of all the inbred parents derived from both populations (**Supplementary Table 1**; Eq. 9, Eq. 15—16; Lamkey and Edwards, 1999). The key reason to partition heterosis into panmictic-midparent, F_2, baseline, and inbred-midparent heterosis is that it allows definition of average heterosis and inbreeding depression at the population level. Furthermore, it provides a framework to contrast heterosis that results from crossing two random-mating subpopulations and heterosis that results from crossing inbred lines that result from the random-mating subpopulations.

Heterosis is often described as the "opposite" or converse of inbreeding depression. However, of the partitions of heterosis described here, only baseline heterosis is strictly the opposite of inbreeding depression. Panmictic-midparent heterosis (and therefore inbred-midparent heterosis) can arise totally from epistatic effects without dominance, whereas inbreeding depression cannot (Falconer and Mackay, 1996; Lamkey and Edwards, 1999; Chen, 2013). Heterosis due to epistasis can only result from additive × additive epistasis, whereas inbreeding depression can result from both additive × dominance and dominance × dominance epistasis (Lynch, 1991; Lynch and Walsh, 1998).

EVOLUTIONARY AND MOLECULAR GENETIC BASES OF HETEROSIS

From the evolutionary perspective, heterosis in quantitative genetics ultimately rests on assumptions of biological dominance and biological epistasis, even though the additive model captures most of the effects of biological dominance and epistasis (Huang and Mackay, 2016). For biological dominance to affect heterosis, dominant alleles should have directional effects on fitness (Lynch and Walsh, 1998). As biologically dominant mutations arise in a population, they affect phenotype regardless of zygosity and are exposed to selection (Falconer and Mackay, 1996). Recessive mutations are only exposed to selection on phenotype in their homozygous state and can propagate in populations as they are not selected as heterozygotes (Falconer and Mackay, 1996). Therefore, deleterious dominant alleles are more likely to be eliminated from populations by selection than deleterious recessives (Falconer and Mackay, 1996). Over evolutionary time, it is expected that biologically dominant alleles tend to be favorable, and deleterious alleles tend to be recessive (Falconer and Mackay, 1996). If dominant alleles have directional effects on fitness, the effect is then often positive (i.e., the sign of dominance occurs in the same direction as the measure of fitness). In maize, alleles identified as likely deleterious via genomic evolutionary rate profiling were found more likely to be recessive (Yang et al., 2017). The likelihood of purging recessive deleterious alleles is reduced as effective population

size increases, and deleterious alleles may be shielded from selection in genomic regions with low recombination rates, such as the centromere, regardless of dominance (Barrett and Charlesworth, 1991; Rodgers-Melnick et al., 2015; Yang et al., 2017). Evolutionary mechanisms besides directional selection have also been proposed to explain the emergence of dominance, such as stabilizing selection (Manna et al., 2011).

Heterosis can also result from overdominance, a type of biological dominance in which heterozygotes have more extreme phenotypes than both homozygotes, and alleles persist in populations at intermediate frequencies (Crow, 1999). One overdominant locus alone is sufficient to cause heterosis (Falconer and Mackay, 1996; Krieger et al., 2010). However, detection of overdominance is complicated because it requires inbred parents to be identical at all loci except the locus of interest. If they are not, then parents can carry biologically dominant alleles of opposite effects on fitness linked in repulsion, and pseudooverdominance results: the loci are never observed in their uncoupled state, and they appear as one overdominant locus. In absence of linkage, pseudooverdominance would not exist.

Finally, biological epistasis may contribute to heterosis as interactions of multiple loci contribute to fitness. Ample evidence of biological epistasis is available; for example, genes encoding transcription factor proteins may physically bind to DNA sequence motifs to activate or repress other genes which affect phenotype, among other mechanisms (Phillips, 2008; Lehner, 2011; Burdo et al., 2014). However, detecting all types of both statistical and biological epistasis in regular experimental samples is often not feasible because the number of combinations of alleles is much larger than the number of individual genotypes in a population (Wei et al., 2014). Epistasis also cannot be detected in a population if the experimental sample is not segregating for both interacting genes (Stitzer and Ross-Ibarra, 2018).

It is possible that heterosis can be explained fully by biologically additive, dominant, and epistatic gene action and that no single gene, class of genes, or physiological phenomenon causes heterosis (Birchler et al., 2010; Fiévet et al., 2018). If so, searching for the genetic basis of heterosis would lead to the genetic basis of the specific trait in question in a particular experimental sample, and heterosis would be conferred by biological dominance, overdominance, or epistasis of those genes which controlled the trait (Fiévet et al., 2018). For example, consider inquiry into the genetic basis of heterosis for grain yield in maize and biomass yield in sorghum (*Sorghum bicolor*). By the explanation of heterosis above, maize grain yield and sorghum biomass yield could be controlled by completely different genes and classes of genes, and dominant, overdominant, or epistatic action of the genes involved would lead to heterosis. If more individuals were sampled, which presented more combinations of genes and/or more genetic variants, then the genetic basis of observed heterosis could change.

It has been further hypothesized that actions of particular classes of genes or physiological effects of genes cause heterosis universally across traits and species (Birchler et al., 2010; Fiévet et al., 2018). These proposed unifying mechanisms include organellar complementation, circadian rhythm changes, changes in hormone expression, genome-wide changes in chromatin state and/or changes in small RNA expression, dosage effects, regulatory incompatibility, parent-specific gene expression, and changes in signaling in response to heterozygosity (Auger et al., 2005; Reif et al., 2005; Lippman and Zamir, 2007; Kaeppler, 2012; Chen and Birchler, 2013; Bar-Zvi et al., 2017; Herbst et al., 2017; Li et al., 2020). It is challenging to detangle whether each of these actions of gene classes and physiological effects are themselves causes of heterosis, or instead the effect of a true unobserved cause of heterosis. None has been demonstrated to universally explain heterosis, but some have been demonstrated to be associated with heterosis in some cases and have been incorporated into predictive models with varying effects on prediction accuracy (Kaeppler, 2012; Westhues et al., 2017; Schrag et al., 2018; Seifert et al., 2018). At the transcriptome level, hybrids generally display transcript levels near their mid-parent value, with some exceptions (Swanson-Wagner et al., 2006; Hochholdinger and Hoecker, 2007; Springer and Stupar, 2007). At the proteome level, hybrids generally show protein levels which deviate from the mid-parent, particularly in functional classes related to central metabolism and stress responses (Marcon et al., 2010). Efforts to map heterosis for various traits generally do not reveal loci for which the association holds universally across genotypes, even for single traits within a species (Huang et al., 2016; Liu et al., 2020). If a particular universal mechanism of heterosis were ultimately revealed, the genes involved would still have biologically additive, dominant, or epistatic gene action. The genes may not be identical at the sequence level, but it would be expected that the mechanism would be common to all cases of heterosis.

PHENOTYPIC BASES OF HETEROSIS

In hybrid individuals, not all traits are necessarily heterotic (Kaeppler, 2011). Nor is there correlation in levels of heterosis for different traits (Longin et al., 2013; Huang et al., 2015). For example, a hybrid individual might show heterosis in yield and height, but not root angle, and the amount of heterosis for yield and height may differ. The sign of heterosis can vary among traits; inter-subspecific hybrids of *indica* and *japonica* rice (*Oryza sativa*) show increased vigor, but reduced fertility, as do interspecific hybrids of donkeys (*Equus asinus*) and horses (*Equus caballus*; Troyer, 2006; Fu et al., 2014). The degree of heterosis can also depend on environment. Maize hybrids usually show more heterosis in stressful than non-stressful environments, even as overall performance is decreased (Duvick et al., 2004). The lack of consistent levels of heterosis across traits may indicate that heterosis cannot be explained by a unifying, systems-wide mechanism—the reasoning being that all traits would then be affected equally.

Heterosis is also found in complex traits that are a function of multiple component traits, even if the component traits can be fully explained by an additive genetic model. If component traits diverge phenotypically in parents, then heterosis in the complex trait is often detected in progeny even as the component traits remain near the mid-parent (Powers, 1944; Williams, 1959;

Grafius, 1961; Coyne, 1965; Melchinger et al., 1994; Dan et al., 2015; Fiévet et al., 2018). For example, in the heterotic pools of oil palm, one pool has a few heavy fruit bunches, and the other has many light fruit bunches (Cros et al., 2015). Their hybrids exhibit substantial heterosis (25%) for fruit production—the product of bunch number and bunch weight—but the hybrid values for bunch number and bunch weight remain near the mid-parent. Notably from the genetics perspective, biological dominance is not required to explain heterosis in multiplicative complex traits (Schnell and Cockerham, 1992; Cros et al., 2015). In this example, it is possible that all of the heterosis in fruit production of oil palms can be fully explained by biologically additive gene action in bunch number and bunch weight (in which case hybrid breeding would not be the optimal strategy to increase fruit production), but in practice the true genetic bases of these traits are unknown.

At the biochemical level, complex phenotypes are a function of multiple component metabolites over time (Fiévet et al., 2018). Metabolite levels or concentrations are themselves a complex phenotype, because they are the product of enzyme amounts and activities within pathways, as well as flux (i.e., rate of turnover) through the pathway (Marshall-Colón et al., 2010; Fiévet et al., 2018). Heterosis can emerge because enzyme activities often affect metabolic flux non-linearly—i.e., halving the activity or concentration of a given enzyme does not necessarily halve metabolic flux (Fiévet et al., 2018; Govindaraju, 2019; Vacher and Small, 2019). The non-linear relationships of enzyme activity and metabolic flux has been proposed as the molecular basis of dominance (Kacser and Burns, 1981). Even if hybrids have enzyme concentrations near the mid-parent, as would be expected under additive inheritance, whether flux or the product metabolite is also at the mid-parent depends on the biochemistry of the pathway (Vacher and Small, 2019). For example, the product metabolite concentration in hybrids may deviate from the mid-parent as enzymes with activities at the mid-parent interact along a pathway and change the flux, or as a rate-limiting step of the pathway is saturated at lower levels than the mid-parent enzyme activity and further increases in enzyme activity do not affect flux (Fiévet et al., 2018). A key conclusion, then, is that non-additive phenotypes such as metabolite concentrations may arise from component additive phenotypes such as enzyme concentrations or activities. Since metabolites are component traits of even more integrated traits, like grain yield, non-additivity in metabolite concentrations can reverberate across levels of phenotype and can lead to heterosis in the integrated trait (Fiévet et al., 2018). Whether heterosis is detected can depend on the choice of phenotype. The metabolome is a phenotype, and using metabolomics data as component traits in multi-trait prediction then has instant appeal, despite current limits in metabolomics on throughput, cost, and the number of metabolites which can be sampled.

ALTERNATIVE DEFINITIONS OF HETEROSIS

There are several alternative definitions of heterosis which are not equivalent to mid-parent heterosis and do not have a well-defined genetic interpretation. Better parent heterosis (heterobeltiosis) and commercial heterosis, in which either the phenotypic value of the better-performing parent or a commercial check, respectively, is taken from the progeny phenotypic value, may be useful measures for varietal development but have no immediate relevance to genetic improvement of a population by selection, except perhaps to define selection targets (Flint-Garcia et al., 2009; Schnable and Springer, 2013). Better-parent and commercial heterosis might be more informatively described as better-parent and commercial relative performance to avoid equating these measurements with mid-parent heterosis.

Heterosis has also been restricted to describing only increases in progeny vigor relative to parents, i.e., positive heterosis (Shull, 1948). Negative heterosis is observed, as in the progeny of outbreeding depressed parents (Lynch and Walsh, 1998). However, the sign of heterosis can also be a simple artifact of the investigator's choice of phenotypic measurement (Falconer and Mackay, 1996). For example, positive heterosis for days to flowering is equivalent to negative heterosis for speed of development—a plant which flowers later would have a more positive value for days to flowering, but it would have a less positive value for speed of development since it matures more slowly (Falconer and Mackay, 1996). Therefore, a progeny that flowers later than its mid-parent would show positive heterosis for days to flowering, but negative heterosis for speed of development even though the character measured (when the progeny flowers) is identical. Another common example is that severity of disease can also be viewed as plant health status, and the investigator chooses whether a more positive number represents more or less severe disease symptoms. In the case that a progeny is more resistant to disease that its mid-parent, it will show positive heterosis if less severe disease is measured as a more positive value but negative heterosis if less severe disease is measured as a less positive value.

Finally, heterosis has been conceptualized as a systems-wide phenomenon is which "the increased vigor, size, fruitfulness, speed of development, resistance to disease and to insect pests, or to climactic rigors of any kind" is observed (Shull, 1952). It is perhaps this perspective of heterosis which has fueled the search for a unifying theory of heterosis as well as investigation into its functional genomic causes (Birchler et al., 2003, 2010). Understanding biological bases of heterosis is valuable, but further investigation is needed to use biological insights into heterosis in hybrid crop breeding programs (Ramstein et al., 2019).

GENOMIC PREDICTION

As parents, individuals transmit neither their phenotype nor their full genotype to their offspring. The allele, or more broadly, the gamete is the unit of inheritance. Only the additive portion of genetic value is heritable in the narrow sense if mating is random, because it does not depend on intra- or inter-locus combinations of loci which are potentially disrupted upon mating (Falconer and Mackay, 1996). If mating is random, additive genetic value is all that is maximizable or "breedable" cyclically over generations,

and an individual's additive genetic value is its breeding value (Falconer and Mackay, 1996; Huang and Mackay, 2016).

Despite its name, the concept of breeding value was not developed specifically for the purpose of breeding, but rather to explain the inheritance of quantitative traits: because Mendel discovered inheritance in traits which had discrete classes, it was initially unknown whether continuous, quantitative traits were also controlled by genes that could be transmitted from parent to progeny (Bernardo, 2020). Fisher (1918) not only conceptualized that quantitative traits could be the effect of many genes, but also connected the partial inheritance of parental alleles to observed patterns of resemblances among relatives (Bernardo, 2020). In applied breeding programs, some of the assumptions that define breeding value—such as random mating—are routinely unmet (Falconer, 1985; Bernardo, 2020). Recently, it has been suggested to move away from referring to estimates of transmissible variance as breeding values in applied plant breeding programs for clarity and because non-additive variance can be transmitted via cross selection (Bernardo, 2020; Werner et al., 2020). Here, we refer to breeding values even though at times the definition is not strictly met.

True breeding value cannot be measured or even observed in the individual alone, since true measure of breeding value requires errorless observations of every possible progeny resulting from the individual mated to every possible member of the population to which it belongs. Therefore, breeding values are estimated. The estimation of breeding values was first accomplished by progeny testing. With random mating, the average performance of an individual's progeny is an estimate of its breeding value (**Supplementary Table 1**; Eq. 17). Many mating schemes were developed to more accurately estimate breeding values by use of more types of relatives (Hallauer et al., 2010). However, best linear unbiased prediction (BLUP) was developed to estimate breeding values without the need for mating designs, as pedigree-based variance-covariance relationship matrices describe the resemblance between relatives in a linear mixed model (**Supplementary Table 1**; Eq. 18; Henderson, 1975). Assuming no fixed effects, BLUP of breeding value can be thought of as a linear combination of observed phenotypes weighted by the degree of their relationship with the individual for which breeding value is predicted. The pedigree-based relationship matrix is often referred to as the numerator relationship matrix, A, and pedigree-based BLUP is sometimes called ABLUP (Bernardo, 2002; Gianola et al., 2018). Interestingly, BLUP was slow to gain traction in plant breeding, but quickly became popular in animal breeding due to the standing practical impossibility of replicating animal genotypes (Piepho et al., 2008). It was perhaps here that the two fields decoupled in their study of genomic prediction and selection, and the benefits of cross-disciplinary synchronization of methods are recognized in both fields (Schön and Simianer, 2015; Hickey et al., 2017).

Well in advance of the sequencing technologies that would make markers cheaper, less biased, and more representative of the genome, the framework for genomic prediction of EBV using molecular markers was developed. Bernardo (1994; 1996) used genome-wide markers to estimate breeding values from kinship rather than pedigree in the first instance of genomic prediction. Whittaker et al. (2000) addressed the problem of selection of marker subsets for linear regression in marker-assisted selection (MAS) by ridge regression, which is a regularization method that shrinks normalized effects for all markers equally toward an assumed mean of zero by an optimized parameter, λ (**Supplementary Table 1**; Eq. 19). This was a crucial advance for the use of genomic markers in selection, because markers generally outnumber phenotypes and cause the "large p, small n" problem: linear regression by OLS is not possible if predictors (p) outnumber responses (n), and subsampling is usually suboptimal (Whittaker et al., 2000). Ridge regression, like other regularization methods, addresses the problem of selecting predictors by shrinking their coefficients instead of subsampling. Regularization also reduces model overfitting, in which models capture noise (i.e., residual error) as well as signal (i.e., effects of predictors). Both model overfitting and poor choice of predictors reduce prediction accuracies. Meuwissen et al. (2001) realized that if markers in linkage with every quantitative trait locus (QTL) affecting a trait were to become available, then additive effects per marker (estimated by ridge regression or other methods) could be summed to calculate individuals' genomic estimated breeding values (GEBVs). Use of GEBVs or any other value estimated using genome-wide information for selection is referred to as genomic selection.

Since 2001, tens of methods for genomic prediction of breeding values, as well as the genetic values of lines used for production rather than breeding, have been developed (Gianola et al., 2006, 2009; Legarra et al., 2009; Habier et al., 2011; Momen et al., 2018; Wang et al., 2018; Howard et al., 2019; Kadam and Lorenz, 2019). These methods include both frequentist and Bayesian, as well as parametric and non-parametric, methods (Gianola et al., 2018). The parametric method of Whittaker et al. (2000), ridge regression of marker effects, is called RR-BLUP and assumes marker effects are drawn from a normal distribution. Hayes et al. (2009) later showed that estimation of breeding values by RR-BLUP is equivalent to estimation by genomic BLUP (GBLUP), in which markers are used to compute a genomic relationship matrix. The genomic relationship matrix (often denoted G) replaces the pedigree relationship matrix (A) to calculate BLUPs of GEBVs (VanRaden, 2008). GBLUP is generally more accurate than ABLUP, because realized genetic relationships deviate from pedigree expectations following Mendelian sampling, selection, and other events (VanRaden, 2008). RR-BLUP and GBLUP are widely used for genomic prediction because they are relatively straightforward to interpret, often more computationally efficient, and often as accurate as other methods with more realistic assumptions (Zhao et al., 2015b; Howard et al., 2019). For reviews of genomic prediction methods, see Gianola et al. (2018) and Howard et al. (2019).

In hybrid breeding, genomic prediction can be used to (a) predict the combining abilities of inbred lines, and (b) predict the performance of new hybrid genotypes (Bernardo and Yu, 2007; Technow et al., 2012). To predict the combining abilities of inbred lines, phenotypes of their hybrid progeny are used to estimate inbred combining abilities, then the combining abilities

of the inbred lines are modeled as a function of their inbred genotypes (Bernardo and Yu, 2007). To predict the performance of new hybrid genotypes, hybrid phenotypes are modeled as a function of hybrid genotypes (Technow et al., 2012). However, hybrid genotypes are usually not sequenced directly, and are inferred instead by genotyping their inbred parents, which reduces the total number of individuals for genotyping. Even though within hybrid genotypes a given allele may be specific to a particular population, modeling population-specific effects of alleles has not been shown to greatly increase accuracy in predicting hybrid performance (Technow et al., 2012).

Prediction accuracy is an important determinant of whether genomic prediction will lead to effective selection across environments, years, and genotypes. Factors which influence accuracy of GEBVs include choice of statistical model, trait heritability, precision in geno- and phenotyping, size the of the training set, and relatedness/common LD structure of the training and testing set (Heslot et al., 2015; Rutkoski et al., 2017). Modeling non-additive effects is an active area of research with particular relevance to prediction of GEBV or performance (Vitezica et al., 2017; Varona et al., 2018b; Voss-Fels et al., 2019). Dominance deviations are by definition zero in genetic values of homozygous inbred lines; only additive and epistatic additive effects are non-zero. In non-inbred individuals, all non-additive effects contribute to genetic value. Product development, in contrast to population improvement, is concerned with total genetic value, which includes non-additive effects.

Though modeling non-additive effects would be expected, then, to improve prediction accuracies, a reminder is warranted: modeling non-additive genetic effects will only improve prediction accuracies if non-additive genetic effects exist for the traits of interest and non-additive genetic effects can be estimated accurately in the populations of interest (Hill et al., 2008). In light of these considerations, it is perhaps unsurprising that in practice classical models which fit non-additive effects rarely outperform accuracies of additive models (Varona et al., 2018a; Werner et al., 2018). Interestingly, though, if dominance effects are fit in absence of underlying dominance, Duenk et al. (2017) observed no change in accuracy of estimating additive effects. In fact, accuracy of estimation of additive effects was always improved or unchanged by models which incorporated dominance, even in small sample sizes and/or in cases that genetic variance explained low proportions of phenotypic variance (Duenk et al., 2017). Though no similar study has been conducted for epistasis to our knowledge, there appears to be no penalty to fitting dominance effects. In crossbred (hybrid) and pure-line animals, incorporating positive directional dominance effects and inbreeding depression effects (which are posited to underlie heterosis) sometimes improves prediction accuracies relative to assuming dominance effects centered at zero or ignoring inbreeding (Xiang et al., 2016; Varona et al., 2018a; Christensen et al., 2019). Inclusion of non-additive effects can also improve choice of parents for crossing by estimates of their progeny genetic value (Aliloo et al., 2017; Werner et al., 2020).

Multivariate genomic prediction methods are promising for improving prediction accuracy when traits under selection with low heritability are genetically correlated with traits with high heritability (Jia and Jannink, 2012; Neyhart et al., 2017; Okeke et al., 2017; Sun et al., 2017; Wang et al., 2017; Fernandes et al., 2018; Watson et al., 2019). Because heterosis in complex traits can sometimes be explained by component traits which are negatively complementary in the parents, multivariate genomic prediction could potentially improve predictions of hybrid performance and EBVs if such component traits are included. Hybrid production also faces a constraint on the performance of the inbred parents, in that inbred parents must have good *per se* performance and specific male and female morphotypes for hybrid seed production (Hallauer et al., 2010). Generally, inbred parents are selected for these traits separately from their selection as hybrid parents (Hallauer et al., 2010). Treating inbred and hybrid performance as different but genetically correlated traits in multivariate genomic selection may improve selection accuracy for hybrid performance, but this has not been reported to date.

GENOMIC SELECTION IN HYBRID BREEDING

Breeding for hybrid performance can benefit from the incorporation of genomic selection, and in a few cases genomic prediction could be used to develop new breeding strategies (Xu et al., 2017). Hybrid breeding primarily involves inter-population improvement, in which recurrent selection of individuals within populations is effected between populations by selecting on individuals' performance as parents in between-population crosses (Hallauer et al., 2010). Unlike intra-population improvement, in which performance of crosses within populations is used to recurrently select individuals in the same population, inter-population recurrent improvement is not only of the populations themselves but also of the performance of their hybrid crosses in combination (Hallauer et al., 2010). Hybrid breeding can be considered to have three main modules: selecting founders of heterotic pools, breeding heterotic pools, and selecting parents of crosses for production pipelines (**Figure 2**).

Heterotic pools are distinct groups of lines which reliably produce heterosis upon crossing; the lines may or may not be related (Melchinger and Gumber, 1998). Breeding distinct heterotic pools is more effective in consistently producing high-performing hybrids than making random crosses, because heterotic pools are improved by recurrent selection for average line performance in hybrid crosses with the opposite heterotic pool, which is termed general combining ability (GCA; Sprague and Tatum, 1942; Reif et al., 2005). Hybrid performance is modeled as the sum of each parental GCA and the specific combining ability, or SCA, of the parent pair (**Supplementary Table 1**; Eq. 20; Griffing, 1956b). GCA corresponds to additive effects, whereas SCA corresponds to dominance effects (Griffing, 1956a). The process of breeding heterotic pools increases the ratio of GCA to SCA effects over time, so the parents' performance in crosses becomes more heritable in the narrow sense (Schulthess et al., 2017). In addition, distinguishing heterotic pools addresses the practical need for lines to have specific traits for use

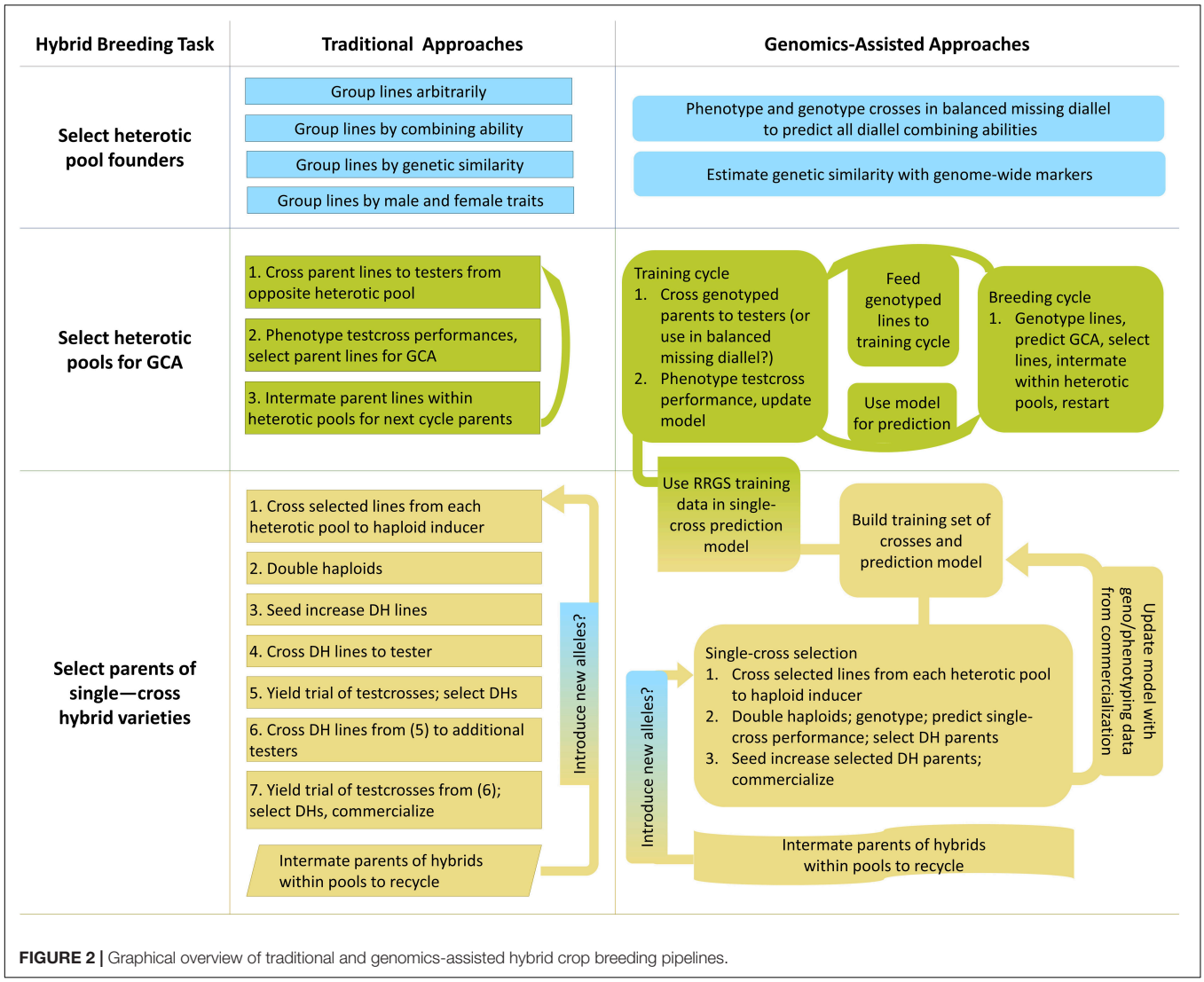

FIGURE 2 | Graphical overview of traditional and genomics-assisted hybrid crop breeding pipelines.

as males or females, as male and female traits such as high pollen production or male sterility can be pool-specific (Zhao et al., 2015b).

Heterotic pools are developed by choosing founders, then recurrently improving the pools for combining ability. Historically, hybrid breeding was first systematically conducted in maize in North America, and maize breeding programs are the longest-running among hybrid crops (Shull, 1908). Though it is a common misconception that the founders of the first heterotic pools of maize were chosen for their origin in geographically and genetically distinct groups, in fact the archetypal Reid-Lancaster heterotic pattern was developed empirically by trial-and-error in crossing (Melchinger and Gumber, 1998; Tracy and Chandler, 2006). Later, successful commercial maize heterotic pools arose upon separating of lines into groups for use as males or females; the initial pools used have been posited to have shared around half of their genetic background (Tracy and Chandler, 2006). Observed genetic divergence between the first North American maize heterotic pools developed in response to selection and drift during breeding, rather than by selection of divergent founders (Duvick et al., 2004; van Heerwaarden et al., 2012).

Heterotic pools have not been widely established for major crops such as wheat and rice, and there is interest in methods to choose founders of heterotic pools (Wang et al., 2015; Zhao et al., 2015a). Based on evidence from maize, some authors suggest that any split of available germplasm will allow development of heterotic pools by breeding, and founder grouping is relatively unimportant (Cress, 1966; Lee and Tracy, 2009). Others suggest systematic approaches. To choose founders of heterotic pools, Melchinger and Gumber (1998) proposed to form genetically similar groups of individuals, then cross a manageable number of representatives of each divergent genetic subgroup and test their progeny performance in replicated field trials. Founders can then be chosen for high *per se* performance, high average progeny performance and progeny genetic variance, and—as applicable—suitability for use as males or females (Melchinger and Gumber, 1998; Melchinger, 1999). Use of genetically diverged founders increases the ratio of GCA to SCA, and

heterosis due to dominance is expected to be positively correlated with increasing genetic distance of parents (Falconer and Mackay, 1996; Reif et al., 2007). In practice, positive heterosis is often observed with increasing genetic distance until a point, at which outbreeding depression prevails and heterosis is negative (East, 1936). A simulation study of heterotic pool formation in an autogamous crop compared randomly splitting the founder population, splitting the founder population by genetic distance, and optimizing the founder split by performance of their F_1 hybrids, and found no differences in future hybrid performance among the formation strategies, suggesting that the initial split can be arbitrary (Cowling et al., 2020). However, no population structure in the founder population was assumed, and testing the formation strategies given a structured founder population would be of interest. Even though existing heterotic pools of maize, for example, were not established from genetically distinct lines, testing the optimal strategy to form heterotic pools in an allogamous species would be interesting as well (Duvick et al., 2004).

Genomics-assisted approaches to choose founders have been proposed (Zhao et al., 2015a; Boeven et al., 2016). One approach is to simply extend the aforementioned method by using genome-wide markers to identify genetic subgroups (Boeven et al., 2016). Another approach to select founders of heterotic pools is to use a training set of observed hybrid crosses of founders to predict performance of unobserved crosses (Zhao et al., 2015a). Then, groups of lines which are heterotic in combination can be identified algorithmically for selection (Zhao et al., 2015a). The long-term potential of the groups of lines can be further assessed by simulation for their genetic representativeness of the base population, usefulness (in terms of the initial population mean and expected response to selection), and long-term selection limits (Zhao et al., 2015a). Though this approach has not been empirically validated, it has been initialized in rice and wheat (Zhao et al., 2015a; Beukert et al., 2017). In the wheat population surveyed, only sixteen of the 135 individuals surveyed were needed to maximize usefulness and the long-term selection limit (Zhao et al., 2015a). Therefore, if effective, this approach could dramatically reduce resources needed to screen potential founders of heterotic pools. It would be interesting to test whether this particular method would discern the founders of North American maize heterotic pools as optimal or near-optimal.

Once founders have been chosen, heterotic pools must be developed by breeding. Heterotic pools can be recurrently improved for their ability to combine into hybrids with high performance by reciprocal recurrent selection (RRS; Comstock et al., 1949). In the first generation, lines from each heterotic pool are at once selfed and crossed to the opposite heterotic pool (Comstock et al., 1949). Rather than making every line-by-line cross, one or more random testers are chosen to represent each heterotic pool and used for crossing to all lines of the opposite heterotic pool, hereafter referred to as testcrossing (Comstock et al., 1949). In the next season, the testcrosses are grown and the testcross phenotypes are used to determine GCA of their parents (Comstock et al., 1949). Parents are selected for their GCA, and in the final season of the cycle, the selected parents are grown from saved seed and randomly intermated within heterotic pools (Comstock et al., 1949). The cycle begins again, with no need to inbreed the parents (Comstock et al., 1949). Overall, RRS can be thought of as a special case of standard phenotypic selection, where the phenotype in question is combining ability, and measuring combining ability requires progeny testing.

Reciprocal recurrent selection, then, is an ideal candidate for genomic selection. Phenotyping GCA is expensive and time-consuming. If GCA could be predicted in the first generation of the RRS cycle, then the selected lines could be intermated immediately, reducing cycle time by two-thirds. Reciprocal recurrent genomic selection (RRGS) has been studied by simulation in oil palm (Ibáñz-Escriche et al., 2009; Kinghorn et al., 2010; Cros et al., 2015). Cros et al. (2015) investigated the effects of training set composition, frequency of model calibration by progeny test, and number of selection candidates on annual selection response. If number of available selection candidates was controlled, RRGS showed a 48% advantage in annual selection response over RRS because genomic predictions replaced phenotyping by progeny testing. If RRGS was assumed to permit evaluation of twice as many candidates for selection relative to RRS as progeny testing was reduced, then the advantage increased to 72%. Interestingly, Cros et al. (2015) tested whether including hybrid geno- and phenotypes in the training set improved accuracy more than including the parents alone and found accuracy to be sensitive to the frequency of model calibration and the number of hybrids included. They posited that optimal number of F_1 hybrid genotypes in the training set should increase with heterozygosity of the parents of the F_1 hybrids, because more heterozygous parents may produce more within-cross variance than less heterozygous parents (Cros et al., 2015). In a follow-up study, Cros et al. (2018) also investigated whether prediction accuracies from a training set of genotypes from either only the previous breeding cycle, or both the previous two breeding cycles, was superior in RRGS. They found that training on two previous breeding cycles was superior because of both increased in prediction accuracy and slightly decreased loss of additive genetic variance (Cros et al., 2018).

A key consideration in both studies was that dominance was not simulated even though RRGS was used, so the mean genetic values of the hybrids were equal to the mean of their parents (Cros et al., 2015, 2018). Further simulations of RRGS with dominance would be valuable. If dominance were simulated, then the F_1 hybrids' mean genetic value would differ from the parents' mean genetic value. To use both parent and hybrid geno- and phenotypes in the same training model when directional dominance is present, it would likely be necessary to include a fixed effect for the average heterozygosity of each individuals' genotype following Xiang et al. (2016) and Vitezica et al. (2016), or alternately to estimate hybrid and parent BLUPs of phenotype separately following Liang et al. (2018), in order to accurately predict parental GCA or hybrid genetic value.

Reciprocal recurrent selection differs slightly from recurrent selection within populations in that breeding values depend on allele frequencies in both heterotic pools (Stuber and Cockerham, 1966). Two issues then arise. Rembe et al. (2019) note that in its current implementation, RRGS does not optimize frequencies of overdominant alleles (either positive or negative) and in some

cases of negative overdominance will fix unfavorable alleles. RRGS also cannot optimize frequencies of alleles which have a frequency of zero in either population, nor predict their effects, and maximum genetic potential cannot then be achieved (Cress, 1966; Kinghorn et al., 2010). Periodic introduction of new germplasm to refresh heterotic pools might overcome the latter issue if it is in fact significant, though care must be taken not to disrupt the heterotic pattern. Another interesting but unexplored possibility is to avoid unwanted fixation from the start of RRGS; methods developed to control long-term inbreeding under genomic selection might be adaptable for the latter purpose.

As heterotic pools are developed, the next module is initiated: parents of crosses for production pipelines are selected. First, individuals are selected from each heterotic pool (Lee and Tracy, 2009). In established commercial maize hybrid breeding programs, within-pool lines which have a past record of producing high-performing cross-pool hybrids are recycled by crossing, and it is their progeny which are selected (Mikel and Dudley, 2006). As of 2006, only seven inbred founder lines (though from four heterotic pools) were thought to be the origin of the commercial North American breeding pool (Mikel and Dudley, 2006).

Next, selected individuals from each pool are used to develop inbred lines by doubled haploid (DH) production or selfing. During inbreeding of lines from two-parent crosses or upon availability of DH lines, lines are selected for *per se* performance, often for traits such as disease resistance (Lee and Tracy, 2009; Kadam and Lorenz, 2018). Then, the selected lines are testcrossed, usually to a single tester, and selected by the performance of their hybrids in a few environments (Lee and Tracy, 2009). Only these selected lines are crossed again (after selfing to homozygosity if DH lines are not used) to multiple testers, and their hybrids are advanced to multi-environment trials (METs; Lee and Tracy, 2009). Parental inbred lines which produce outstanding hybrids can then be commercialized, and their hybrids may be used in production (Lee and Tracy, 2009).

Some authors have proposed to reduce or eliminate preliminary testing by use of genomic prediction (Lee and Tracy, 2009; Kadam et al., 2016). With sufficient prediction accuracies, genomic predictions of all possible two-parent, single crosses of a set of inbred lines (i.e., a diallel) could replace testcrossing, then crosses predicted to have outstanding performance could immediately be tested in METs (Hallauer et al., 2010; Kadam et al., 2016). Genomic prediction could save time and resources as well as retaining useful lines which happen to perform poorly with chosen testers (Kadam et al., 2016). The primary challenges in doing so are generating an adequate training set of crosses and predicting SCA; it remains difficult to predict the performance of hybrids for which neither parent is observed in the training set (Kadam et al., 2016; Kadam and Lorenz, 2019). Notably, the ideal training set to predict performance of single crosses that would be obtained from a diallel is thought to be not a set of testcrosses, but rather the North Carolina II (NCII) design, in which inbred parents are grouped into males and females, then crossed factorially across groups (Hallauer et al., 2010; Fristche-Neto et al., 2018). However, the number of crosses needed for training can be reduced from NCII by using various algorithms which rely on estimates of relationship (Fristche-Neto et al., 2018; Akdemir and Isidro-Sánchez, 2019; Guo et al., 2019). Inclusion of historical single cross information can also improve prediction accuracies, though in some studies this benefit was only realized if the crosses were from recent cycles, even within the same breeding program (Dias et al., 2019; Schrag et al., 2019). If the production of F_1 hybrid seed by cross-pollination is too expensive on a large scale with many hybrid combinations (as in self-pollinated crops), then $F_{1:2}$ individuals can be substituted into the training set with only modest reductions in prediction accuracies (Technow, 2019). If entirely eliminating testcrossing is perceived as too risky, selections from testcrossing can be supplemented with predicted exceptional single crosses (Kadam and Lorenz, 2018; Viana et al., 2018). Another cost-reducing alternative is to testcross a subset of several related lines, and predict combining abilities for their relatives (Windhausen et al., 2012). Once exceptional single crosses are identified, with or without testcrossing, their seed can be increased, advanced through preliminary and multi-environment yield trials, and eventually released as varieties for production (Kadam and Lorenz, 2018).

Other approaches to utilizing heterosis, besides by inbred development and testing, deserve consideration. The cost and time required for traditional and genomic hybrid breeding is substantial, and thus the rate of genetic gain is generally less for hybrid than inbred breeding (Longin et al., 2012). Furthermore, hybrid seed production is generally more expensive that inbred seed production, and hybrid genotypes cannot be replicated by selfing (Schulthess et al., 2017).

One alternative approach to using heterosis is to systematically reproduce desirable non-inbred genotypes. A major barrier to utilization of superior genotypes in non-inbred populations is that they cannot be repeatedly reproduced identically by crossing, since their parents are not fully inbred (Wricke and Weber, 1986). However, recent proof-of-concept "reverse breeding" in *Arabidopsis thaliana* offers an alternative to fixing heterosis by crossing inbred lines (Wijnker et al., 2012). In reverse breeding, recombination is suppressed in non-inbred lines, and DH lines are generated from their gametes (Wijnker et al., 2012). The DH lines can then be maintained, genotyped, and crossed at will to reconstitute the original non-inbred line (Wijnker et al., 2012). However, this method has not been tested in crop species or applied in crop breeding. A similar approach is synthetic apomixis, in which seeds identical to the parent plant are produced without meiosis or fertilization (Wang et al., 2019). In rice, apomictic seeds can be produced by editing only four genes, but fertility issues leading to low seed set also result (Wang et al., 2019).

Clonal propagation methods also reproduce non-inbred genotypes. Many non-inbred economically important crops, including sugarcane, potato, and cassava, are propagated asexually as clones rather than from seed (McKey et al., 2010). The drawbacks of clonal propagation, however, include the accumulation of deleterious somatic mutations, disease, costs of propagule production, and the recalcitrance of some species to clonally propagate (McKey et al., 2010). Use of polyploidy

has also been viewed as a way to "immortalize" hybrids, as allopolyploids can maintain heterozygosity across their subgenomes at individual loci even upon selfing (Santantonio et al., 2019). Dosage effects, or changes in phenotype due to increases in allele copy number, independent of allele state, have also been posited to contribute to the genetic values of polyploids relative to genotypes of lesser ploidies (Gianinetti, 2013; Yao et al., 2013; Fort et al., 2016). Unlike in diploids, heterosis in polyploids is not maximized in a single cross; this phenomenon is termed progressive heterosis (Washburn and Birchler, 2014). Progressive heterosis is expected in polyploids because the number of gametes inherited exceeds the number of parents. Going beyond single crosses permits combining gametes from more than two parents into a single individual genotype, so additional heterosis results. For example, in autotetraploids, heterosis would be maximized by a four-way cross. In diploids, the number of gametes inherited by the F_1 progeny in a single cross (two) is equal to the number of parents (two), so heterosis is maximized in single crosses.

If a uniform population is not necessary, then breeding open-pollinated varieties (OPVs) can lead to effective utilization of heterosis. For much of human history, OPVs were the only varieties available. This is still true in regions which the commercial breeding sector does not yet serve, and OPVs can outperform hybrids in some low-input environments (Pixley, 2006; Masuka et al., 2017; Andorf et al., 2019). In United States maize production, OPVs were abandoned in the early 1920s due to the difficulty of improving their quantitative traits (i.e., yield) as well as lack of uniformity (Duvick, 1999). However, it is unknown whether OPVs could outperform hybrids today if they had been as intensively developed (Duvick, 1999). OPV breeding could be advanced by genomic selection; the breeding cycle for OPVs is shorter and less costly than for hybrids. Furthermore, if used as part of a reverse breeding pipeline, sufficiently outstanding individuals from OPVs could be reproduced indefinitely as uniform "hybrid" varieties.

If heterosis is largely due to dominance rather than overdominance, then inbred lines which perform as well as hybrids must be possible, although they may take time to develop due to linkage disequilibrium (Werner et al., 2020). There is some evidence from commercial maize programs that inbred lines bred conventionally are already beginning to approach hybrid line performance, though likely because of the longer hybrid breeding cycle (Troyer and Wellin, 2009). Heterotic effects of yield have decreased as a percentage of mean yield over a short time—100 years—perhaps also because some favorable dominant alleles have been fixed in inbreds. Continued purging of deleterious recessive alleles from the genome by genome editing has been proposed, especially in regions hard to reach by recombination such as the centromere (Wallace et al., 2018; Valluru et al., 2019). If overdominance also affects hybrid performance, and overdominant loci can be identified, then arguably copy number variation could be induced to fix overdominance in inbred lines. These genomics-assisted approaches are reminiscent of genetic ideotype building, but until they are possible, genomics-assisted inbred line breeding may be a good start to genomics-assisted ideotype building of inbred lines (Trethowan, 2014).

LONG-TERM OPTIMIZATION OF SELECTION IN GENOMIC SELECTION PROGRAMS

All plant breeding programs require genetic variance for continued progress. Within any breeding population, reducing effective population size by selection early in the program may limit long-term genetic gain (Comstock et al., 1949; Robertson, 1960; Woolliams et al., 2015). Though genomic selection leads to less inbreeding than pedigree-based selection methods, inbreeding must be controlled (Rodríguez-Ramilo et al., 2015; Woolliams et al., 2015). Direct selection on GEBV maximizes gain in the subsequent cycle only and does not necessarily maximize long-term gain (Sonesson et al., 2012). Fortunately, data collected routinely in genomic selection programs allow monitoring and optimization of loss of diversity and inbreeding. Genomic selection strategies which seek to balance rates of genetic gain and loss of diversity include:

(a) Optimum contribution selection: genetic value is maximized while inbreeding is constrained to give the optimal contributions of parents to the next generation, i.e., number of progeny (Meuwissen and Sonesson, 1998).
(b) Weighting of rare alleles: allelic effects are weighted by their frequency such that rare favorable alleles are preserved (Goddard, 2009).
(c) Weighted genomic selection: allelic effects are weighted by their frequency, but also the magnitude of their effect, such that rare favorable alleles which tend to have large effects on EBV are preserved (Jannink, 2010).
(d) Genotype building: a subpopulation is selected algorithmically to segregate for maximal haplotype values, then intermated such that the two best segments ultimately segregate with equal frequency (Kemper et al., 2012).
(e) Optimal cross selection: selection intensity, inbreeding, and cross allocation are simultaneously optimized (Gorjanc et al., 2018).
(f) Usefulness criterion parental contribution: overall and within-family selection intensity, inbreeding, and cross allocation are simultaneously optimized (Allier et al., 2019).
(g) Genomic mating: genetic value, inbreeding, and risk (calculated from variability in breeding value estimates) are simultaneously optimized (Akdemir and Sánchez, 2016).
(h) Optimal haploid value selection: outbred individuals are selected for the predicted value of the best DH lines they could produce, then used to make DH lines for which breeding or genetic values are then predicted (Daetwyler et al., 2015).
(i) Optimal population value selection: sets of individuals are selected for their collective rather than individual maximum possible haploid value (Goiffon et al., 2017).
(j) Expected maximum haploid breeding value selection: individuals are selected for their maximum possible haploid value (Müller et al., 2018).
(k) IND-HE: genetic gain and expected heterozygosity are balanced in selection (De Beukelaer et al., 2017).

(l) Look-ahead selection: sets of individuals are selected for their collective rather than individual maximum possible haploid value, with the maximum value occurring in a user-specified target generation (Moeinizade et al., 2019).
(m) Optimal contribution selection to update the reference population: assuming a breeding population is used to update the training set for prediction, selecting the training set candidates by optimal contribution selection balances genetic gain and inbreeding (Eynard et al., 2018).
(n) Optimal contribution selection with branching: the population mating scheme is branched into two paths which maintain genetic diversity and maximize genetic gain (Santantonio and Robbins, 2020).

A simulation comparing all methods of long-term selection optimization in hybrid breeding programs is not yet available (Rembe et al., 2019). Development of genomic selection strategies to optimally introgress novel variation are also ongoing (Rembe et al., 2019). A recent comparison of introducing genetic donors with varying performance levels either using or omitting a bridging population to increase mean genetic values of introgression lines found that use of a bridging step was more useful when considering low-value donors, and that controlled introduction of diversity increased gain relative to a completely closed population (Allier et al., 2020a). Though the field of long-term selection optimization developed in response to need to avoid inbreeding and maintain genetic variance, the techniques developed can also be used to improve short- and medium-term gain (Müller et al., 2018).

For hybrid programs specifically, selection optimization methods to prevent unintentional allelic fixation during RRGS in opposite heterotic groups could be useful (Cowling et al., 2020). Another issue in hybrid breeding over time is introducing new germplasm and assigning it to a heterotic pool. Traditionally, new individuals are assigned to a heterotic pool by their phenotypic similarity to existing members or observed performance in testcrosses with representatives of each pool (Melchinger, 1999). Alternatively, individuals can be assigned to pools by genetic resemblance (Melchinger, 1999; Boeven et al., 2016). However, in practice, genetic distance is not consistently useful in assigning individuals to heterotic pools (Fischer et al., 2010; Brauner et al., 2019).

Though advanced commercial maize hybrid breeding programs should not be construed as resulting from long-term genomic selection, the genetic base of North American and European commercial maize is narrow, prompting concern that limiting loss of diversity has occurred (Brauner et al., 2019; Allier et al., 2020b). Some approaches, such as Germplasm Enhancement of Maize (GEM), have proposed adaptation of exotic germplasm to commercial inbred backgrounds by public-private collaboration. Several inbred lines have been released as a result of GEM (Samayoa et al., 2018). Other efforts based on generating DH lines of maize landraces and characterizing them for their *per se* and testcross performance with European testers have also demonstrated a 15% yield gap between mean testcross yield and mean commercial yield (Brauner et al., 2019; Hölker et al., 2019). Genomic selection for line adaptation has been proposed but is largely untested (Bernardo, 2009; Samayoa et al., 2018; Allier et al., 2020b). Additionally, commercial breeding programs may reduce loss of useful diversity by targeted introgression of QTL or transgenes into elite lines, which improves the lines without drastically changing their genomic makeup or disrupting the heterotic pattern (Samayoa et al., 2018).

The limits of long-term selection within closed breeding populations are unknown (Dudley and Lambert, 2004; Paixão and Barton, 2016). Breeding progress for high grain oil and protein content, which was initiated in a maize OPV, has continued for over 100 cycles of selection without introduction of new germplasm (Dudley and Lambert, 2004; Moose et al., 2004). In the same experiment, breeding progress for low oil and protein content ceased due to measurement and physiological constraints, respectively (Dudley and Lambert, 2004). Surprisingly, when the direction of selection on lines bred for low oil and protein content was experimentally reversed at 48 generations, selection response in the opposite direction occurred rapidly (Dudley and Lambert, 2004). Though not conclusive, these results suggest that it is difficult to exhaust response to selection even in completely closed or selected populations using conventional recurrent selection strategies. The cost of testing and adapting vast quantities of new germplasm may not be worth the short-term benefits for advanced commercial hybrid programs if sufficient genetic variance remains for selection gain, even among very few lines.

DISCUSSION

Exact recommendations for crop hybrid breeding programs are situation-dependent, including whether and how to apply genomic selection. Factors to consider in implementation of genomic selection strategies include budget, trait heritability, cost and accuracy of phenotyping, length of the breeding cycle, and infrastructure for genomic selection (e.g., marker availability, marker cost, bioinformatics software, statistical expertise, etc.; Heslot et al., 2015). General factors that affect the success of a hybrid breeding program relative to an inbred breeding program include (a) mating system, including whether selfing is possible, (b) existence of heterotic pools, (c) the degree of heterosis, (d) the cost of hybrid seed production, including availability of hybridization systems, and (e) the number of seeds needed in the cropping system (Longin et al., 2013). Another rationale for hybrid breeding has been that the sale of hybrid seeds generates a sustainable funding model for breeding, with built-in variety protection (Schulthess et al., 2017). However, this argument is beyond the scope of plant breeding and requires economics research.

Mating system is a major factor driving the use of hybrid breeding systems (Longin et al., 2013). Most hybrid crops (e.g., maize, sugarbeet, rye, and sunflower) are allogamous, or outcrossing, rather than autogamous, or selfing. Though both present difficulties in breeding programs—autogamous crops may be difficult to cross, and allogamous crops may be difficult or nearly impossible to self—in general autogamous

crops are less amenable to hybrid breeding due to higher costs of seed production and less observed heterosis (Wricke and Weber, 1986; Longin et al., 2012). Less heterosis in autogamous than allogamous crops may have an evolutionary basis. Over time, deleterious recessive mutations are more exposed to selection in selfing than outcrossing species—ultimately leading to reduced inbreeding depression (Moose et al., 2004). Selfing genotypes also have more opportunities for selection on epistatic networks, perhaps leading to increased outbreeding depression (Fenster et al., 1997).

For a breeding program, the question then remains whether the gains of heterosis are outweighed by the costs of breeding hybrids in autogamous crops. The costs of breeding hybrids can be reduced by developing male-female heterotic pools, scalable male sterility systems, and hybridization systems. The gains of heterosis can be increased by breeding heterotic pools. Thus, initial investment to establish a hybrid breeding program may be high, but it could provide higher returns over time than an inbred program. A case study of hybrid wheat, for example, found that although hybrids are currently competitive with inbred varieties, whether long-term improvement of hybrids keeps pace with lines strongly depends on budget, cost of hybrid seed production, and GCA variance (Longin et al., 2014). Use of genomic prediction can increase the relative efficiency of hybrid breeding to line breeding (Longin et al., 2015). If sufficient budgets to cover the start-up costs of hybrid breeding (e.g., heterotic pool development, male sterility systems) are available at no cost to line breeding, then hybrid breeding is worth investigating.

Whether for autogamous or allogamous species, genomic selection methods have potential to increase rates of genetic gain at every stage of hybrid breeding. Use of genomic selection, for example, to rapidly develop heterotic pools in crops in which they are not well-established—e.g., rice and wheat— is worth trying (Rembe et al., 2019). Further reports on RRGS programs which have been initiated in oil palm, which has a long generation interval, high phenotyping costs, and high environmental impact, are anticipated (Cros et al., 2017; Nyouma et al., 2019). In selection of single crosses, genomic prediction has potential to reduce the need for testcrossing and field evaluation (Longin et al., 2015; Kadam et al., 2016). Longin et al. (2015) considered optimal allocation of resources to number of DH lines, test locations, and tester lines used for inbred and hybrid breeding programs with different degrees of reliance on genomic prediction and different prediction accuracies. After DH production, testcrosses or lines were either immediately subject to genomic selection, advanced through one round of field testing, or advanced through two rounds of field testing (Longin et al., 2015). The importance of field testing strongly depended on accuracies of genomic predictions, but for hybrid breeding, even prediction with low accuracies improved rate of genetic gain (Longin et al., 2015). If DH lines underwent a round of phenotypic selection before advancing, the relative merits of incorporating genomic selection did not change (Marulanda et al., 2016). The best scenario was genomic prediction followed by a round of phenotyping (Longin et al., 2015; Marulanda et al., 2016).

The era of genomic selection offers new opportunities in hybrid breeding. Genomic selection methods can jumpstart the establishment of heterotic pools by founder selection and use of RRGS to unlock heterosis in new hybrid breeding programs. Genomic selection can also shorten the notoriously long hybrid breeding cycle by reducing the need for testcrosses and their phenotypic evaluation. Though implementing genomic selection methods requires optimization to specific hybrid breeding situations, a sufficient framework for breeders to make genomics-assisted decisions already exists.

AUTHOR CONTRIBUTIONS

ML conducted the literature review and wrote the manuscript. AS and JR conceived of the review topic, edited the manuscript, and generated cross-disciplinary insights through discussion. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Jonathan Baldwin Turner Fellowship and the Crop Sciences Department of the College of Agricultural, Consumer, and Environmental Sciences at the University of Illinois at Urbana–Champaign.

ACKNOWLEDGMENTS

We thank R. Chris Gaynor for editing the manuscript and providing the population-level quantitative genetic framework of heterosis. We thank Martin Bohn and Amy Marshall-Colón for helpful discussion. We thank our reviewers for their critical insights.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.643761/full#supplementary-material

REFERENCES

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci. Rep.* 9:1146.

Akdemir, D., and Sánchez, J. I. (2016). Efficient breeding by genomic mating. *Front. Genet.* 7:210. doi: 10.3389/fgene.2016.00210

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., Goddard, M. E., and Hayes, B. J. (2017). Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. *J. Dairy Sci.* 100, 1203–1222. doi: 10.3168/jds.2016-11261

Allier, A., Lehermeier, C., Charcosset, A., Moreau, L., and Teyssèdre, S. (2019). Improving short and long term genetic gain by accounting for within family

variance in optimal cross selection. *Front. Genet.* 10:1006. doi: 10.3389/fgene.2019.01006

Allier, A., Teyssèdre, S., Lehermeier, C., Charcosset, A., and Moreau, L. (2020b). Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs. *Theor. Appl. Genet.* 133, 201–215. doi: 10.1007/s00122-019-03451-9

Allier, A., Teyssèdre, S., Lehermeier, C., Moreau, L., and Charcosset, A. (2020a). Optimized breeding strategies to harness Genetic Resources with different performance levels. *BMC Genomics* 21:349. doi: 10.1186/s12864-020-6756-0

Andorf, C., Beavis, W. D., Hufford, M., Smith, S., Suza, W. P., Wang, K., et al. (2019). Technological advances in maize breeding: past, present and future. *Theor. Appl. Genet.* 132, 817–849. doi: 10.1007/s00122-019-03306-3

Auger, D. L., Peters, E. M., and Birchler, J. A. (2005). A genetic test of bioactive gibberellins as regulators of heterosis in maize. *J. Hered.* 96, 614–617. doi: 10.1093/jhered/esi102

Barrett, S. C. H., and Charlesworth, D. (1991). Effects of a change in the level of inbreeding on the genetic load. *Nature* 352, 522–524. doi: 10.1038/352522a0

Bar-Zvi, D., Lupo, O., Levy, A. A., and Barkai, N. (2017). Hybrid vigor: the best of both parents, or a genomic clash? *Curr. Opin. Syst. Biol.* 6, 22–27. doi: 10.1016/j.coisb.2017.08.004

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci.* 34, 20–25. doi: 10.2135/cropsci1994.0011183x003400010003x

Bernardo, R. (1996). Best linear unbiased prediction of maize single-cross performance given erroneous inbred relationships. *Crop Sci.* 36, 862–866. doi: 10.2135/cropsci1996.0011183x003600040007x

Bernardo, R. (2002). *Breeding for Quantitative Traits in Plants*, Vol. 1. Woodbury, MN: Stemma Press, 369.

Bernardo, R. (2009). Genomewide selection for rapid introgression of exotic germplasm in maize. *Crop Sci.* 49, 419–425. doi: 10.2135/cropsci2008.08.0452

Bernardo, R. (2020). Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity* 125, 375–385. doi: 10.1038/s41437-020-0312-1

Bernardo, R., and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47, 1082–1090. doi: 10.2135/cropsci2006.11.0690

Beukert, U., Li, Z., Liu, G., Zhao, Y., Ramachandra, N., Mirdita, V., et al. (2017). Genome-based identification of heterotic patterns in rice. *Rice* 10, 1–10.

Birchler, J. A., Auger, D. L., and Riddle, N. C. (2003). In search of the molecular basis of heterosis. *Plant Cell* 15, 2236–2239. doi: 10.1105/tpc.151030

Birchler, J. A., Yao, H., Chudalayandi, S., Vaiman, D., and Veitia, R. A. (2010). Heterosis. *Plant Cell* 22, 2105–2112.

Boeven, P. H., Longin, C. F. H., and Würschum, T. (2016). A unified framework for hybrid breeding and the establishment of heterotic groups in wheat. *Theor. Appl. Genet.* 129, 1231–1245. doi: 10.1007/s00122-016-2699-x

Brauner, P. C., Schipprack, W., Utz, H. F., Bauer, E., Mayer, M., Schön, C. C., et al. (2019). Testcross performance of doubled haploid lines from European flint maize landraces is promising for broadening the genetic base of elite germplasm. *Theor. Appl. Genet.* 132, 1897–1908. doi: 10.1007/s00122-019-03325-0

Burdo, B., Gray, J., Goetting−Minesky, M. P., Wittler, B., Hunt, M., Li, T., et al. (2014). The Maize TF ome–development of a transcription factor open reading frame collection for functional genomics. *Plant J.* 80, 356–366. doi: 10.1111/tpj.12623

Charlesworth, D., and Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.* 18, 237–268. doi: 10.1146/annurev.es.18.110187.001321

Charlesworth, D., and Willis, J. H. (2009). The genetics of inbreeding depression. *Nat. Rev. Genet.* 10, 783–796.

Chen, Z. J. (2013). Genomic and epigenetic insights into the molecular bases of heterosis. *Nat. Rev. Genet.* 14, 471–482. doi: 10.1038/nrg3503

Chen, Z. J., and Birchler, J. A. (2013). *Polyploid and Hybrid Genomics*. Hoboken, NJ: John Wiley & Sons.

Cheverud, J. M., and Routman, E. J. (1995). Epistasis and its contribution to genetic variance components. *Genetics* 139, 1455–1461. doi: 10.1093/genetics/139.3.1455

Christensen, O. F., Nielsen, B., Su, G., Xiang, T., Madsen, P., Ostersen, T., et al. (2019). A bivariate genomic model with additive, dominance and inbreeding depression effects for sire line and three-way crossbred pigs. *Genet. Select. Evol.* 51:45.

Comstock, R. E., Robinson, H. F., and Harvey, P. H. (1949). A breeding procedure designed to make maximum use of both general and specific combining ability 1. *Agron. J.* 41, 360–367. doi: 10.2134/agronj1949.00021962004100080006x

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* 11, 2463–2468. doi: 10.1093/hmg/11.20.2463

Cowling, W. A., Gaynor, R. C., Antolín, R., Gorjanc, G., Edwards, S. M., Powell, O., et al. (2020). In silico simulation of future hybrid performance to evaluate heterotic pool formation in a self-pollinating crop. *Sci. Rep.* 10, 1–8. doi: 10.1002/csc2.20033

Coyne, D. P. (1965). Component interaction in relation to heterosis for plant height in *Phaseolus vulgaris* L. Variety crosses 1. *Crop Sci.* 5, 17–18. doi: 10.2135/cropsci1965.0011183x000500010007x

Cress, C. E. (1966). A comparison of recurrent selection systems. *Genetics* 54, 1371. doi: 10.1093/genetics/54.6.1371

Cros, D., Bocs, S., Riou, V., Ortega-Abboud, E., Tisné, S., Argout, X., et al. (2017). Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. *BMC Genomics* 18:839. doi: 10.1186/s12864-017-4179-3

Cros, D., Denis, M., Bouvet, J. M., and Sánchez, L. (2015). Long-term genomic selection for heterosis without dominance in multiplicative traits: case study of bunch production in oil palm. *BMC Genomics* 16:651. doi: 10.1186/s12864-015-1866-9

Cros, D., Tchounke, B., and Nkague-Nkamba, L. (2018). Training genomic selection models across several breeding cycles increases genetic gain in oil palm in silico study. *Mol. Breed.* 38:89.

Crow, J. F. (1999). "Dominance and overdominance," in *Genetics and Exploitation of Heterosis in Crops*, eds J. G. Coors and S. Pandey (Madison, WI: American Society of Agronomy, Inc), 49–58. doi: 10.2134/1999.geneticsandexploitation.c5

Crow, J. F. (2010). On epistasis: why it is unimportant in polygenic directional selection. *Philos. Trans. R. Soc. B Biol. Sci.* 365, 1241–1244. doi: 10.1098/rstb.2009.0275

Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., and Hayes, B. J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics* 200, 1341–1348. doi: 10.1534/genetics.115.178038

Dan, Z., Hu, J., Zhou, W., Yao, G., Zhu, R., Huang, W., et al. (2015). Hierarchical additive effects on heterosis in rice (*Oryza sativa* L.). *Front. Plant Sci.* 6:738. doi: 10.3389/fpls.2015.00738

Davenport, C. B. (1908). Degeneration, albinism and inbreeding. *Science* 28, 454–455. doi: 10.1126/science.28.718.454-b

De Beukelaer, H., Badke, Y., Fack, V., and De Meyer, G. (2017). Moving beyond managing realized genomic relationship in long-term genomic selection. *Genetics* 206, 1127–1138. doi: 10.1534/genetics.116.194449

de la Luz Gutiérrez-Nava, M., Warren, C. A., León, P., and Walbot, V. (1998). Transcriptionally active MuDR, the regulatory element of the mutator transposable element family of Zea mays, is present in some accessions of the Mexican land race *Zapalote chico*. *Genetics* 149, 329–346.

de los Campos, G., Sorensen, D., and Gianola, D. (2015). Genomic heritability: what is it? *PLoS Genet.* 11:e1005048. doi: 10.1371/journal.pgen.1005048

Dias, K. O. G., Piepho, H. P., Guimarães, L. J. M., Guimarães, P. E. O., Parentoni, S. N., Pinto, M. O., et al. (2019). Novel strategies for genomic prediction of untested single-cross maize hybrids using unbalanced historical data. *Theor. Appl. Genet.* 133, 443–455. doi: 10.1007/s00122-019-03475-1

Dobzhansky, T. (1941). *Genetics and the Origin of Species*. New York, NY: Columbia University Press.

Dudley, J. W., and Lambert, R. J. (2004). 100 generations of selection for oil and protein in corn. *Plant Breed. Rev.* 24, 79–110. doi: 10.1002/9780470650240.ch5

Duenk, P., Calus, M. P., Wientjes, Y. C., and Bijma, P. (2017). Benefits of dominance over additive models for the estimation of average effects in the presence of dominance. *G3* 7, 3405–3414. doi: 10.1534/g3.117.300113

Duvick, D. N. (1999). "Heterosis: feeding people and protecting natural resources," in *Genetics and Exploitation of Heterosis in Crops*, eds J. G. Coors and S. Pandey

(Madison, WI: American Society of Agronomy, Inc), 19–29. doi: 10.2134/1999.geneticsandexploitation.c3

Duvick, D. N. (2005). Genetic progress in yield of United States maize (*Zea mays* L.). *Maydica* 50:193.

Duvick, D. N., Smith, J. S. C., and Cooper, M. (2004). Long-term selection in a commercial hybrid maize breeding program. *Plant Breed. Rev.* 24, 109–152. doi: 10.1002/9780470650288.ch4

East, E. M. (1908). Inbreeding in corn. *Rep. Conn. Agric. Exp. Stn* 1907, 419–428.

East, E. M. (1936). Heterosis. *Genetics* 21:375.

Eynard, S. E., Windig, J. J., Hulsegge, I., Hiemstra, S. J., and Calus, M. P. (2018). The impact of using old germplasm on genetic merit and diversity—A cattle breed case study. *J. Anim. Breed. Genet.* 135, 311–322. doi: 10.1111/jbg.12333

Falconer, D. S. (1985). A note on Fisher's 'average effect'and 'average excess'. *Genet. Res.* 46, 337–347. doi: 10.1017/s0016672300022825

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Essex: Longman Group.

Fenster, C. B., Galloway, L. F., and Chao, L. (1997). Epistasis and its consequences for the evolution of natural populations. *Trends Ecol. Evol.* 12, 282–286. doi: 10.1016/s0169-5347(97)81027-0

Fernandes, S. B., Dias, K. O., Ferreira, D. F., and Brown, P. J. (2018). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theor. Appl. Genet.* 131, 747–755. doi: 10.1007/s00122-017-3033-y

Fiévet, J. B., Nidelet, T., Dillmann, C., and de Vienne, D. (2018). Heterosis is a systemic property emerging from non-linear genotype-phenotype relationships: evidence from in vitro genetics and computer simulations. *Front. Genet.* 9:159. doi: 10.3389/fgene.2018.00159

Fischer, S., Melchinger, A. E., Korzun, V., Wilde, P., Schmiedchen, B., Möhring, J., et al. (2010). Molecular marker assisted broadening of the Central European heterotic groups in rye with Eastern European germplasm. *Theor. Appl. Genet.* 120, 291–299. doi: 10.1007/s00122-009-1124-0

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edinburgh* 52, 399–433. doi: 10.1017/s0080456800012163

Flint-Garcia, S. A., Buckler, E. S., Tiffin, P., Ersoz, E., and Springer, N. M. (2009). Heterosis is prevalent for multiple traits in diverse maize germplasm. *PLoS One* 4:e7433. doi: 10.1371/journal.pone.0007433

Fort, A., Ryder, P., McKeown, P. C., Wijnen, C., Aarts, M. G., Sulpice, R., et al. (2016). Disaggregating polyploidy, parental genome dosage and hybridity contributions to heterosis in *Arabidopsis thaliana*. *New Phytol.* 209, 590–599. doi: 10.1111/nph.13650

Fristche-Neto, R., Akdemir, D., and Jannink, J. L. (2018). Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. *Theor. Appl. Genet.* 131, 1153–1162. doi: 10.1007/s00122-018-3068-8

Fu, D., Xiao, M., Hayward, A., Fu, Y., Liu, G., Jiang, G., et al. (2014). Utilization of crop heterosis: a review. *Euphytica* 197, 161–173. doi: 10.1007/s10681-014-1103-7

Gianinetti, A. (2013). A criticism of the value of midparent in polyploidization. *J. Exp. Bot.* 64, 4119–4129. doi: 10.1093/jxb/ert263

Gianola, D., Cecchinato, A., Naya, H., and Schön, C. C. (2018). Prediction of complex traits: robust alternatives to best linear unbiased prediction. *Front. Genet.* 9:195. doi: 10.3389/fgene.2018.00195

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183, 347–363. doi: 10.1534/genetics.109.103952

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776. doi: 10.1534/genetics.105.049510

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Goiffon, M., Kusmec, A., Wang, L., Hu, G., and Schnable, P. S. (2017). Improving response in genomic selection with a population-based selection strategy: optimal population value selection. *Genetics* 206, 1675–1682. doi: 10.1534/genetics.116.197103

Gorjanc, G., Gaynor, R. C., and Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theor. Appl. Genet.* 131, 1953–1966. doi: 10.1007/s00122-018-3125-3

Govindaraju, D. R. (2019). An elucidation of over a century old enigma in genetics—Heterosis. *PLoS Biol.* 17:e3000215. doi: 10.1371/journal.pbio.3000215

Grafius, J. E. (1961). The complex trait as a geometric construct. *Heredity* 16, 225–228. doi: 10.1038/hdy.1961.24

Griffing, B. (1956a). A generalised treatment of the use of diallel crosses in quantitative inheritance. *Heredity* 10, 31–50. doi: 10.1038/hdy.1956.2

Griffing, B. (1956b). Concept of general and specific combining ability in relation to diallel crossing systems. *Aust. J. Biol. Sci.* 9, 463–493. doi: 10.1071/bi9560463

Grindeland, J. M. (2008). Inbreeding depression and outbreeding depression in Digitalis purpurea: optimal outcrossing distance in a tetraploid. *J. Evol. Biol.* 21, 716–726. doi: 10.1111/j.1420-9101.2008.01519.x

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., et al. (2019). Optimal designs for genomic selection in hybrid crops. *Mol. Plant* 12, 390–401. doi: 10.1016/j.molp.2018.12.022

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinform.* 12:186. doi: 10.1186/1471-2105-12-186

Hallauer, A. R., Carena, M. J., and Miranda Filho, J. D. (2010). *Quantitative Genetics in Maize Breeding*, Vol. 6. Berlin: Springer Science & Business Media.

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91, 47–60. doi: 10.1017/s0016672308009981

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.1007/978-3-319-63170-7_1

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447. doi: 10.2307/2529430

Herbst, R. H., Bar-Zvi, D., Reikhav, S., Soifer, I., Breker, M., Jona, G., et al. (2017). Heterosis as a consequence of regulatory incompatibility. *BMC Biol.* 15:38. doi: 10.1186/s12915-017-0373-7

Heslot, N., Jannink, J. L., and Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Sci.* 55, 1–12. doi: 10.2135/cropsci2014.03.0249

Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., Eggen, A., Kilian, A., et al. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* 49:1297. doi: 10.1038/ng.3920

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* 4:e1000008. doi: 10.1371/journal.pgen.1000008

Hochholdinger, F., and Hoecker, N. (2007). Towards the molecular basis of heterosis. *Trends Plant Sci.* 12, 427–432. doi: 10.1016/j.tplants.2007.08.005

Hölker, A. C., Mayer, M., Presterl, T., Bolduan, T., Bauer, E., Ordas, B., et al. (2019). European maize landraces made accessible for plant breeding and genome-based studies. *Theor. Appl. Genet.* 132, 3333–3345. doi: 10.1007/s00122-019-03428-8

Howard, R., Gianola, D., Montesinos-López, O., Juliana, P., Singh, R., Poland, J., et al. (2019). Joint use of genome, pedigree, and their interaction with environment for predicting the performance of wheat lines in new environments. *G3* 9, 2925–2934. doi: 10.1534/g3.119.400508

Huang, W., and Mackay, T. F. (2016). The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS Genet.* 12:e1006421. doi: 10.1371/journal.pgen.1006421

Huang, X., Yang, S., Gong, J., Zhao, Q., Feng, Q., Zhan, Q., et al. (2016). Genomic architecture of heterosis for yield traits in rice. *Nature* 537, 629–633. doi: 10.1038/nature19760

Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., et al. (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat. Commun.* 6, 1–9. doi: 10.1007/978-3-642-85115-5_1

Ibáñz-Escriche, N., Fernando, R. L., Toosi, A., and Dekkers, J. C. (2009). Genomic selection of purebreds for crossbred performance. *Genet. Select. Evol.* 41:12. doi: 10.1186/1297-9686-41-12

Jannink, J. L. (2010). Dynamics of long-term genomic selection. *Genet. Select. Evol.* 42:35.

Jia, Y., and Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 192, 1513–1522. doi: 10.1534/genetics.112.144246

Kacser, H., and Burns, J. A. (1981). The molecular basis of dominance. *Genetics* 97, 639–666.

Kadam, D. C., and Lorenz, A. J. (2018). "Toward redesigning hybrid maize breeding through genomics-assisted breeding," in *The Maize Genome*, eds J. Bennetzen, S. Flint-Garcia, C. Hirsch, and R. Tuberosa (Cham: Springer), 367–388. doi: 10.1007/978-3-319-97427-9_21

Kadam, D. C., and Lorenz, A. J. (2019). Evaluation of nonparametric models for genomic prediction of early-stage single crosses in maize. *Crop Sci.* 59, 1411–1423. doi: 10.2135/cropsci2017.11.0668

Kadam, D. C., Potts, S. M., Bohn, M. O., Lipka, A. E., and Lorenz, A. J. (2016). Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. *G3* 6, 3443–3453. doi: 10.1534/g3.116.031286

Kaeppler, S. (2011). Heterosis: one boat at a time, or a rising tide? *New Phytol.* 189, 900–902. doi: 10.1111/j.1469-8137.2010.03630.x

Kaeppler, S. (2012). Heterosis: many genes, many mechanisms—end the search for an undiscovered unifying theory. *ISRN Bot.* 2012:682824.

Kemper, K. E., Bowman, P. J., Pryce, J. E., Hayes, B. J., and Goddard, M. E. (2012). Long-term selection strategies for complex traits using high-density genetic markers. *J. Dairy Sci.* 95, 4646–4656. doi: 10.3168/jds.2011-5289

Kinghorn, B. P., Hickey, J. M., and Van Der Werf, J. H. J. (2010). "Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals," in *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, (Leipzig: German Society for Animal Science), 1–6.

Krieger, U., Lippman, Z. B., and Zamir, D. (2010). The flowering gene SINGLE FLOWER TRUSS drives heterosis for yield in tomato. *Nat. Genet.* 42:459. doi: 10.1038/ng.550

Lamkey, K. R., and Edwards, J. W. (1999). "Quantitative genetics of heterosis," in *Genetics and Exploitation of Heterosis in Crops*, eds J. G. Coors and S. Pandey (Madison, WI: American Society of Agronomy, Inc), 31–48. doi: 10.2134/1999.geneticsandexploitation.c4

Lee, E. A., and Tracy, W. F. (2009). "Modern maize breeding," in *Handbook of Maize*, eds J. L. Bennetzen and S. Hake (New York, NY: Springer), 141–160. doi: 10.1007/978-0-387-77863-1_7

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92, 4656–4663. doi: 10.3168/jds.2009-2061

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. *Trends Genet.* 27, 323–331. doi: 10.1016/j.tig.2011.05.007

Li, Z., Zhou, P., Della Coletta, R., Zhang, T., Brohammer, A. B., O'Connor, C., et al. (2020). Single−parent expression drives dynamic gene expression complementation in maize hybrids. *Plant J.* doi: 10.1111/tpj.15042 [Epub ahead of print].

Liang, Z., Gupta, S. K., Yeh, C. T., Zhang, Y., Ngu, D. W., Kumar, R., et al. (2018). Phenotypic data from inbred parents can improve genomic prediction in pearl millet hybrids. *G3* 8, 2513–2522. doi: 10.1534/g3.118.200242

Lippman, Z. B., and Zamir, D. (2007). Heterosis: revisiting the magic. *Trends Genet.* 23, 60–66. doi: 10.1016/j.tig.2006.12.006

Liu, H., Wang, Q., Chen, M., Ding, Y., Yang, X., Liu, J., et al. (2020). Genome−wide identification and analysis of heterotic loci in three maize hybrids. *Plant Biotechnol. J.* 18, 185–194. doi: 10.1111/pbi.13186

Longin, C. F. H., Gowda, M., Mühleisen, J., Ebmeyer, E., Kazman, E., Schachschneider, R., et al. (2013). Hybrid wheat: quantitative genetic parameters and consequences for the design of breeding programs. *Theor. Appl. Genet.* 126, 2791–2801. doi: 10.1007/s00122-013-2172-z

Longin, C. F. H., Mi, X., and Würschum, T. (2015). Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. *Theor. Appl. Genet.* 128, 1297–1306. doi: 10.1007/s00122-015-2505-1

Longin, C. F. H., Mühleisen, J., Maurer, H. P., Zhang, H., Gowda, M., and Reif, J. C. (2012). Hybrid breeding in autogamous cereals. *Theor. Appl. Genet.* 125, 1087–1096. doi: 10.1007/s00122-012-1967-7

Longin, C. F. H., Reif, J. C., and Würschum, T. (2014). Long-term perspective of hybrid versus line breeding in wheat based on quantitative genetic theory. *Theor. Appl. Genet.* 127, 1635–1641. doi: 10.1007/s00122-014-2325-8

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., et al. (2011). Genomic selection in plant breeding: knowledge and prospects. *Adv. Agron.* 110, 77–123.

Lynch, M. (1991). The genetic interpretation of inbreeding depression and outbreeding depression. *Evolution* 45, 622–629. doi: 10.2307/2409915

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*, Vol. 1. Sunderland, MA: Sinauer, 535–557.

Manfredi, E., Tusell, L., and Vitezica, Z. G. (2017). Prediction of complex traits: conciliating genetics and statistics. *J. Anim. Breed. Genet.* 134, 178–183. doi: 10.1111/jbg.12269

Manna, F., Martin, G., and Lenormand, T. (2011). Fitness landscapes: an alternative theory for the dominance of mutation. *Genetics* 189, 923–937. doi: 10.1534/genetics.111.132944

Marcon, C., Schützenmeister, A., Schütz, W., Madlung, J., Piepho, H. P., and Hochholdinger, F. (2010). Nonadditive protein accumulation patterns in maize (*Zea mays* L.) hybrids during embryo development. *J. Proteome Res.* 9, 6511–6522. doi: 10.1021/pr100718d

Marshall-Colón, A., Sengupta, N., Rhodes, D., Dudareva, N., and Morgan, J. (2010). A kinetic model describes metabolic response to perturbations and distribution of flux control in the benzenoid network of *Petunia hybrida*. *Plant J.* 62, 64–76. doi: 10.1111/j.1365-313x.2010.04127.x

Marulanda, J. J., Mi, X., Melchinger, A. E., Xu, J. L., Würschum, T., and Longin, C. F. H. (2016). Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. *Theor. Appl. Genet.* 129, 1901–1913. doi: 10.1007/s00122-016-2748-5

Masuka, B., Magorokosho, C., Olsen, M., Atlin, G. N., Bänziger, M., Pixley, K. V., et al. (2017). Gains in maize genetic improvement in Eastern and Southern Africa: II. CIMMYT open-pollinated variety breeding pipeline. *Crop Sci.* 57, 180–191. doi: 10.2135/cropsci2016.05.0408

McKey, D., Elias, M., Pujol, B., and Duputié, A. (2010). The evolutionary ecology of clonally propagated domesticated plants. *New Phytol.* 186, 318–332. doi: 10.1111/j.1469-8137.2010.03210.x

Melchinger, A. E. (1999). "Genetic diversity and heterosis," in *Genetics and Exploitation of Heterosis in Crops*, eds J. G. Coors and S. Pandey (Madison, WI: American Society of Agronomy, Inc), 99–118. doi: 10.2134/1999.geneticsandexploitation.c10

Melchinger, A. E., and Gumber, R. K. (1998). "Overview of heterosis and heterotic groups in agronomic crops," in *Concepts and Breeding of Heterosis in Crop Plants*, Vol. 25, eds K. R. Larnkey and J. E. Staub (Madison, WI: Crop Science Society of America, Inc), 29–44. doi: 10.2135/cssaspecpub25.c3

Melchinger, A. E., Singh, M., Link, W., Utz, H. F., and Von Kittlitz, E. (1994). Heterosis and gene effects of multiplicative characters: theoretical relationships and experimental results from *Vicia faba* L. *Theor. Appl. Genet.* 88, 343–348. doi: 10.1007/bf00223643

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

Meuwissen, T. H. E., and Sonesson, A. K. (1998). Maximizing the response of selection with a predefined rate of inbreeding: overlapping generations. *J. Anim. Sci.* 76, 2575–2583. doi: 10.2527/1998.76102575x

Mikel, M. A., and Dudley, J. W. (2006). Evolution of North American dent corn from public to proprietary germplasm. *Crop Sci.* 46, 1193–1205. doi: 10.2135/cropsci2005.10-0371

Moeinizade, S., Hu, G., Wang, L., and Schnable, P. S. (2019). Optimizing selection and mating in genomic selection with a look-ahead approach: an operations research framework. *G3* 9, 2123–2133. doi: 10.1534/g3.118.200842

Momen, M., Mehrgardi, A. A., Sheikhi, A., Kranis, A., Tusell, L., Morota, G., et al. (2018). Predictive ability of genome-assisted statistical models under various forms of gene action. *Sci. Rep.* 8, 1–11.

Moose, S. P., Dudley, J. W., and Rocheford, T. R. (2004). Maize selection passes the century mark: a unique resource for 21st century genomics. *Trends Plant Sci.* 9, 358–364. doi: 10.1016/j.tplants.2004.05.005

Müller, D., Schopp, P., and Melchinger, A. E. (2018). Selection on expected maximum haploid breeding values can increase genetic gain in recurrent genomic selection. *G3* 8, 1173–1181. doi: 10.1534/g3.118.200091

Neyhart, J. L., Tiede, T., Lorenz, A. J., and Smith, K. P. (2017). Evaluating methods of updating training data in long-term genomewide selection. *G3* 7, 1499–1510. doi: 10.1534/g3.117.040550

Nyouma, A., Bell, J. M., Jacob, F., and Cros, D. (2019). From mass selection to genomic selection: one century of breeding for quantitative yield components of oil palm (*Elaeis guineensis* Jacq.). *Tree Genet. Genomes* 15:69.

Oakley, C. G., Ågren, J., and Schemske, D. W. (2015). Heterosis and outbreeding depression in crosses between natural populations of Arabidopsis thaliana. *Heredity* 115, 73–82. doi: 10.1038/hdy.2015.18

Okeke, U. G., Akdemir, D., Rabbi, I., Kulakow, P., and Jannink, J. L. (2017). Accuracies of univariate and multivariate genomic prediction models in African cassava. *Genet. Select. Evol.* 49:88.

Paixão, T., and Barton, N. H. (2016). The effect of gene interactions on the long-term response to selection. *Proceedings of the National Academy of Sciences* 113, 4422–4427. doi: 10.1073/pnas.1518830113

Phillips, P. C. (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* 9, 855–867. doi: 10.1038/nrg2452

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228. doi: 10.1007/s10681-007-9449-8

Pixley, K. V. (2006). "Hybrid and open−pollinated varieties in modern agriculture," in *Plant Breeding: The Arnel R. Hallauer International Symposium*, eds K. R. Lamkey and M. Lee (Ames, IA: Blackwell Publishing), 234–250. doi: 10.1002/9780470752708.ch17

Powers, L. (1944). An expansion of Jones's theory for the explanation of heterosis. *Am. Nat.* 78, 275–280. doi: 10.1086/281199

Ramstein, G. P., Jensen, S. E., and Buckler, E. S. (2019). Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theor. Appl. Genet.* 132, 559–567. doi: 10.1007/s00122-018-3267-3

Reif, J. C., Gumpert, F. M., Fischer, S., and Melchinger, A. E. (2007). Impact of interpopulation divergence on additive and dominance variance in hybrid populations. *Genetics* 176, 1931–1934. doi: 10.1534/genetics.107.074146

Reif, J. C., Hallauer, A. R., and Melchinger, A. E. (2005). Heterosis and heterotic patterns in maize. *Maydica* 50:215.

Rembe, M., Zhao, Y., Jiang, Y., and Reif, J. C. (2019). Reciprocal recurrent genomic selection: an attractive tool to leverage hybrid wheat breeding. *Theor. Appl. Genet.* 132, 687–698. doi: 10.1007/s00122-018-3244-x

Robertson, A. (1960). A theory of limits in artificial selection. *Proc. R. Soc. Lond. B. Biol. Sci.* 153, 234–249. doi: 10.1098/rspb.1960.0099

Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., Mitchell, S. E., et al. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. U.S.A.* 112, 3823–3828. doi: 10.1073/pnas.1413864112

Rodríguez-Ramilo, S. T., García-Cortés, L. A., and de Cara, M. ÁR. (2015). Artificial selection with traditional or genomic relationships: consequences in coancestry and genetic diversity. *Front. Genet.* 6:127. doi: 10.3389/fgene.2015.00127

Rutkoski, J. E., Crain, J., Poland, J., and Sorrells, M. E. (2017). "Genomic selection for small grain improvement," in *Genomic Selection for Crop Improvement*, eds R. K. Varshney, M. Roorkiwal, and M. E. Sorrells (Cham: Springer), 99–130. doi: 10.1007/978-3-319-63170-7_5

Samayoa, L. F., Dunne, J. C., Andres, R. J., and Holland, J. B. (2018). "Harnessing maize biodiversity," in *The Maize Genome*, eds J. Bennetzen, S. Flint-Garcia, C. Hirsch, and R. Tuberosa (Cham: Springer), 335–366. doi: 10.1007/978-3-319-97427-9_20

Santantonio, N., Jannink, J. L., and Sorrells, M. (2019). Homeologous epistasis in wheat: the search for an immortal hybrid. *Genetics* 211, 1105–1122. doi: 10.1534/genetics.118.301851

Santantonio, N., and Robbins, K. R. (2020). A hybrid optimal contribution approach to drive short-term gains while maintaining long-term sustainability in a modern plant breeding program. *bioRxiv [Preprint]* doi: 10.1101/2020.01.08.899039

Schnable, P. S., and Springer, N. M. (2013). Progress toward understanding heterosis in crop plants. *Annu. Rev. Plant Biol.* 64, 71–88. doi: 10.1146/annurev-arplant-042110-103827

Schnell, F. W., and Cockerham, C. C. (1992). Multiplicative vs. arbitrary gene action in heterosis. *Genetics* 131, 461–469. doi: 10.1093/genetics/131.2.461

Schön, C. C., and Simianer, H. (2015). Resemblance between two relatives–animal and plant breeding. *J. Anim. Breed. Genet.* 132, 1–2. doi: 10.1111/jbg.12137

Schrag, T. A., Schipprack, W., and Melchinger, A. E. (2019). Across-years prediction of hybrid performance in maize using genomics. *Theor. Appl. Genet.* 132, 933–946. doi: 10.1007/s00122-018-3249-5

Schrag, T. A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., et al. (2018). Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics* 208, 1373–1385. doi: 10.1534/genetics.117.300374

Schulthess, A. W., Zhao, Y., and Reif, J. C. (2017). "Genomic selection in hybrid breeding," in *Genomic Selection for Crop Improvement*, eds R. K. Varshney, M. Roorkiwal, and M. E. Sorrells (Cham: Springer), 149–183. doi: 10.1007/978-3-319-63170-7_7

Seifert, F., Thiemann, A., Schrag, T. A., Rybka, D., Melchinger, A. E., Frisch, M., et al. (2018). Small RNA-based prediction of hybrid performance in maize. *BMC Genomics* 19:371. doi: 10.1186/s12864-018-4708-8

Shull, G. H. (1908). The composition of a field of maize. *J. Hered.* 4, 296–301. doi: 10.1093/jhered/os-4.1.296

Shull, G. H. (1948). What is" heterosis"? *Genetics* 33:439.

Shull, G. H. (1952). "Beginnings of a heterosis concept," in *Heterosis*, ed. W. Gowen (Ames, IA: Iowa State College Press), 14–48.

Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. (2012). Genomic selection requires genomic control of inbreeding. *Genet. Select. Evol.* 44:27.

Sprague, G. F., and Tatum, L. A. (1942). General vs. specific combining ability in single crosses of corn 1. *Agron. J.* 34, 923–932. doi: 10.2134/agronj1942.00021962003400100008x

Springer, N. M., and Stupar, R. M. (2007). Allelic variation and heterosis in maize: how do two halves make more than a whole? *Genome Res.* 17, 264–275. doi: 10.1101/gr.5347007

Stitzer, M. C., and Ross-Ibarra, J. (2018). Maize domestication and gene interaction. *New Phytol.* 220, 395–408. doi: 10.1111/nph.15350

Stuber, C. W., and Cockerham, C. C. (1966). Gene effects and variances in hybrid populations. *Genetics* 54:1279. doi: 10.1093/genetics/54.6.1279

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J. L., and Sorrells, M. E. (2017). Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *Plant Genome* 10, 1–12.

Swanson-Wagner, R. A., Jia, Y., DeCook, R., Borsuk, L. A., Nettleton, D., and Schnable, P. S. (2006). All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6805–6810. doi: 10.1073/pnas.0510430103

Technow, F. (2019). Use of F2 bulks in training sets for genomic prediction of combining ability and hybrid performance. *G3* 9, 1557–1569. doi: 10.1534/g3.118.200994

Technow, F., Riedelsheimer, C., Schrag, T. A., and Melchinger, A. E. (2012). Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor. Appl. Genet.* 125, 1181–1194. doi: 10.1007/s00122-012-1905-8

Tracy, W. F., and Chandler, M. A. (2006). "The historical and biological basis of the concept of heterotic patterns in corn belt dent maize," in *Plant Breeding: The Arnel R. Hallauer International Symposium*, eds K. R. Lamkey and M. Lee (Ames, IA: Blackwell Publishing), 219–233. doi: 10.1002/9780470752708.ch16

Trethowan, R. M. (2014). "Defining a genetic ideotype for crop improvement," in *Crop Breeding*, eds D. Fleury and R. Whitford (New York, NY: Humana Press), 1–20. doi: 10.1007/978-1-4939-0446-4_1

Troyer, A. F. (2006). Adaptedness and heterosis in corn and mule hybrids. *Crop Sci.* 46, 528–543. doi: 10.2135/cropsci2005.0065

Troyer, A. F., and Wellin, E. J. (2009). Heterosis decreasing in hybrids: yield test inbreds. *Crop Sci.* 49, 1969–1976. doi: 10.2135/cropsci2009.04.0170

Vacher, M., and Small, I. (2019). Simulation of heterosis in a genome-scale metabolic network provides mechanistic explanations for increased biomass production rates in hybrid plants. *NPJ Syst. Biol. Appl.* 5, 1–10. doi: 10.1016/b978-0-12-817953-6.00001-4

Valluru, R., Gazave, E. E., Fernandes, S. B., Ferguson, J. N., Lozano, R., Hirannaiah, P., et al. (2019). Deleterious mutation burden and its association with complex traits in sorghum (*Sorghum bicolor*). *Genetics* 211, 1075–1087. doi: 10.1534/genetics.118.301742

van Heerwaarden, J., Hufford, M. B., and Ross-Ibarra, J. (2012). Historical genomics of North American maize. *Proc. Natl. Acad. Sci. U.S.A.* 109, 12420–12425. doi: 10.1073/pnas.1209275109

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Varona, L., Legarra, A., Herring, W., and Vitezica, Z. G. (2018a). Genomic selection models for directional dominance: an example for litter size in pigs. *Genet. Select. Evol.* 50:1.

Varona, L., Legarra, A., Toro, M. A., and Vitezica, Z. G. (2018b). Non-additive effects in genomic selection. *Front. Genet.* 9:78. doi: 10.3389/fgene.2018.00078

Viana, J. M. S., Pereira, H. D., Mundim, G. B., Piepho, H. P., and Silva, F. F. (2018). Efficiency of genomic prediction of non-assessed single crosses. *Heredity* 120, 283–295. doi: 10.1038/s41437-017-0027-0

Vitezica, Z. G., Varona, L., Elsen, J. M., Misztal, I., Herring, W., and Legarra, A. (2016). Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genetics Selection Evolution, 48, 1–8. doi: 10.1186/s12711-016-0185-1

Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics* 206, 1297–1307. doi: 10.1534/genetics.116.199406

Voss-Fels, K. P., Cooper, M., and Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theor. Appl. Genet.* 132, 669–686. doi: 10.1007/s00122-018-3270-8

Wallace, J. G., Rodgers-Melnick, E., and Buckler, E. S. (2018). On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu. Rev. Genet.* 52, 421–444. doi: 10.1146/annurev-genet-120116-024846

Wang, C., Liu, Q., Shen, Y., Hua, Y., Wang, J., Lin, J., et al. (2019). Clonal seeds from hybrid rice by simultaneous genome engineering of meiosis and fertilization genes. *Nat. Biotechnol.* 37, 283–286. doi: 10.1038/s41587-018-0003-0

Wang, J., Zhou, Z., Zhang, Z., Li, H., Liu, D., Zhang, Q., et al. (2018). Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity* 121, 648–662. doi: 10.1038/s41437-018-0075-0

Wang, K., Qiu, F., Larazo, W., dela Paz, M. A., and Xie, F. (2015). Heterotic groups of tropical indica rice germplasm. *Theor. Appl. Genet.* 128, 421–430. doi: 10.1007/s00122-014-2441-5

Wang, X., Li, L., Yang, Z., Zheng, X., Yu, S., Xu, C., et al. (2017). Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity* 118, 302–310. doi: 10.1038/hdy.2016.87

Waser, N. M., and Price, M. V. (1994). Crossing−distance effects in *Delphinium nelsonii*: outbreeding and inbreeding depression in progeny fitness. *Evolution* 48, 842–852. doi: 10.2307/2410491

Washburn, J. D., and Birchler, J. A. (2014). Polyploids as a "model system" for the study of heterosis. *Plant Reprod.* 27, 1–5. doi: 10.1007/s00497-013-0237-4

Watson, A., Hickey, L. T., Christopher, J., Rutkoski, J., Poland, J., and Hayes, B. J. (2019). Multivariate genomic selection and potential of rapid indirect selection with speed breeding in spring wheat. *Crop Sci.* 59, 1945–1959. doi: 10.2135/cropsci2018.12.0757

Wei, W. H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nat. Rev. Genet.* 15, 722–733. doi: 10.1038/nrg3747

Welch, J. J. (2004). Accumulating Dobzhansky−Muller incompatibilities: reconciling theory and data. *Evolution* 58, 1145–1156. doi: 10.1554/03-502

Werner, C. R., Gaynor, R. C., Sargent, D. J., Lillo, A., Gorjanc, G., and Hickey, J. M. (2020). Genomic selection strategies for clonally propagated crops. *bioRxiv [Preprint]* doi: 10.1101/2020.06.15.152017

Werner, C. R., Qian, L., Voss-Fels, K. P., Abbadi, A., Leckband, G., Frisch, M., et al. (2018). Genome-wide regression models considering general and specific combining ability predict hybrid performance in oilseed rape with similar accuracy regardless of trait architecture. *Theor. Appl. Genet.* 131, 299–317. doi: 10.1007/s00122-017-3002-5

Westhues, M., Schrag, T. A., Heuer, C., Thaller, G., Utz, H. F., Schipprack, W., et al. (2017). Omics-based hybrid prediction in maize. *Theor. Appl. Genet.* 130, 1927–1939. doi: 10.1007/s00122-017-2934-0

Whittaker, J. C., Thompson, R., and Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genet. Res.* 75, 249–252. doi: 10.1017/s0016672399004462

Wijnker, E., van Dun, K., de Snoo, C. B., Lelivelt, C. L., Keurentjes, J. J., Naharudin, N. S., et al. (2012). Reverse breeding in *Arabidopsis thaliana* generates homozygous parental lines from a heterozygous plant. *Nat. Genet.* 44:467. doi: 10.1038/ng.2203

Williams, W. (1959). Heterosis and the genetics of complex characters. *Nature* 184, 527–530. doi: 10.1038/184527a0

Windhausen, V. S., Atlin, G. N., Hickey, J. M., Crossa, J., Jannink, J. L., Sorrells, M. E., et al. (2012). Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3* 2, 1427–1436. doi: 10.1534/g3.112.003699

Woolliams, J. A., Berg, P., Dagnachew, B. S., and Meuwissen, T. H. E. (2015). Genetic contributions and their optimization. *J. Anim. Breed. Genet.* 132, 89–99. doi: 10.1111/jbg.12148

Wricke, G., and Weber, E (eds) (1986). "Hybrid varieties," in *Quantitative Genetics and Selection in Plant Breeding*, (Berlin: de Gruyter), 257–280.

Xiang, T., Christensen, O. F., Vitezica, Z. G., and Legarra, A. (2016). Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. *Genet. Select. Evol.* 48:92.

Xu, Y., Li, P., Zou, C., Lu, Y., Xie, C., Zhang, X., et al. (2017). Enhancing genetic gain in the era of molecular breeding. *J. Exp. Bot.* 68, 2641–2666. doi: 10.1093/jxb/erx135

Xue, W., Anderson, S. N., Wang, X., Yang, L., Crisp, P. A., Li, Q., et al. (2019). Hybrid decay: a transgenerational epigenetic decline in vigor and viability triggered in backcross populations of teosinte with maize. *Genetics* 213, 143–160. doi: 10.1534/genetics.119.302378

Yang, J., Mezmouk, S., Baumgarten, A., Buckler, E. S., Guill, K. E., McMullen, M. D., et al. (2017). Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *PLoS Genet.* 13:e1007019. doi: 10.1371/journal.pgen.1007019

Yao, H., Gray, A. D., Auger, D. L., and Birchler, J. A. (2013). Genomic dosage effects on heterosis in triploid maize. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2665–2669. doi: 10.1073/pnas.1221966110

Zhao, Y., Li, Z., Liu, G., Jiang, Y., Maurer, H. P., Würschum, T., et al. (2015a). Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15624–15629. doi: 10.1073/pnas.1514547112

Zhao, Y., Mette, M. F., and Reif, J. C. (2015b). Genomic selection in hybrid breeding. *Plant Breed.* 134, 1–10. doi: 10.1111/pbr.12231

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 March 2021
doi: 10.3389/fgene.2021.600040

A Stacking Ensemble Learning Framework for Genomic Prediction

*Mang Liang, Tianpeng Chang, Bingxing An, Xinghai Duan, Lili Du, Xiaoqiao Wang, Jian Miao, Lingyang Xu, Xue Gao, Lupei Zhang, Junya Li and Huijiang Gao**

Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing, China

Machine learning (ML) is perhaps the most useful tool for the interpretation of large genomic datasets. However, the performance of a single machine learning method in genomic selection (GS) is currently unsatisfactory. To improve the genomic predictions, we constructed a stacking ensemble learning framework (SELF), integrating three machine learning methods, to predict genomic estimated breeding values (GEBVs). The present study evaluated the prediction ability of SELF by analyzing three real datasets, with different genetic architecture; comparing the prediction accuracy of SELF, base learners, genomic best linear unbiased prediction (GBLUP) and BayesB. For each trait, SELF performed better than base learners, which included support vector regression (SVR), kernel ridge regression (KRR) and elastic net (ENET). The prediction accuracy of SELF was, on average, 7.70% higher than GBLUP in three datasets. Except for the milk fat percentage (MFP) traits, of the German Holstein dairy cattle dataset, SELF was more robust than BayesB in all remaining traits. Therefore, we believed that SEFL has the potential to be promoted to estimate GEBVs in other animals and plants.

Keywords: ensemble learning, stacking, genomic prediction, machine learning, prediction accuracy

OPEN ACCESS

Edited by:
Waseem Hussain, International Rice Research Institute (IRRI), Philippines

Reviewed by:
Bor-Sen Chen, National Tsing Hua University, Taiwan
Moyses Nascimento, Universidade Federal de Viçosa, Brazil

****Correspondence:***
Huijiang Gao
gaohuijiang@caas.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *28 August 2020*
Accepted: *12 January 2021*
Published: *04 March 2021*

Citation:
Liang M, Chang T, An B, Duan X, Du L, Wang X, Miao J, Xu L, Gao X, Zhang L, Li J and Gao H (2021) A Stacking Ensemble Learning Framework for Genomic Prediction. Front. Genet. 12:600040. doi: 10.3389/fgene.2021.600040

INTRODUCTION

Genomic selection (GS) was first introduced by Meuwissen et al. (2001), by using whole-genome markers' information to predict the genomic estimated breeding values (GEBVs). The first application of GS was on dairy cattle, to improve the selection of better performing genotypes and accelerate the genetic gain by shortening the breeding cycles (Hayes et al., 2009a; Crossa et al., 2017; Tong et al., 2020). After more than 10 years of development, GS has been wildly used in livestock and plant breeding programs with high prediction accuracy (Hayes et al., 2009a; Heffner et al., 2009). Moreover, GS has been applied to improve the prediction of complex disease phenotypes using genotype data (De Los Campos et al., 2010; Menden et al., 2013). However, a critical concern in genomic prediction is the prediction accuracy calculated by the Pearson's correlation between the estimated breeding values and the corrected phenotypes. Therefore, the exploration of more robust genomic prediction methods is a well-identified searched by breeders. In recent years, there was an increasing interest in applying machine learning (ML) to genomic prediction. Machine learning is a computer program which can optimize a performance criterion using training data, making predictions or decisions without being explicitly programmed (Alpaydin, 2020). The excellent predictive ability for complex problems leads ML to be employed in industries requiring high accuracy, e.g., email filtering, face recognition, natural language processing or stock market forecasting. ML has been used in GS and might have the best performance at the interpretation of large-scale genomic data (De Los Campos et al., 2010). González-Camacho et al. (2018)

suggested that ML was a valuable alternative to well-known parametric methods for genomic selection. Montesinos-López et al. (2018) also found that the predictions of the multi-trait deep learning model were very competitive with the Bayesian multi-trait and multi-environment model. In another study, Jubair and Domaratzki (2019) estimated GEBVs of Iranian wheat landraces by ensemble learning, obtaining better results with those than with single machine learning. It is possible to clearly identify a trend from the literature that more breeders are applying machine learning methods to estimate GEBVs in genomic prediction.

Currently, the machine learning methods applied in animal and plant breeding tend to mainly include: support vector regression (SVR), random forest (RF), kernel ridge regression (KRR), multi-layer prediction (MLP) and convolutional neural network (CNN) (Gianola et al., 2011; Libbrecht and Noble, 2015; González-Camacho et al., 2018; Zou et al., 2019). Those machine learning methods possess the ability to predict GEBVs by building a complex non-linear model, considering the interaction effects and epistatic effects (Gianola et al., 2011). Nevertheless, the prediction accuracy of those single machine learning methods did not improve much when compared to the traditional genomic prediction methods [GBLUP, ridge regression BLUP (rrBLUP), BayesB, etc.]. Ogutu et al. (2011) compared the prediction accuracy of RF, boosting and support vector machine (SVM) with rrBLUP in a simulated dataset, in which rrBLUP outperformed the three machine learning methods. When comparing the prediction performance of multi-layer prediction and the SVM with the Bayesian threshold genomic best linear unbiased prediction (TGBLUP), the reliability of two machine learning methods was comparable to, and in some cases, outperformed that of TGBLUP (Montesinos-López et al., 2019). Albeit that the achievement of ML in GS has not been fantastic, breeders are confident on this promising tool. Moreover, even currently associated with certain limitations, it outstands from the other common available methods in the performance.

One of the available solutions to further improve the prediction accuracy of ML in GS is to simultaneously integrate several machine learning methods in genomic prediction. Ensemble learning is an ML paradigm where multiple learners are trained to solve the same problem, therefore, the obtained robustness is higher when compared to that using single learner (Thomas, 1997; Polikar, 2006). Stacking, boosting and bagging were the most common integration strategies used on ensemble learning, among which stacking has a powerful prediction capability for complex problems. In other research areas, stacking has been applied to date prediction, protein-protein interaction prediction, credit scoring, cancer detection, etc. (Wang et al., 2011; Wang Y. et al., 2019; Sun and Trevor, 2018; Yi et al., 2020). However, the application of stacking in GS has rarely been reported.

Therefore, the objective of this study was to improve genomic predictions by using a stacking ensemble learning framework (SELF). In the experiment, SVR, KRR, and ENET were selected as the base learner, and the ordinary least squares (OLS) linear regression was chosen as the meta learner to construct the SELF model. Subsequently, we evaluated the SELF model using two animal datasets (Chinese Simmental beef cattle dataset and German Holstein dairy cattle dataset) and a plant dataset (Loblolly pine dataset). To assess the performance of SELF, we compared the prediction accuracy of SELF with the base learners, GBLUP and BayesB. Finally, the 20-fold cross-validation was employed to mitigate the impact of the accidental error.

MATERIALS AND METHODS

Dataset

Chinese Simmental Beef Cattle Dataset

The Chinese Simmental beef cattle population included 1,217 individuals; born between 2008 and 2014 in Ulgai, Xilingolia of China, and were slaughtered at 16 to18 months. After slaughtering, the carcass trait was assessed according to the institutional meat purchase specifications for fresh beef guidelines. At the present study, three important economic traits were selected for latter analysis: live weight (LW), carcass weight (CW), and eye muscle area (EMA). The statistics description for each trait included an estimation of component variance, which is presented in **Table 1**. The entire Chinese Simmental beef cattle population was genotyped by Illumina® BovineHD BeadChip (770K). The quality control criteria of genotype data were as follows: minor allele frequency (MAF) > 0.05, call rate (CR) > 0.95 and P-value $> 10^{-5}$ from Hardy-Weinberg equilibrium (HWE). In addition, the fix effects were used to correct the phenotypes of each trait. Among them, age and sex were regarded as a contemporary group; the fattening time and initial weight were regarded as covariates.

German Holstein Dairy Cattle Dataset

The dataset of German Holstein dairy cattle consisted of 5,024 bulls with genotypes and phenotypes (Zhang et al., 2015). The genotype data were generated with the Illumina® Bovine SNP50 BeadChip [42,551 single nucleotide polymorphisms (SNPs)]. All of the SNPs met the following standards: HWE P-value $> 10^{-4}$, CR > 0.95 and MAF > 0.01 (Yin et al., 2020). Because the dataset used at the present study was not original, all the phenotype data had been standardized (mean = 0, standard deviation = 1). More details about the original dataset can be found at Zhang et al. (2015). For the German Holstein dairy cattle dataset, the statistics description was based on Zhang et al. (2015) and can be found in **Table 1**. The phenotypes were described by three traits: milk yield (MY), milk fat percentage (MFP) and somatic cell score (SCS). These three traits may represent three genetic architectures of complex traits composed of: (1) one major gene and a large number of small effect loci (MFP), (2) few moderate effect loci and many small effect loci (MY), and (3) many loci with small effects (SCS), respectively (Zhang et al., 2015; Yin et al., 2020).

Loblolly Pine Dataset

The Loblolly pine dataset comprised 951 individuals from 61 families, having 17 traits systemically recorded from each individual (Resende et al., 2012). For the original dataset, all the individuals were genotyped with an Illumina® Iminium assay (7216 SNPs) (Zhang et al., 2015). After quality control,

TABLE 1 | Descriptive statistics of the phenotype data used in the genomic prediction.

Dataset	Trait	N[a]	h^2	Mean	SD
Beef cattle	LW	1216	0.53	505.26	70.76
	CW	1216	0.44	271.36	45.65
	EMA	1117	0.57	85.21	13.32
Dairy cattle	MY	5024	0.95	370.79	641.60
	MFP	5024	0.94	−0.06	0.28
	SCS	5024	0.88	102.32	11.73
Loblolly pine	HT	861	0.31	20.30	73.31
	CWAL	861	0.27	2.44	27.32
	TS	910	0.37	0.10	1.22

N[a], number of the animal with phenotypes; h2, heritability; SD, standard deviation. LW, live weight; CW, carcass weight; EMA, eye muscle area; MY, milk yield; MFP, milk fat percentage; SCS, somatic cell score; HT, total stem height; CWAL, crown width along the planting beds; TS, tree stiffness.

the genotypes contained 4,853 polymorphic SNPs, which were the same as used by Resende et al. (2012) and Zhang et al. (2015). The phenotypes that were used were a subset of the original phenotype data. Within the traits selected, i.e., growth traits (total stem height, HT), development traits (crown width along the planting beds, CWAL) and wood quality traits (tree stiffness, TS), only one trait was chosen to implement prediction models, respectively. The statistics description for the Loblolly pine dataset is shown in **Table 1**.

Stacking

Stacking is a form of meta-learning which can yield impressive results by designing novel deep learning architectures (Kyriakides and Margaritis, 2019). The core idea of stacking is using the base learners to generate metadata for the inputs and then utilize another learner, generally called the meta-learner, to process metadata. Base learners are usually called level 0 learners, the meta learners are called level 1 learners and the meta learners stacked on the based learners are the so-called stacking (Kyriakides and Margaritis, 2019). In genomic prediction, the SELF is performed in two steps: firstly, a series of single machine learning methods are trained to generate metadata using markers' information; secondly, a meta learner are trained to predict GEBVs using metadata. The data flow of SELF for genomic prediction is shown in **Figure 1**.

The base learners employed to construct SELF at present study, involved SVR, KRR and ENET. SVR and KRR construct a non-linear model to predict GEBVs and ENET estimate the GEBVs by building a linear regression. It is important to highlight that the meta learner should be a relatively simple ML algorithm to (1) avoid overfitting and (2) possess the ability to handle correlated inputs with no assumptions about the independence of features. These two factors will be important because the inputs of meta-learner will be highly correlated (Kyriakides and Margaritis, 2019). Taking into account the above requirements, the OLS linear regression was chosen as the meta-learner in the SELF. During the SELF model training, the genotypes were not taken as the direct inputs, instead, it were replaced by the genomic relationship matrix derived from genotypes (Gianola et al., 2011). Although this might reduce the prediction accuracy of a single machine learning method, it would significantly reduce

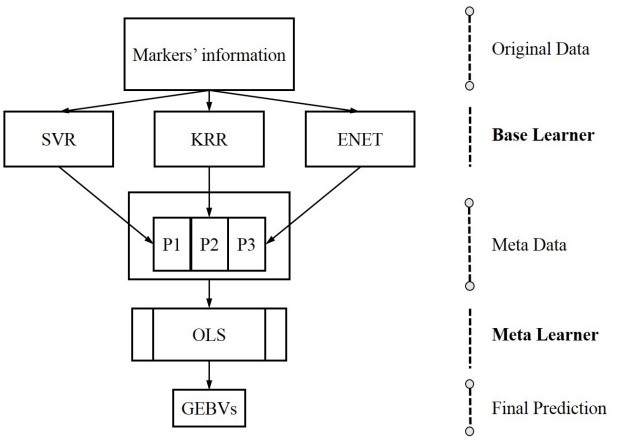

FIGURE 1 | The data flow of stacking ensemble learning framework for genomic prediction, from original data to the base learners, creating metadata for the meta-learner. G, genotypes derived genomic relationship matrix; SVR, support vector regression; KRR, kernel ridge regression; ENET, elastic net; OLS, ordinary least squares linear regression.

the time and the memory required for computation. In theory, the calculation time of SELF will be equivalent to five times of that by a single machine learning method, as five-fold cross-validation was used to generate metadata. It is important to highlight that if the same steps of previous studies were used to apply the genotypes as the inputs, the computation time of SELF would be unacceptable. Finally, SELF was run in Python (V3.7) with the help of *sklearn* (V0.22) package. The genomic relationship matrix G was calculated as described by VanRaden (2008):

$$G = \frac{MM'}{\sum_{l=1}^{m} 2p_j q_j}$$

where M is a n × m matrix (n is the number of individuals, m is the number of markers) and elements of column j in M are $0 - 2p_j$, $1 - 2p_j$ and $2 - 2p_j$ for genotypes A_1A_1, A_1A_2 and A_2A_2; q_j is allele frequency A_1 at locus j, p_j is allele frequency A_2 at locus jth.

Support Vector Regression

Support vector machine (SVM) is grounded in statistical learning theory. SVR is an application of SVM for regression. SVR utilizes a linear or non-linear kernel function to map the original space to a higher dimensional feature space (Müller and Guido, 2016; Li, 2019). Therefore we built a linear prediction model on feature space. The SVR problem was formalized as:

$$\min_{w,b} \frac{1}{2}w^2 + C\sum_{i=1}^{m} L_\varepsilon \left(f\left(x_i\right) - y_i\right)$$

where C is the regularization constant, L_ε is the ε-insensitive loss:

$$L_\varepsilon = \begin{cases} 0 & if\ z < \varepsilon \\ |z| - \varepsilon, & otherwise \end{cases}$$

where $\mathrm{k}\left(x_i,\ x_j\right) = \phi\left(x_i\right)^T \phi\left(x_j\right)$ $\mathrm{z} = f(x_i) - y_i$. Through a series of optimization processes, the SVR can be written as:

$$f(x) = \sum_{i=1}^{m} \left(\widehat{\alpha}_i - \alpha_i\right) k(x, x_i) + b$$

where $\mathrm{k}\left(x_i, x_j\right) = \phi\left(x_i\right)^T \phi\left(x_j\right)$ is the kernel function. In this study, the Gaussian kernel was used to map original data and the most suitable parameters of C and ε for each trait were determined by grid search. The function *SVR* in *sklearn* package (V 0.22) was used to implement SVR methods.

Kernel Ridge Regression

The difference between KRR and ridge regression is that KRR exploits the kernel trick to define a higher dimensional feature space and then builds the ridge regression model in feature space (Douak et al., 2013; He et al., 2014; Exterkate et al., 2016). For KRR, the final prediction function can be written as the following:

$$f\left(x\right) = k^{'}\left(K + \lambda I\right)^{-1} y$$

where K is the so-called gram matrix with entries$K_{ij} = \phi(x_i) \cdot \phi(x_j)$, k is a vector with entries $k_i = \phi(x) \cdot \phi(x_i = k(x, x_i)$ with$i = 1, 2, 3, \ldots,$ n, n is the number of training samples; I is the identity matrix, λis the ridge parameter. In this study, the kernel was used to transform input data that was selected by the grid search method.

Elastic Net

Elastic net is a linear regression model trained with both ℓ_1 and ℓ_2-norm regularization of the coefficients. This combination leads to the ENET, presenting similar advantages when compared to Lasso and ridge regression simultaneously. Thus, ENET can learn a sparse model where few of the weights are non-zero and maintaining the regularization properties (Pedregosa et al., 2011). The progress of training the ENET model can be seen as an optimization process for:

for this study, X $c$is a matrix of the training section of G matrix, ω is the vector of weights, α and ρ are the parameters that determined by grid search.

Genomic Best Linear Unbiased Prediction

The basic GBLUP method was built by the following equation (VanRaden, 2008; Hayes et al., 2009b):

$$y = 1\mu + Zg + e$$

where y is the vector of the correct phenotype, μ is the overall mean, *1* is a vector of ones, Z is a design matrix that allocates records to breeding values, g is a vector of genomic breeding values, e is a vector of residuals. Random residuals were assumed that $e \sim N(0,\ I\sigma_e^2)$ where σ_e^2 is the residual variance, I is an identity matrix. g assumed that $g \sim N(0, G\sigma_g^2)$ where σ_g^2 is the additive genetic variance, and G is the marker-based genomic relationship matrix. To implement GBLUP, we used the *mixed.solve* function of *rrBLUP* package in the R V3.5.

BayesB

BayesB assumed *a priori* that many markers have no effects, while some have an effect attributed to gamma or exponential distribution (Meuwissen et al., 2009). The formula of BayesB can be written as the following:

$$\mathrm{y} = \sum_{j=1}^{p} m_j\alpha_j + e$$

where y is a vector of phenotypes; $m_j$is the jth maker; $\alpha_j$is the effect of the jth maker and$\alpha_j \sim N(0,\ \sigma_{\alpha_j}^2)$. The variance of α_j is assigned an informative before showing the presence (with the probability of $1 - \pi$) and absence (with the probability of π) of the marker j. The π was determined by the experience before building the BayesB model.

Cross-Validation

The prediction accuracy of the machine learning methods, GBLUP and BayesB was evaluated with K-fold cross-validation (CV). Each dataset under study was randomly divided into twenty folds by the 20-fold cross-validation. Each fold would be the testing set and the remaining nineteen folds were grouped into the training set. The training set was used to teach the SELF model how to predict the GEBVs of individuals in the testing set. The accuracy obtained and shown in the result section was the mean of prediction accuracy of each testing set which was measured as the Pearson correlation between the corrected phenotypes (y) and predicted GEBV (y_{pre}) using the formula

$$r = \frac{cov(y,\ y_{pre})}{\sqrt{var\left(y\right) * var(y_{pre})}}$$

RESULTS

Comparison Between the Prediction Accuracy of Base Learners, GBLUP and BayesB

Firstly, we described the prediction accuracy of base learners, GBLUP and BayesB for three datasets, as shown in **Table 2**.

TABLE 2 | Prediction accuracy of SVR, KRR, ENET, GBLUP, and BayesB for the three datasets.

Dataset	Trait	SVR	KRR	ENET	GBLUP	BayesB
Beef cattle	LW	0.274 ± 0.022	**0.283 ± 0.019**	0.276 ± 0.018	0.256 ± 0.017	0.265 ± 0.016
	CW	0.307 ± 0.016	**0.315 ± 0.015**	**0.315 ± 0.017**	0.292 ± 0.014	0.282 ± 0.012
	EMA	0.280 ± 0.025	0.281 ± 0.022	0.285 ± 0.024	0.292 ± 0.015	0.281 ± 0.015
Dairy cattle	MY	0.764 ± 0.013	**0.781 ± 0.009**	0.762 ± 0.014	0.768 ± 0.006	0.767 ± 0.005
	MFP	0.796 ± 0.012	0.828 ± 0.006	0.797 ± 0.012	0.832 ± 0.003	**0.855 ± 0.003**
	SCS	0.706 ± 0.010	0.751 ± 0.008	0.722 ± 0.019	**0.752 ± 0.006**	0.731 ± 0.003
Loblolly pine	HT	0.340 ± 0.027	0.352 ± 0.011	**0.366 ± 0.014**	0.349 ± 0.012	0.365 ± 0.009
	CWAL	0.352 ± 0.022	0.359 ± 0.018	0.369 ± 0.022	0.384 ± 0.014	**0.400 ± 0.011**
	TS	0.397 ± 0.017	0.407 ± 0.016	0.398 ± 0.015	0.366 ± 0.012	**0.418 ± 0.013**

The accuracy was calculated by the Pearson's correlation. LW, live weight; CW, carcass weight; EMA, eye muscle area; MY, milk yield; MFP, milk fat percentage; SCS, somatic cell score; HT, total stem height; CWAL, crown width along the planting beds; TS, tree stiffness. SVR, support vector regression; KRR, kernel ridge regression; ENET, elastic net; GBLUP, genomic best linear unbiased prediction. The bold values mean the highest prediction accuracy for each trait.

BayesB and KRR outperformed other methods in three traits, showing the best predictive power. The prediction accuracy of GBLUP and ENET was higher than that of other methods in two traits. The prediction performance of SVR was the worst, and the prediction accuracy of SVR was always lower than that of the other methods. For base learners, the prediction accuracy of KRR was the highest. The prediction accuracy gap between these methods was not significant, however, the ability to estimate the GEBVs was comparable.

Comparison Between the Prediction Accuracy of SELF and Base Learners

Figure 2 shows the comparison between the prediction accuracy of the base learners and SELF for nine traits. The red one represents the prediction accuracy of SELF. SELF performed better than all the other base learners for each trait. Particularly for CWAL, HT, and EMA, the prediction accuracy of SELF was improved by 9.97, 7.36, and 6.40%, respectively, when compared to the highest prediction accuracy of base learners. Among the three base learners, the prediction ability of KRR was comparable to SELF in German Holstein dairy cattle dataset.

Comparison Between the Prediction Accuracy of SELF, GBLUP and BayesB

Figure 3 demonstrates the prediction accuracy of GBLUP, BayesB and SELF for the three datasets. For the Chinese Simmental beef cattle dataset, the prediction accuracy of SELF was higher than GBLUP and BayesB, showing an average improvement of 11.68% from SELF to GBLUP. For the German Holstein daily cattle, except for the trait of MFP, SELF performed better than BayesB and GBLUP. For the Loblolly pine dataset, SELF predicted GEBVs more accurately than GBLUP and BayesB, showing an improvement of 14.18% for TS, when compared with GBLUP. Comparing the prediction accuracy between SELF and GBLUP, the average prediction accuracy of SELF was increased by 7.70% in nine traits.

DISCUSSION

The previous large number of studies had tried to apply single machine learning methods into genomic prediction (Long et al., 2011; Jubair and Domaratzki, 2019; Montesinos-López et al., 2019; Lenz et al., 2020). However, the single machine learning methods applicatied in most of the previous studies, only performed well on certain traits (Long et al., 2011; Ogutu et al., 2011; González-Camacho et al., 2018; Montesinos-López et al., 2019). Therefore, the present study proposed a new strategy to utilize machine learning methods in genomic prediction by using a stacking ensemble learning framework integrating three machine learning methods to predict GEBVs simultaneously. To examine the prediction ability of SELF, we compared the prediction accuracy of SELF with GBLUP and BayesB in animal and plant datasets with a variety of genetic architecture. Considering the computation time and that overfitting was employed, the genotypes derived relationship matrix as the inputs rather than using the genotypes directly (Gianola et al., 2011).

The Prediction Accuracy of Base Learners, GBLUP, and BayesB

Using GBLUP and BayesB to predict GEBV for the three dataset had been reported early which provided a reference for verifying our results. Therefore, this study compared the prediction accuracy of GBLUP and BayesB with the prediction accuracy obtained from Wang X. et al. (2019), Zhang et al. (2015), and Resende et al. (2012). Wang X. et al. (2019) compared GBLUP with BayesB in the Chinese Simmental beef cattle dataset. Zhang et al. (2015) and Resende et al. (2012) compared the prediction accuracy of different methods on the German Holstein dairy cattle dataset and the Loblolly pine dataset, respectively. Overall, the results were consistent. Since the method was slightly different from that was used in the previous studies, the accuracy differed in individual traits. Although, the application of a single machine learning method to estimate GEBVs on the three datasets has not been reported, the vast majority of studies has compared the prediction accuracy of the single machine learning method with GBLUP or Bayesian family methods on other populations. Therefore, it provided a practical reference

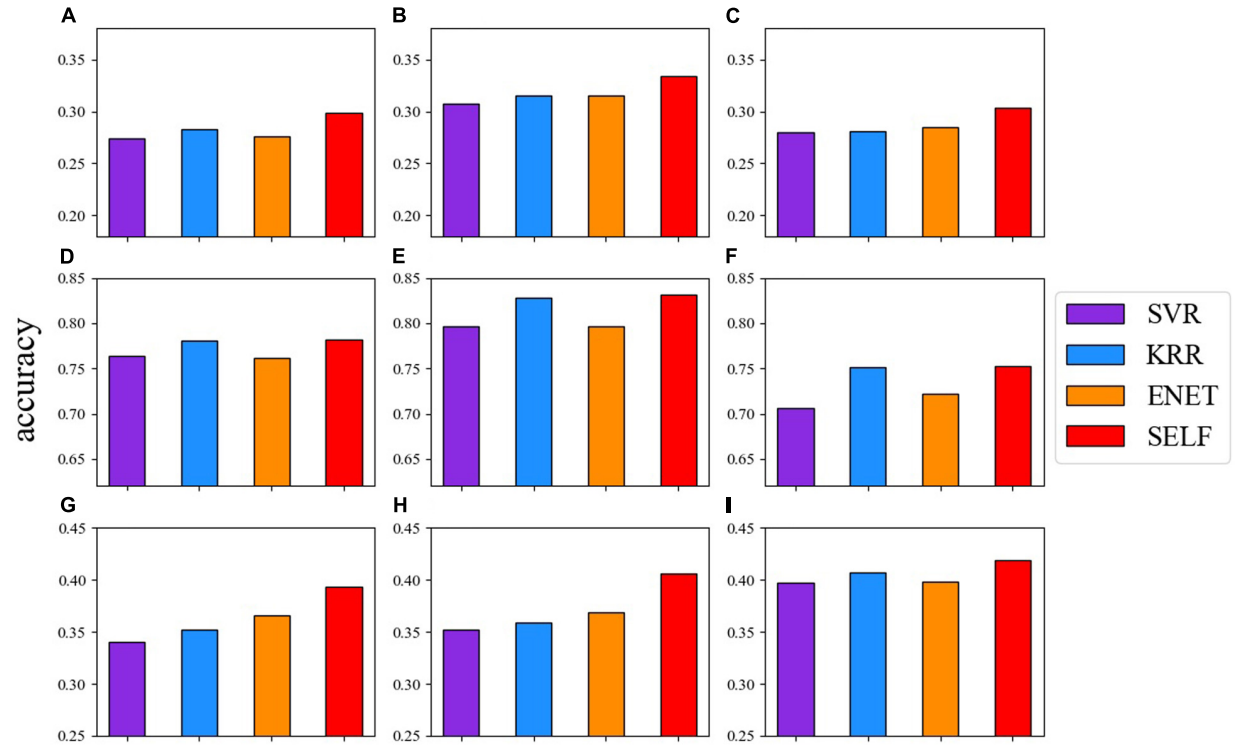

FIGURE 2 | Comparison of the prediction accuracy among: SVR (blue violet), KRR (dodger blue), ENET (dark orange) and SELF for nine traits. **(A)** live weight; **(B)** carcass weight; **(C)** eye muscle area; **(D)** milk yield; **(E)** milk fat percentage; **(F)** somatic cell score; **(G)** total stem height; **(H)** crown width along the planting beds; **(I)** tree stiffness.

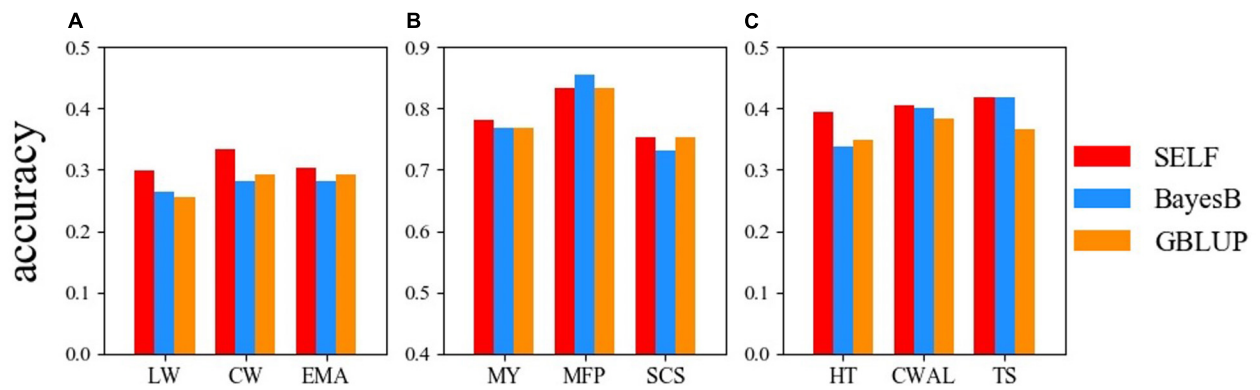

FIGURE 3 | Comparison of the prediction accuracy among: SELF (red), GBLUP (dodger blue) and BayesB (dark orange) for three datasets. **(A)** Chinese Simmental beef cattle dataset; **(B)** German Holstein dairy cattle dataset; **(C)** Loblolly pine dataset. LW, live weight; CW, carcass weight; EMA, eye muscle area; MY, milk yield; MFP, milk fat percentage; SCS, somatic cell score; HT, total stem height; CWAL, crown width along the planting beds; TS, tree stiffness. GBLUP, genomic best linear unbiased prediction; SELF, a stacking ensemble learning framework.

when evaluating the performance of single machine learning methods. The results of Ghafouri-Kesbi et al. (2017) and Long et al. (2011) indicated that GBLUP presented better prediction accuracy when compared to SVR and RF. Furthermore, in most cases, the performance of SVR with Gaussian kernel was comparable to that of Bayesian Lasso (Long et al., 2011; Ghafouri-Kesbi et al., 2017). Similar to previously reported studies, the results from the present study also confirmed that single machine learning did not perform significantly better than GBLUP and Bayes methods.

Excellent Predictive Performance of SELF

Compared to GBLUP, the average prediction accuracy of SELF was increased by 7.70% for the nine traits, which is significant

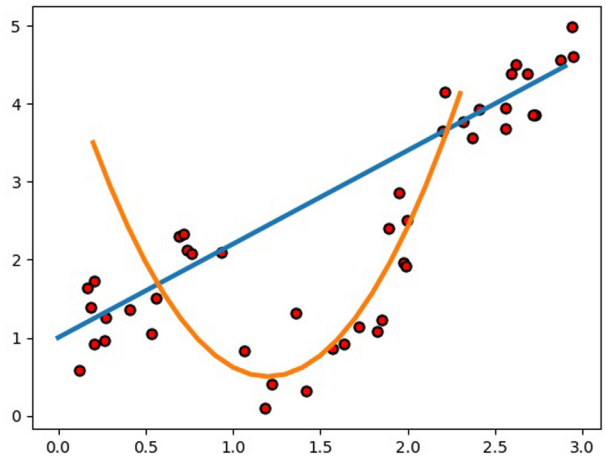

FIGURE 4 | Example of how an integrated model using a stacked ensemble with linear and non-linear regression can significantly outperform either a single linear or non-linear model.

for animal and plant breeding. Particularly for the beef cattle with a longer generation intervals, such considerable prediction accuracy improvement will greatly accelerate the genetic gain. Actually, it is very difficult to build a SELF model to predict a specific problem with higher accuracy, since the composition of SELF model is so flexible. Therefore, the present study referred to previous studies that using machine learning methods to estimated GEBVs, and combined with our experience to select the candidate base learner. Besides, a single-layer framework or multi-layer framework also should be premeditated carefully when constructing frameworks. Considering the overfitting always accompanied by the machine learning methods in GS and the calculating time of SELF, we determined a single layer stacking framework. Before constructing the model of SELF, RF, SVR, KRR, and ENET were chosen as the candidates for base learners, in which RF and SVR had been performed to predict GEBV in previous studies (Long et al., 2011; Ogutu et al., 2011; González-Recio et al., 2014; Libbrecht and Noble, 2015; Ghafouri-Kesbi et al., 2017). Although the utilization of KRR in genomic prediction had been rarely reported, it was frequently utilized to classification and regression task for other research areas (Douak et al., 2013; Avron et al., 2017; Chang et al., 2017; Naik et al., 2018). In addition, ENET was chosen to achieve more diversification of SELF model due to the reason that SVR, RF, and KRR predicted GEBV by building a non-linear model and ENET was a liner model (Wang Y. et al., 2019). Subsequently to the prediction of GEBVs using four base learners, we decided to exclude RF from the SELF, because RF greatly increased the computation time of SELF. Consequently, the final SELF model was constructed by SVR, KRR and ENET, in which the base learners were used to build different types of models to estimate the GEBVs. Generally, it was reasonable to employ different learning algorithms to explore the relationship between the feature and the target variable (Kyriakides and Margaritis, 2019). For the regression example (**Figure 4**), we used a stacked ensemble with linear and non-linear regression, showing the possibility to significantly outperform either a single linear or non-linear model. Even though we directly utilized the best prediction of the linear and non-linear models as the outputs of the integrated model without stacking, the performance of the integrated model was greatly improved. Therefore, the constructed SELF could learn more characteristics in different aspects of the input data, and it performed better than either of the base learners.

Besides, the form of input data in this study might be another momentous reason contributed to the higher prediction accuracy of SELF model. The majority of published studies directly employed genotypes as the inputs of machine learning methods. Nevertheless, the number of markers was considerably larger than the number of individuals. In this case, if we used genotypes with no transformed, the number of variables in the prediction model would be an astronomical figure compared to group size. Despite that single machine learning methods were able to solve the problem of "big P and small N," stronger overfitting was inevitable, which also decreased the prediction accuracy of the SELF. The application of genomic relationship matrix as the input data was completely different, as the genomic relationship matrix was a n × n matrix, whose size is determined by the group sizen. Therefore, the number of variables in the prediction model would be consistent with the number of individuals. Although it might reduce the prediction accuracy of the base learners, it simultaneously and dramatically reduces the risk of overfitting, which potentially improves the prediction accuracy of the SELF.

CONCLUSION

The present study proposes a stacking ensemble learning framework integrating SVR, KRR, and ENET to predict GEBVs. The excellent performance of SELF in a variety of genetic architecture datasets indicates that SELF possesses a significant potential to improve genomic predictions in other animal and plant populations.

DATA AVAILABILITY STATEMENT

Chinese Simmental Beef Cattle dataset: Data is available from the Dryad Digital Repository: doi: 10.5061/dryad.4qc06. German Holstein dairy cattle dataset: Data can be obtained at: https://www.g3journal.org/content/5/4/615.supplemental. Loblolly pine dataset: The quality-controlled genotypes can be gotten at: https://www.genetics.org/highwire/filestream/412827/field_highwire_adjunct_files/1/FileS1.zip and the complete phenotypes at: https://www.genetics.org/highwire/filestream/412827/field_highwire_adjunct_files/4/FileS4.xlsx.

ETHICS STATEMENT

The animal study was reviewed and approved by the Science Research Department of the Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing, China (approval number: RNL09/07).

AUTHOR CONTRIBUTIONS

HG and JL conceived and designed the study. ML and BA performed statistical analyses and wrote the manuscript. ML, JM, and XW wrote the code. TC, BA, XD, LD, and JM participated in data analyses. LZ, LX, and XG participated in the design of the study and contributed to acquisition of data. All authors read and approved the final manuscript.

FUNDING

This work was supported by funds from the National Natural Science Foundation of China (31872975) and the Program of National Beef Cattle and Yak Industrial Technology System (CARS-37).

ACKNOWLEDGMENTS

The authors would like to thank HG for guidance and for proofreading the manuscript. Furthermore, the authors acknowledge all staff at the experimental cattle unit in Beijing for supporting animal care and sample collection during experimental period.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.600040/full#supplementary-material

REFERENCES

Alpaydin, E. (2020). *Introduction to Machine Learning*. Cambridge, MA: MIT press.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. (2017). "Random fourier features for kernel ridge regression: approximation bounds and statistical guarantees," in *International Conference on Machine Learning*, (Sydney, Australia), 253–262.

Chang, X., Lin, S.-B., and Zhou, D.-X. (2017). Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res.* 18, 1493–1514.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de Los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975.

De Los Campos, G., Gianola, D., and Allison, D. B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.* 11, 880–886. doi: 10.1038/nrg2898

Douak, F., Melgani, F., and Benoudjit, N. (2013). Kernel ridge regression with active learning for wind speed prediction. *Appl. Energy* 103, 328–340. doi: 10.1016/j.apenergy.2012.09.055

Exterkate, P., Groenen, P. J., Heij, C., and van Dijk, D. (2016). Nonlinear forecasting with many predictors using kernel ridge regression. *Int. J. Forecast.* 32, 736–753. doi: 10.1016/j.ijforecast.2015.11.017

Ghafouri-Kesbi, F., Rahimi-Mianji, G., Honarvar, M., and Nejati-Javaremi, A. (2017). Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. *Anim. Prod. Sci.* 57, 229–236. doi: 10.1071/AN15538

Gianola, D., Okut, H., Weigel, K. A., and Rosa, G. J. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12:87. doi: 10.1186/1471-2156-12-87

González-Recio, O., Rosa, G. J., and Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest. Sci.* 166, 217–231. doi: 10.1016/j.livsci.2014.05.036

González-Camacho, J. M., Ornella, L., Pérez-Rodríguez, P., Gianola, D., Dreisigacker, S., and Crossa, J. (2018). Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *Plant Genome* 11, 1–15.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009a). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91, 47–60. doi: 10.1017/S0016672308009981

He, J., Ding, L., Jiang, L., and Ma, L. (2014). "Kernel ridge regression classification," in *2014 International Joint Conference on Neural Networks (IJCNN)*, Piscataway, NJ: IEEE, 2263–2267.

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Jubair, S., and Domaratzki, M. (2019). "Ensemble supervised learning for genomic selection," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Piscataway, NJ: IEEE, 1993–2000. doi: 10.1109/BIBM47256.2019.8982998

Kyriakides, G., and Margaritis, K. G. (2019). *Hands-On Ensemble Learning with Python*. Sebastopol, CA: O REILLY.

Lenz, P. R., Nadeau, S., Mottet, M. J., Perron, M., Isabel, N., Beaulieu, J., et al. (2020). Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce. *Evol. Appl.* 13, 76–94.

Li, H. (2019). *Statistical Learning Methods*, 2nd Edn. Beijing: Tsinghua University Press.

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Long, N., Gianola, D., Rosa, G. J., and Weigel, K. A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet.* 123:1065. doi: 10.1007/s00122-011-1648-y

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., et al. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* 8:e61318. doi: 10.1371/journal.pone.0061318

Meuwissen, T. H., Solberg, T. R., Shepherd, R., and Woolliams, J. A. (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Select. Evol.* 41:2.

Meuwissen, T. H. E., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C. M., and Martín-Vallejo, J. (2018) Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits *G3: Genes, Genomes, Genetics* 8, 3829–3840. doi: 10.1534/g3.118.200728

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., et al. (2019). A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding *G3 Genes Genomes Genet.* 9, 601–618.

Müller, A. C., and Guido, S. (2016). *Introduction to Machine Learning With Python: a Guide for Data Scientists*. Sebastopol, CA: O'Reilly Media, Inc.

Naik, J., Satapathy, P., and Dash, P. (2018). Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression. *Appl. Soft Comput.* 70, 1167–1188. doi: 10.1016/j.asoc.2017.12.010

Ogutu, J. O., Piepho, H.-P., and Schulz-Streeck, T. (2011). "A comparison of random forests, boosting and support vector machines for genomic selection," in *BMC Proceedings*, Vol. S3, BioMed Central, 1–5

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and Thirion, B. (2011). Scikit-learn: machine learning in python the journal of machine. *Learn. Res.* 12, 2825–2830.

Polikar, R. (2006). Ensemble based systems in decision. *Making IEEE Circ. Syst. Mag.* 6, 21–45. doi: 10.1109/MCAS.2006.1688199

Resende, M. F., Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., et al. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics* 190, 1503–1510.

Sun, W., and Trevor, B. (2018). A stacking ensemble learning framework for annual river ice breakup dates. *J. Hydrol.* 561, 636–650. doi: 10.1016/j.jhydrol.2018.04.008

Thomas, G. D. (1997). Machine learning research: four current directions Artificial Intelligence. *Magazine* 18, 97–136.

Tong, H., Küken, A., and Nikoloski, Z. (2020). Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth. *Nat. Commun.* 11, 1–9. doi: 10.1038/s41467-020-16279-5

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Wang, G., Hao, J., Ma, J., and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Syst. Appl.* 38, 223–230.

Wang, X., Miao, J., Chang, T., Xia, J., An, B., Li, Y., et al. (2019). Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese Simmental beef cattle. *PLoS One* 14:e0210442. doi: 10.1371/journal.pone.0210442

Wang, Y., Wang, D., Geng, N., Wang, Y., Yin, Y., and Jin, Y. (2019). Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Appl. Soft Comput.* 77, 188–204.

Yi, H.-C., You, Z.-H., Wang, M.-N., Guo, Z.-H., Wang, Y.-B., and Zhou, J.-R. (2020). RPI-SE: a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information. *BMC Bioinformatics* 21:60. doi: 10.1186/s12859-020-3406-0

Yin, L., Zhang, H., Zhou, X., Yuan, X., Zhao, S., Li, X., et al. (2020). KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. *Genome Biol.* 21, 1–22.

Zhang, Z., Erbe, M., He, J., Ober, U., Gao, N., Zhang, H., et al. (2015). Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix *G3 Genes Genomes Genet.* 5, 615–627.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. *Nat. Genet.* 51, 12–18.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 31 March 2021
doi: 10.3389/fgene.2021.643733

Improving Genomic Prediction for Seed Quality Traits in Oat (Avena sativa L.) Using Trait-Specific Relationship Matrices

Malachy T. Campbell[1], Haixiao Hu[1], Trevor H. Yeats[1], Lauren J. Brzozowski[1], Melanie Caffe-Treml[2], Lucía Gutiérrez[3], Kevin P. Smith[4], Mark E. Sorrells[1], Michael A. Gore[1] and Jean-Luc Jannink[1,5]*

[1] Plant Breeding & Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, United States, [2] Seed Technology Lab 113, Agronomy, Horticulture & Plant Science, South Dakota State University, Brookings, SD, United States, [3] Department of Agronomy, University of Wisconsin-Madison, Madison, WI, United States, [4] Department of Agronomy & Plant Genetics, University of Minnesota, St. Paul, MN, United States, [5] R.W. Holley Center for Agriculture & Health, US Department of Agriculture, Agricultural Research Service, Ithaca, NY, United States

The observable phenotype is the manifestation of information that is passed along different organization levels (transcriptional, translational, and metabolic) of a biological system. The widespread use of various omic technologies (RNA-sequencing, metabolomics, etc.) has provided plant genetics and breeders with a wealth of information on pertinent intermediate molecular processes that may help explain variation in conventional traits such as yield, seed quality, and fitness, among others. A major challenge is effectively using these data to help predict the genetic merit of new, unobserved individuals for conventional agronomic traits. Trait-specific genomic relationship matrices (TGRMs) model the relationships between individuals using genome-wide markers (SNPs) and place greater emphasis on markers that most relevant to the trait compared to conventional genomic relationship matrices. Given that these approaches define relationships based on putative causal loci, it is expected that these approaches should improve predictions for related traits. In this study we evaluated the use of TGRMs to accommodate information on intermediate molecular phenotypes (referred to as endophenotypes) and to predict an agronomic trait, total lipid content, in oat seed. Nine fatty acids were quantified in a panel of 336 oat lines. Marker effects were estimated for each endophenotype, and were used to construct TGRMs. A multikernel TRGM model (MK-TRGM-BLUP) was used to predict total seed lipid content in an independent panel of 210 oat lines. The MK-TRGM-BLUP approach significantly improved predictions for total lipid content when compared to a conventional genomic BLUP (gBLUP) approach. Given that the MK-TGRM-BLUP approach leverages information on the nine fatty acids to predict genetic values for total lipid content in unobserved individuals, we compared the MK-TGRM-BLUP approach to a multi-trait gBLUP (MT-gBLUP) approach that jointly fits phenotypes for fatty acids and total lipid content. The MK-TGRM-BLUP approach significantly outperformed MT-gBLUP. Collectively, these results highlight the utility of using TGRM to accommodate information on endophenotypes and improve genomic prediction for a conventional agronomic trait.

Keywords: genomic prediction, Bayesian regression, lipids, metabolomics, genomics, plant breeding, oats

OPEN ACCESS

Edited by:
Diego Jarquin, University of Nebraska-Lincoln, United States

Reviewed by:
Palle Duun Rohde, Aalborg University, Denmark
Kui Zhang, Michigan Technological University, United States

****Correspondence:***
Malachy T. Campbell
campbell.malachy@gmail.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 December 2020*
Accepted: *04 March 2021*
Published: *31 March 2021*

Citation:
*Campbell MT, Hu H, Yeats TH, Brzozowski LJ, Caffe-Treml M, Gutiérrez L, Smith KP, Sorrells ME, Gore MA and Jannink J-L (2021) Improving Genomic Prediction for Seed Quality Traits in Oat (*Avena sativa *L.) Using Trait-Specific Relationship Matrices. Front. Genet. 12:643733. doi: 10.3389/fgene.2021.643733*

1. INTRODUCTION

The observable phenotype is the manifestation of numerous biological process that occur across organizational levels (DNA, transcript, protein, and metabolite) in the plant. In the last 20 years significant progress has been made to query phenotypes at these levels and elucidate the molecular mechanisms (e.g., regulatory networks, biochemical pathways, and physiological mechanisms) that shape variation in conventional traits like plant architecture, disease resistance, productivity and fitness. Omics technologies have provided a means to query the phenotypic space at a molecular level and quantify these phenotypes across organizational levels and query these mechanisms in large populations that are typically required in genetic studies. The term "endophenotype" has been coined to describe these molecular phenotypes (Kremling et al., 2019). Nonetheless, efficiently leveraging these resources to improve prediction of the classical traits that are typically the focus of breeding programs remains a significant challenge.

The widespread use of various omics technologies has motivated many studies to develop approaches that integrate these data types to predict complex traits (Rincent et al., 2018; Morgante et al., 2020). Dense omics data can be used to create relationship matrices, much like genomic relationship matrices, that describe the relatedness among individuals based on the endophenotypes. Best linear unbiased prediction (BLUP) frameworks can then be used to predict complex traits using these kernels. Using these frameworks, Morgante et al. (2020) showed that BLUP models that included relationship matrices derived from transciptome data, as well as transcriptome and genome-wide marker data improved prediction accuracies compared to models that used only genome-wide markers. Several other studies have reported similar improvements in prediction accuracies when omics-based kernels are used for prediction, suggesting that these omics-based kernels capture some component of the phenotype that is not explained by genome-wide markers (environmental or non-additive genetic variance) (Westhues et al., 2017; Rincent et al., 2018; Schrag et al., 2018; Krause et al., 2019; Li et al., 2019; Rohde et al., 2020; Zhou et al., 2020). Despite these promising studies, these improv2gfgements seem to be dependent on the trait, methodologies and datatype (Guo et al., 2016; Schrag et al., 2018; Zhou et al., 2020). Moreover, these approaches require omics phenotypes for all individuals, which would be a burden for many plant breeding programs due to the cost of growing-out and quantifying endophenotypes on these materials.

Information flows from the genotypic space to endophenotypes and eventually to the focal trait. Given this relationship, rather than using these data to create omics-based relationship matrices, knowledge about quantitative trait loci (QTL) that affect these endophenotypes can instead be directly introduced into the prediction frameworks. Predictions for the focal traits should be improved by allowing variance components to be estimated separately for putative functional (causal loci and markers in linkage with these loci) and non-functional markers. This approach would also remove the requirement to have endophenotypes measured on the population used for prediction. Of course, this assumes that effects will be somewhat consistent across populations and locations, and does not account for genotype-by-environment effects. Several studies have used domain/prior knowledge to partition genomic markers into potentially functional (associated with endophenotypes or proximal to causal genes) and non-functional sets (Gusev et al., 2014; Speed and Balding, 2014; Edwards et al., 2016; MacLeod et al., 2016; Xiang et al., 2019). The limitation with these approaches is that they require a means to link endophenotypes to the genome, whether that is through association or linkage mapping or physical positions in the genome, thus favoring traits with simple genetic architecture and large-effect QTL. Since many traits of agronomic importance follow a complex genetic architecture, this approach is somewhat limiting for research programs in plant genetics.

An alternative to these set-based genomic prediction approaches is to use estimated marker effects to construct trait-specific genomic relationship matrices (TGRM). Unlike the genomic relationship matrices defined by VanRaden (2008), which assume that the trait is affected by many small effect loci distributed throughout the genome, TGRMs differentially weight markers according to their effects on the phenotype (Zhang et al., 2010; Sun et al., 2012; de los Campos et al., 2013; Karaman et al., 2018; Gianola et al., 2020; Turner-Hissong et al., 2020). Given this differential weighting, TGRM should better reflect the relationships between individuals at causal, or potentially casual loci.

Zhang et al. (2010) used a two-step approach where marker effects are predicted using Bayes B or Ridge Regression and each marker is weighted by its corresponding genetic variance (in Ridge Regression markers have the same variance) when constructing the relationship matrices. The authors simulated traits controlled by 50 QTL of varying effect sizes, and showed that genomic predictions using the TGRM outperformed conventional genomic prediction approaches that assume an infinitesimal architecture (i.e., genomic BLUP and Ridge Regression), but performed slightly worse than a genomic prediction model that better accommodates large effect QTL (i.e., Bayes B). The results from this early study highlighting the potential benefits of using TGRMs has been supported by several more recent studies (Su et al., 2014; Tiezzi and Maltecca, 2015; Ren et al., 2020). The advantages of these approaches is that information on endophenotypes can be transferred to new populations through marker effects, eliminating the need to quantify endophenotypes in these populations as required for approaches that directly use these data to develop relationship matrices.

These statistical frameworks that use TGRM offer opportunities to improve selection for conventional traits by including genetic effects for related endophenotypes. In this study, we evaluated the potential of TGRM to improve genomic prediction of seed composition traits in oat. We measured endophenotypes in a large diverse population, allowing inferences on these endophenotypes to be leveraged to improve predictions for related phenotypes in new populations. The abundances of nine fatty acid methyl esters were quantified in the mature seed of 336 oat lines using gas chromatography-mass

spectrometry (GC-MS). These data were used to estimate marker effects for TGRMs using five Bayesian regression approaches: Bayesian ridge regression, Bayes A, Bayes B, Bayes Cπ, and Bayesian LASSO. Two datasets were used for validation. The first dataset consists of fatty acid abundances measured on an independent population of 213 elite oat lines. The second study quantified seed protein and lipid content using near-infrared spectroscopy (NIRS) in 210 elite oat lines. These datasets allow us to answer two questions: (1) Are estimated marker effects consistent across populations? (2) Can predictions for a trait be improved by using TGRM for component traits (i.e., endophenotypes)? The utility of these TGRM prediction frameworks is demonstrated through comparisons with single-trait genomic best linear unbiased prediction (gBLUP) and multi-trait gBLUP approaches (MT-gBLUP). This work broadly tests if endophenotype relationships are transferable between populations. Further, it assesses the efficiency of endophenotyping for plant breeding: the cost of such phenotyping will make it efficient only if knowledge obtained from core populations can be transferred to multiple breeding populations.

2. MATERIALS AND METHODS

2.1. Plant Materials

This study used three datasets. The first dataset was used to infer marker effects for nine fatty acids. These data consist of fatty acid phenotypes measured on an oat diversity panel of 375 lines derived from breeding programs in North America and Europe. We refer to this panel as the "Diversity Panel." The Diversity Panel was grown in an augmented field design in Ithaca, NY, in 2018. A total of 368 unreplicated entries were randomly allocated to 18 blocks with 21–23 plots per block. One primary check, "Corral," and one of six secondary checks were included in each of the blocks. These secondary checks were replicated four times in total, while the primary check was replicated 19 times (one block had two "Corral" plots). A total of 336 lines with genotypic data were used for downstream analyses.

The second dataset consists of fatty acid measurements on 227 lines from a second oat panel, and was used to validate marker effects estimated in the Diversity Panel. This panel is constructed from breeding materials and varieties that were used to develop oat varieties for the northern Midwestern United States, which will be referred to as the "Elite Panel" throughout the remainder of this manuscript. The panel was grown in three locations (Crookston, MN; Volga, SD; and Madison, WI) using an augmented block design. Each experiment included 220–224 unreplicated entries and three check lines.

The third experiment measured total lipid content using Near Infrared Spectroscopy (NIRS) in six trials for 210 lines in the Elite Panel. The experiments followed an augmented block design. Entry means were downloaded from the Triticeae Toolbox (Blake et al., 2016). Links to each trial are provided in **Supplementary Table 1**.

2.2. Genotyping and Marker Imputation

Single-nucleotide polymorphism (SNP) data were collected from 11 genotyping experiments for 539 lines (Campbell et al., 2020). The `glmnet` approach was used to impute missing marker data (Chan et al., 2016). Markers were excluded based on the following criteria before performing imputation: allele frequency < 0.02, proportion of missing data across individuals > 0.6, and heterozygosity > 0.1. Individuals where more than 70% of markers were missing or more than 10% of the markers were heterozygous were removed. Genotypic data for individuals in each study were extracted from these data, and markers with a minor allele frequency < 0.05 were removed. This resulted in a total of 62,002 markers used to estimate marker effects for fatty acid traits in the Diversity Panel, 58,123 markers used for prediction of fatty acid phenotypes in the Elite Panel, and 54,220 markers used to predict lipid content measured via NIRS in the Elite Panel.

2.3. Metabolite Profiling for Fatty Acid Methyl Esters (FAME)

The following protocol was used for all experiments that measured fatty acid phenotypes. The methods are described in detail in Campbell et al. (2020) and Carlson et al. (2019). Briefly, dehulled seeds were homogenized, and 100 mg of pulverized tissue was used to separate polar and non-polar compounds using a biphasic extraction method. A set of quality control (QC) samples was created by combining 60 μL of the upper organic layer from each sample, as well as 60 μL of the lower aqueous phase. A total of 600 μL of the upper organic layer was transferred to new glass vials and was dried under nitrogen gas overnight. Organic fractions were re-suspended in 0.7 mL of 50% methanol 50% methyl tert-butyl ether and a 70 μL aliquot was transferred to a 2 mL glass vial. Solvent was completely removed by nitrogen evaporation at ambient temperature. To the dry sample, 100 μL of toluene containing 2.5 mg/mL of internal standard, glyceryl triheptadecanoate, and 200 μL of 3N methanolic HCl were added. The mixture was incubated at 60°C for 1 h, and 0.5 mL of hexane and 700 μL of water were added to the cooled sample. The samples were vortexed, centrifuged at 2,000 rpm for 5 min at 4°C, and the upper hexane layer was diluted 2× with 100% hexane.

One micro-liter of the upper hexane layer containing FAME was injected into a TG-WAXMS column (30mm × 0.25 mm × 0.25 μm, Thermo Scientific) in a Trace1310 GC (Thermo Scientific) coupled to a Thermo Scientific ISQ-LT mass spectrometer. The injector temperature was 260°C, and the split ratio was 15:1. A constant flow rate of the carrier gas (He) was controlled at 1.2 mL $\cdot$ min^{-1}. The initial oven temperature was 200°C and held for 1 min, then increased to 260°C at 10°C$\cdot$min^{-1} and held for 3 min. Detection was completed under electron impact mode, with a scan range of 50–650 amu and scan rate 5 scans$\cdot$s^{-1}. Transfer line and source temperature were both at 250°C. Data processing was completed with Chromeleon 7 software (Thermo Scientific). QC sample were injected after every 6 samples. Standard curves for C14:0, C16:1, C16:0, C18:0, C18:1, C18:2, C18:3, C20:0, and C20:1 were established.

2.4. Calculation of Best Linear Unbiased Predictors for FAMEs

Best linear unbiased predictors (BLUPs) were calculated to remove systematic effects for each fatty acid phenotype. Given that both experiments that quantified fatty acids followed the same type of experimental design (augmented block), the linear mixed model is nearly identical and is given by

$$y = \mu + DTH + check + new : entry + block + batch + e \quad (1)$$

where *check* is a fixed effect for each of the check varieties; *new* is an indicator variable where 0 indicates a check variety and 1 indicates an unreplicated entry, and is nested within entry; *DTH* is a fixed covariate that provides days to heading for each observation; *block* and *batch* are random effects to account for field blocks and injection batch for GC-MS, respectively. Heading dates were only available for the experiments performed in Ithaca, so the linear model used to compute BLUPs for fatty acid phenotypes in the Elite Panel did not include this term. The terms μ and e represent the overall mean and the vector of residuals, respectively. We assume entries are unrelated in this step. The above model was fitted using the `sommer` package in R (Covarrubias-Pazaran, 2016). Deregressed BLUPs for each entry i and fatty acid j were calculated following Edriss et al. (2017) using

$$\hat{g}^*_{ij} = \frac{\hat{g}_{ij}}{1 - \frac{PEV_{ij}}{\sigma^2_{g_j}}} \quad (2)$$

where $\hat{g}_{ij}$ is the BLUP for entry i and metabolite j, PEV_{ij} is the prediction error variance, and $\sigma^2_{g_j}$ is the total genetic variance.

2.5. Prediction of Marker Effects for Fatty Acid Traits

Five Bayesian whole-genome regression approaches were used to estimate marker effects for each of the fatty acid phenotypes. The linear model for all approaches is identical. The methods differ in how the priors for the marker effects are defined. The linear model is

$$\mathbf{y} = \mu + \sum_{p=1}^{P} w_p a_p + \mathbf{e} \quad (3)$$

where w_p is a vector of allele dosages for marker p and a_p is the corresponding additive genetic effect, $\mathbf{y}$ is a vector of fatty acid phenotypes (endophenotypes), and $\mathbf{e}$ is a vector of residuals. In all cases, we assume $\mathbf{e} \sim N(0, \sigma^2_e)$. This linear model was fitted using the BGLR package in R using 20,000 iterations for the Gibbs sampler and the first 5,000 samples were discarded (Pérez and de Los Campos, 2014). Every fifth sample was used to compute the posterior means of marker effects.

The five Bayesian approaches use different prior distributions for the marker effects and are described in detail in Meuwissen et al. (2001) and Gianola (2013). Briefly, Bayesian Ridge Regression (BRR) is analogous to genomic BLUP (gBLUP) and samples marker effects from a Normal distribution. In Bayes A, marker effects are sampled from a scaled-t density, allowing differential shrinkage of marker effects. Scale-mixture densities are used as priors for Bayes B and Bayes Cπ. Some effects are sampled from a point mass at zero and others are sampled from a scaled-t density, as is the case in Bayes B, or a Normal distribution in Bayes Cπ. The mixing parameter specifies the probability of a marker being sampled from either density and is treated as an unknown in implementations of Bayes B and Bayes Cπ used in this study (Pérez and de Los Campos, 2014). Markers are sampled from a point mass at zero with a probability π and a non-zero density with probability $(1 - \pi)$. Thus, in the extreme case where $\pi = 0$ Bayes B will behave like Bayes A and Bayes Cπ will behave similar to BRR. Bayesian LASSO (BL) samples marker effects from a LaPlace density. This prior has thicker tails compared to the Normal density used in BRR, but will shrink small-effect markers toward zero much stronger than BRR. These frameworks provide a means to estimate marker effects for a range of traits with different genetic architectures, which is consistent with what has been reported for fatty acid traits in oat (Carlson et al., 2019) (**Supplementary Figures 1–18**).

2.6. Construction of Trait Specific Genomic Relationship Matrices

Trait-specific genomic relationship matrices (TGRM) were constructed using the estimated marker effects for each of the nine fatty acid phenotypes in the Diversity Panel. For each fatty acid phenotype, the TGRM are defined as

$$\mathbf{G}^* = \frac{\mathbf{MDM}'}{P} \quad (4)$$

where $\mathbf{M}$ is an $n \times P$ scaled and centered matrix of allele dosages with n being the number of individuals and P the number of markers. $\mathbf{D}$ is an $P \times P$ diagonal matrix that contains the marker weights. The weight for marker p is given by $\frac{a^2_p}{\sum_{p=1}^{P} a^2_p}$ where a_p is the additive effect.

2.7. Genomic Prediction

2.7.1. Prediction of Fatty Acid Phenotypes in the Elite Panel

To predict each fatty acid trait the following model was fitted

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z_u u} + \mathbf{Z_e s} + \mathbf{e} \quad (5)$$

where $\mathbf{y}$ is a vector of deregressed BLUPs for each line in the six trials; $\mathbf{Z_u}$ is an $n \times q$ incidence matrix that assigns the q genomic values to n observations; $\mathbf{u}$ is a vector of genomic values; and $\mathbf{Z_e}$ is an $n \times e$ incidence matrix that assigns observations to trials and $\mathbf{s}$ are the corresponding effects. Both TGRM-BLUP and gBLUP follow the same model, what separates the two methods are the assumptions on $\mathbf{u}$. For TGRM-BLUP, we assume $\mathbf{u} \sim N(0, \sigma^2_{g^*}\mathbf{G}^*)$ where $\mathbf{G}^*$ is the TGRM defined above, and for gBLUP we assume $\mathbf{u} \sim N(0, \sigma^2_g\mathbf{G})$ where $\mathbf{G}$ is a genomic relationship matrix calculated using VanRaden's second definition (VanRaden, 2008). All models were fitted using the `BGLR` package in R using the settings mentioned above (Pérez and de Los Campos, 2014). Prediction

accuracies were assessed using five-fold cross validation with 50 independent resampling runs. In each resampling run, the dataset was randomly split into five-folds. The models were trained on 80% of the data and predictions were made on the remaining 20%. This process was repeated until each fold was used as the testing set. Prediction accuracies were computed using Pearson's correlation between observed phenotypes and predicted genomic values for lines in the testing set. Correlation coefficients were averaged across folds. Comparisons were made between gBLUP and TGRM-BLUP, and significant differences in the two methods were declared if TGRM-BLUP increased prediction accuracy in 90% of the resampling runs. We used this approach to compare methods over a *t*-test for two reasons: (1) in cross-validation each sample is drawn from the same dataset and are not independent, which violates one of the assumptions of the *t*-test; and (2) the magnitude of the *t*-statistic is dependant on the sample size, which is the number of resampling runs. Our approach is not dependent on the sample size and should be a more robust alternative to the *t*-test.

2.7.2. Prediction of Total Lipid Content in the Elite Panel

Prediction of total lipid content was performed using multi-kernel TGRM-BLUP (MK-TGRM-BLUP), multi-trait gBLUP, and gBLUP approaches. The model for MK-TGRM-BLUP is given by

$$\mathbf{y} = \boldsymbol{\mu} + \sum_{t}^{T} \mathbf{Z_u u_t} + \mathbf{Z_e s} + \mathbf{e} \tag{6}$$

with all matrices and vectors defined as above; however, $\mathbf{u_t}$ is a vector of genomic breeding values computed using the TGRM for fatty acid trait *t*. Prediction accuracy was assessed using Pearson's correlation between the predicted genomic estimated breeding values and the BLUPs for each trial. Prediction accuracies from the model above were compared to gBLUP to determine if TGRM affected genomic predictions.

The multi-trait BLUP model is

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{Z_U U} + \mathbf{e} \tag{7}$$

here $\mathbf{Y}$ is a $n \times T$ matrix of phenotypes and $\mathbf{U}$ is a $n \times T$ matrix of genomic breeding values. BLUPs were averaged across the six trials and were used to construct $\mathbf{Y}$. These data were also used to fit MK-TGRM-BLUP models that were compared to multi-trait gBLUP and are given by $\mathbf{y} = \boldsymbol{\mu} + \sum_{t}^{T} \mathbf{Z_u u_t} + \mathbf{e}$. Prediction accuracy was assessed in the Elite Panel using five-fold cross validation. Since 12 lines were included in both the Diversity and Elite panels, and had phenotypes for both fatty acid and NIRS traits, these lines were always included in the training data. The testing set included lines that only had NIRS phenotypes. All models were fitted using the BGLR package as described earlier (Pérez and de Los Campos, 2014).

3. RESULTS

Nine fatty acid phenotypes were quantified in a panel of 336 diverse oat lines (referred to hereafter as the Diversity Panel) using targeted GC-MS (**Supplementary File 1**). Generally, the fatty acid phenotypes were highly correlated at both the genetic and phenotypic levels and correlation patterns were reflective of the biochemical relationships between compounds (**Figure 1**). For instance, we observed strong positive correlations among C18-type and C20-type fatty acids. Moreover, shorter chain fatty acids (e.g., C14 and C16) which are synthesized in the early steps of fatty acid biosynthesis also exhibited strong positive correlations (Ohlrogge and Jaworski, 1997; Brown et al., 2009; Li-Beisson et al., 2013). There were exceptions to these patterns, specifically for C16:1 and C18:3. These fatty acids showed much lower positive correlations with all other fatty acid phenotypes. Narrow-sense heritability estimates were moderate to high and ranged from 0.38 to 0.69, with the lowest and highest h^2 observed for C18:3 and C18:0, respectively. Collectively, these results suggest that these lipid phenotypes are genetically interrelated and are under additive genetic control.

3.1. Construction of Trait-Specific Genomic Relationship Matrices (TGRMs)

Given that a significant portion of phenotypic variation in these lipid phenotypes could be explained by additive genetic effects, we sought to leverage these effects to better predict lipid-related traits in an independent population. We constructed trait-specific genomic relationship matrices (TGRMs), which differentially weight markers based on their additive genetic effects on the phenotype. Since the genetic architectures of the fatty acid phenotypes differ, we used five Bayesian whole-genome regression approaches to estimate marker effects: Bayesian ridge regression (BRR), Bayes A, Bayes B, Bayes Cπ, and Bayesian LASSO (BL; **Supplementary Figures 1–18**). These approaches sample marker effects from various prior densities and can accommodate a wide range of genetic architectures (see section 2). We evaluated whether the signal captured by these TGRMs are transferable across populations by predicting the same fatty acid phenotypes measured in an independent population (Elite Panel) and environment. Predictive ability was assessed using five-fold cross validation with 50 independent resampling runs. Genomic BLUP (gBLUP) using VanRaden's second GRM was used as a baseline model. The TGRM-BLUP approaches were deemed to significantly improve prediction accuracies if the TGRM out-performed gBLUP in 90% of the resampling runs (**Table 1**, **Figure 2**).

With the exception of C18:1 and C18:3, prediction accuracies were significantly improved by using a TGRM, indicating that the signal captured by TGRMs is relevant in this second independent population (**Table 1**, **Figure 2**). Comparisons between TGRM approaches showed small, often non-significant differences between methods used to estimate marker effects (**Figure 2**, **Supplementary Table 2**). On average, Bayes B showed higher predictive abilities for more traits compared to other methods. For instance, Bayes B significantly outperformed at least one approach for six of the nine fatty acid traits

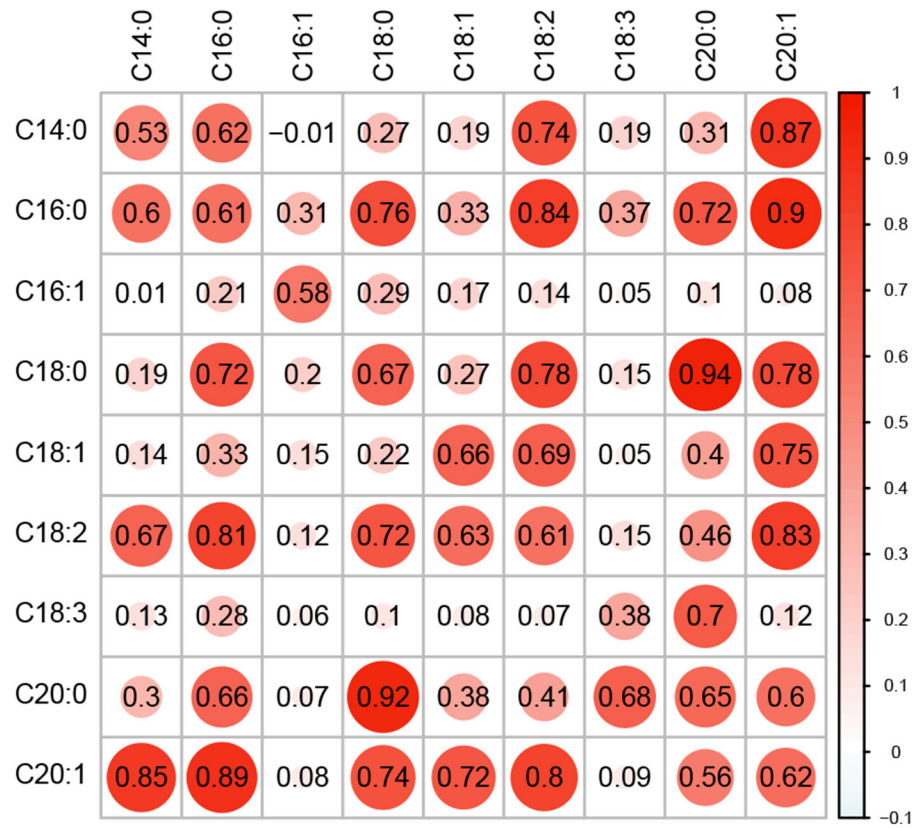

FIGURE 1 | Correlation and heritability for nine fatty acid traits. Genomic correlation between fatty acid phenotypes is shown in the upper triangle of the matrix, while the lower triangle shows the phenotypic correlations. Narrow-sense heritability estimates (h^2) are provided along the diagonal. All values were estimated using a multi-trait BLUP model using phenotypes recorded in the Diversity Panel. The size of each circle is proportional to the magnitude of the estimate.

TABLE 1 | Proportion of resampling runs where BLUP using trait-specific genomic relationship matrices (TGRM-BLUP) outperformed genomic BLUP (gBLUP).

Method	C14:0	C16:0	C16:1	C18:0	C18:1	C18:2	C18:3	C20:0	C20:1
BRR	**0.96**	**1.00**	**0.92**	**1.00**	0.48	**1.00**	0.62	**1.00**	0.68
Bayes A	0.82	**1.00**	0.80	**1.00**	0.38	**0.98**	0.28	**1.00**	0.54
Bayes B	**1.00**	**1.00**	**0.96**	**1.00**	0.54	**1.00**	0.58	**1.00**	**0.92**
Bayes Cπ	**1.00**	**1.00**	**0.96**	**1.00**	0.58	**0.98**	0.62	**1.00**	0.86
BL	0.74	**1.00**	**0.94**	**1.00**	0.52	**0.98**	0.50	**1.00**	0.74

Marker effects were estimated using five Bayesian whole-genome regression approaches for each of the nine fatty acid traits in the Diversity Panel (336 lines). Predicted marker effects were used to construct TGRMs for each trait. The predictive ability of TGRM-BLUP was assessed using nine fatty acid phenotypes measured in a population of 213 oat lines (Elite Panel). Five-fold cross validation was performed with 50 independent resampling runs. TGRM-BLUP was deemed to significantly improve genomic predictions in a TGRM-BLUP approach that outperformed gBLUP in 90% or more of the resampling runs, and are indicated by boldfaced text. BRR, Bayesian ridge regression; BL, Bayesian LASSO.

(**Supplementary Table 2**). Bayes Cπ also showed significantly higher predictive abilities relative to other approaches, and significantly outperformed at least one TGRM approach for four of the nine traits (**Supplementary Table 2**). Bayesian LASSO generally showed the lowest predictive ability among the TGRM approaches and did not outperform any approach for any trait. Collectively, these results show that the predicted marker effects are transferable across populations and can improve genomic prediction for endophenotypes for such seed traits as total lipid content. Moreover, the Bayesian whole-genome regression approaches that use a scale mixture prior may better capture genetic signal for traits with different genetic architectures, and may be a robust approach to estimate marker effects and create TGRMs.

3.2. Using TGRMs to Predict Total Lipid Content

The previous analyses showed that TGRMs can be used to improve genomic prediction for fatty acid traits in an independent population. While these outcomes provide support

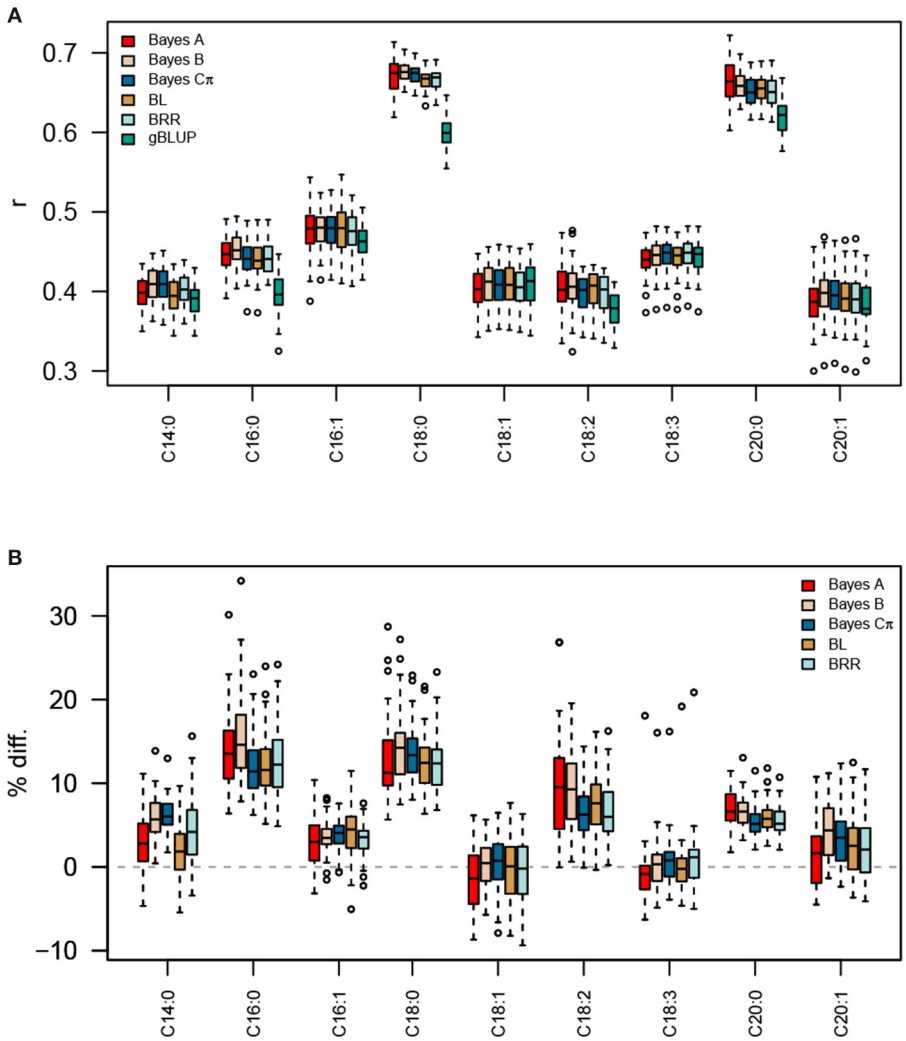

FIGURE 2 | Prediction accuracies for fatty acid traits using TGRM-BLUP and gBLUP. Five Bayesian whole-genome regression approaches (Bayes A, Bayes B, Bayes Cπ, BRR, and BL) were used to estimate marker effects for each fatty acid trait in the Diversity Panel. These marker effects were used to construct trait-specific genomic relationship matrices (TGRM) and were used to predict fatty acid abundances in the Elite Panel. Prediction accuracy was assessed using five-fold cross validation with 50 resampling runs. The correlation between predicted genomic breeding values in the testing population and the observed phenotypes is shown in **(A)**. Panel **(B)** shows the percent improvement relative to genomic BLUP (gBLUP) for each trait. BL, Bayesian LASSO; BRR, Bayesian ridge regression; *r*, Pearson's correlation coefficient.

for the use of TGRMs in breeding programs, the quantification of these compounds may not be feasible in breeding programs due to the high cost of GC-MS. Seed compositional traits measured via indirect methods, e.g., near-infrared spectroscopy (NIRS), is a more feasible approach to quantify total seed lipids in a large breeding program (Melchinger et al., 1986; Rosales et al., 2011; Diepenbrock and Gore, 2015). With this in mind, we used the TGRMs for each of the nine fatty acid traits to predict total seed lipid content measured through NIRS using a multi-kernel genomic prediction model (MK-TGRM-BLUP). Prediction accuracies for each multi-kernel model were compared to gBLUP and the TRGM methods were determined to significantly improve prediction accuracies if it outperformed gBLUP in at least 90% of sampling runs.

All MK-TGRM-BLUP approaches significantly increased prediction accuracies compared to gBLUP (**Figure 3**). Improvements in prediction accuracies ranged from 11.8 to 13.8%. Differences between MK-TGRM-BLUP approaches were minimal and non-significant. In contrast to the predictions for fatty acid traits, BRR showed slightly higher prediction accuracies on average ($r = 0.481$) compared to other approaches, while Bayes A showed the lowest prediction accuracy among the MK-TGRM-BLUP approaches ($r = 0.473$).

Given that the MK-TGRM-BLUP leverages information on related traits to improve prediction accuracies, we also compared

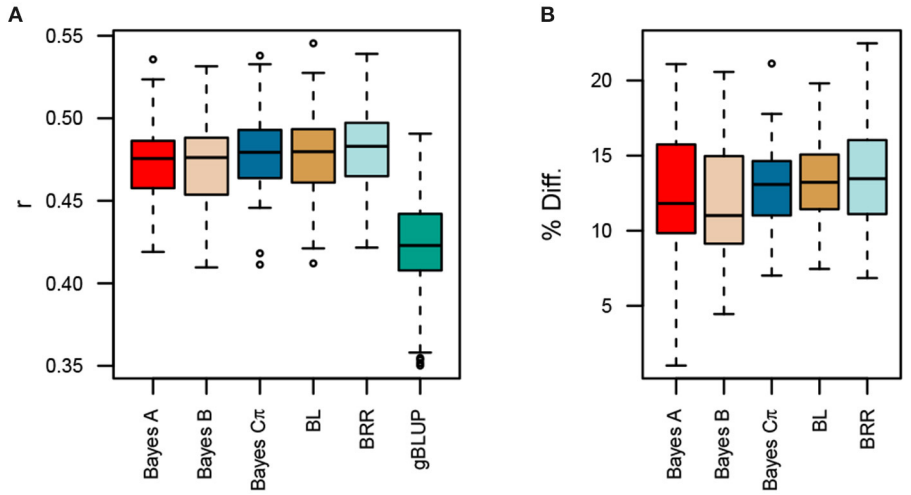

FIGURE 3 | Comparison of prediction accuracies for multi-kernel trait-specific BLUP models (MK-TGRM-BLUP) and a genomic BLUP approach (gBLUP). The multi-kernel models used TGRMs constructed from estimated marker effects for the nine fatty acid traits. Prediction accuracy was assessed using five-fold cross validation with 50 resampling runs. The correlation between predicted genomic breeding values in the testing population and the observed phenotypes at each location is shown in **(A)**. Panel **(B)** shows the percent improvement relative to gBLUP for each MK-TGRM-BLUP approach. BL, Bayesian LASSO; BRR, Bayesian ridge regression; *r*, Pearson's correlation coefficient.

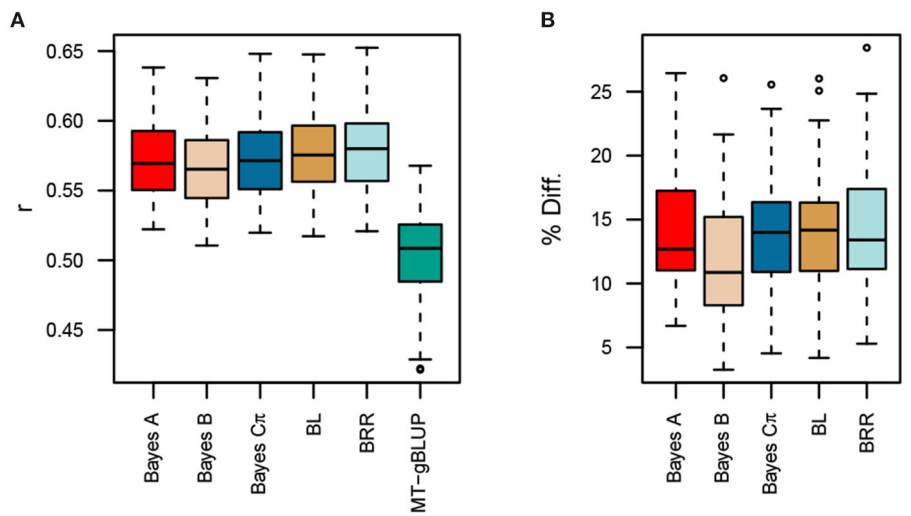

FIGURE 4 | Comparison of prediction accuracies for multi-kernel trait-specific BLUP models (MK-TGRM-BLUP) and a multi-trait gBLUP approach (MT-gBLUP). The multi-trait gBLUP model used phenotypes for the nine fatty acid traits and total lipid content measured via near-infrared spectroscopy (NIRS) to predict total lipid content. Prediction accuracy was assessed using five-fold cross validation with 50 resampling runs. Since there is a small overlap between lines in the diversity panel, which have fatty acid phenotypes, and lines in the Elite Panel, these common lines were always included in the training set. The testing set is then 20% of the lines that only have NIRS phenotypes. The correlation between predicted genomic breeding values in the testing population and the average of observed phenotypes across locations is shown in **(A)**. Panel **(B)** shows the percent improvement relative to MT-gBLUP for each MK-TGRM-BLUP approach. BL, Bayesian LASSO; BRR, Bayesian ridge regression; *r*, Pearson's correlation coefficient.

the MK-TGRM-BLUP approach to a multitrait gBLUP (MT-gBLUP) model that jointly modeled all nine fatty acid traits in the Diversity Panel and total lipid content in the Elite Panel. Thus, MT-gBLUP contains all the data that was used to compute the TGRM for fatty acids used in the MK-TGRM-BLUP model. A total of 12 lines in the Elite Panel had phenotypes for individual fatty acids and their sum. Five-fold cross validation was used for the remaining 198 lines in the Elite Panel with phenotypes for total lipid content. All TGRM-BLUP approaches showed significant improvements in prediction accuracies over the MT-gBLUP approach (**Figure 4**). Prediction accuracies were highest on average for BRR ($r = 0.578$), which showed a 14.41%

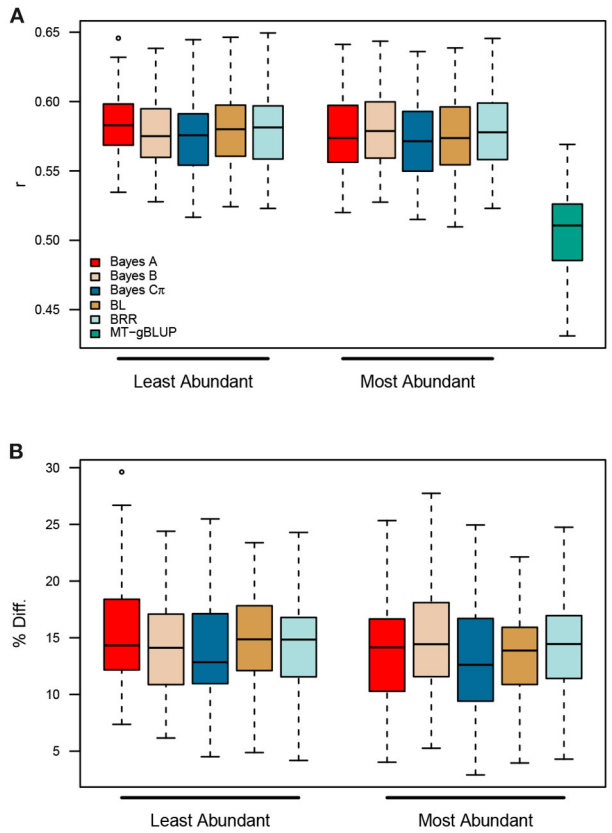

FIGURE 5 | Prediction accuracies for two multi-kernel trait-specific BLUP models (MK-TGRM-BLUP) that use TGRM for the three most abundant and three least abundant fatty acids. Estimated marker effects for the most abundant fatty acids (C16:0, C18:1, and C18:2) were used to create the three TGRM and were used in a multi-kernel gBLUP framework to predict total lipid content. A similar approach was used with the three least abundant fatty acids (C14:0, C16:1, and C20:0). Predictions for each MK-TGRM-BLUP approach were compared to a multi-trait gBLUP approach. Five-fold cross validation was performed using 50 resampling runs. BL, Bayesian LASSO; BRR, Bayesian ridge regression; *r*, Pearson's correlation coefficient.

increase in prediction accuracy over MT-gBLUP. Collectively, these results suggest that the use of a TGRM approach can significantly improve prediction accuracies over conventional genomic prediction approaches, even when information on related phenotypes is included in the prediction model.

Finally, we asked whether it was necessary to quantify and construct TGRM for all fatty acids, or whether similar improvements in prediction accuracy could be achieved by using kernels for the most abundant fatty acids. In both panels, C16:0, C18:1, and C18:2 were the most abundant fatty acids, while C14:0 C16:1 and C20:0 were present at much lower levels (**Supplementary Figure 20**). Two MK-TGRM-BLUP models were constructed using kernels for the top three most abundant fatty acids and the three least abundant fatty acids. These MK-TGRM-BLUP approaches were compared to the MT-gBLUP model described above using five-fold cross validation. Both MK-TGRM-BLUP approaches outperformed MT-gBLUP in all resampling runs, indicating that including genetic signal for a subset of fatty acid traits is sufficient to significantly improve prediction for total lipid content (**Figure 5**). Comparisons between the two MK-TGRM-BLUP approaches did not show any significant differences in prediction accuracies, which may be due to QTL that are shared between fatty acids (**Supplementary Figures 2, 3, 5, 6, 8**).

4. DISCUSSION

Omics technologies provide an easy and effective way to measure thousands of endophenotypes in large mapping populations. Many research groups are using these approaches to improve prediction for complex traits (Guo et al., 2016; Westhues et al., 2017; Rincent et al., 2018; Schrag et al., 2018; Li et al., 2019; Xiang et al., 2019; Rohde et al., 2020; Zhou et al., 2020). While several studies have reported improvements in prediction accuracies when these data were used to create relationship matrices, the results are often mixed and inconsistent (Guo et al., 2016; Schrag et al., 2018; Zhou et al., 2020). More importantly, such approaches can be costly to implement in a breeding program since individuals in the testing population require records for endophenotypes. TGRMs offer an alternative approach to use relevant information on endophenotypes to improve prediction for conventional traits.

In this study, we show that data on endophenotypes can be used to create TGRMs that majorly improve prediction for related higher level focal traits. The TGRM improved prediction accuracies for most traits by as much as 15%. The greatest improvements among fatty acid traits was observed for C16:0 when marker effects were estimated using Bayes Cπ. C16:0 showed moderate to high heritabilities in the Diversity and Elite Panels ($h^2 = 0.68$ and 0.64, respectively), and it seemed to be affected by at least one large-effect QTL in both panels (**Supplementary Figures 2, 11**). Thus, predictions for this trait can be improved by placing greater emphasis on putative causal markers when defining the genomic relationships among lines. These results are in agreement with other studies that evaluated TGRMs (Tiezzi and Maltecca, 2015; Karaman et al., 2018; Ren et al., 2020). Improvements over gBLUP were most pronounced for high heritability traits that were regulated by a few large-effect QTL, which is expected given that such traits are far from the infinitesimal model assumed by gBLUP (Tiezzi and Maltecca, 2015; Karaman et al., 2018; Ren et al., 2020). This likely explains the improvements in prediction accuracies observed for C16:0 with TGRM-BLUP. Ren et al. (2020) used several TGRM-BLUP approaches to predict both simulated and real traits in three species. Marker weights were estimated using methods with priors that impose local or global shrinkage, and several types of TGRM were constructed using these weights. The authors reported the greatest improvements in prediction accuracies for simulated traits with moderate heritability and 200 QTL when TGRM were constructed using weights estimated using Bayes Cπ. The authors did not estimate marker effects using Bayes B; however, both Bayes B and Bayes Cπ use scale mixture densities to accommodate large-effect QTL (Gianola,

2013). With these approaches, estimates for small-effect QTL are shrunk heavily toward zero, while effects for markers that are in linkage disequilibrium with large-effect QTL are shrunk less. These approaches are more effective to estimate marker effects and construct TGRMs for traits that exhibit oligogenic architectures compared to methods that impose uniform shrinkage.

Predictions for two fatty acid traits, C18:1 and C18:3, were not significantly improved with TGRM-BLUP. C18:3 had the lowest heritability in the Diversity and Elite Panels (h^2 = 0.38 and 0.42, respectively) and exhibited a much more complex genetic architecture compared to other fatty acids (**Figure 1**, **Supplementary Figures 7, 16**). On average, prediction accuracies were improved by −0.73 to 1.0% over gBLUP, but only outperformed gBLUP in 28 to 62% of the resampling runs. These are not unexpected findings given that other studies that simulated traits with complex architectures and low heritabilities also failed to see much of an improvement with TGRM-BLUP (Tiezzi and Maltecca, 2015; Karaman et al., 2018; Ren et al., 2020). Compared to C18:3, heritability estimates were much higher for C18:1 and a large-effect QTL was detected in both panels on chromosome 3D, which explained about 6% of variation in C18:1 in the Diversity Panel, but predictions were not improved with TGRM-BLUP (**Figure 1**, **Supplementary Figures 5, 14**). Although the minor allele at this locus was common in the Diversity Panel (MAF = 0.40), the top marker was rare in the Elite Panel and was below the MAF threshold (MAF < 0.05) used when computing the TGRM.

Compared to other approaches that have created relationship matrices using endophenotype values, the TGRM approach should be more feasible to implement in a breeding program since predictions on the testing population can be performed without records for endophenotypes. Pertinent genetic information are passed between populations through marker effects for the endophenotypes. Of course, this assumes that relevant markers are still segregating in the testing population; therefore, it is important to carefully select a population to estimate marker effects. Fatty acid phenotypes were initially measured in the Diversity Panel which consists primarily of breeding materials from European and North American breeding programs, while the Elite Panel used for genomic prediction is comprised of materials used in oat breeding programs in the Upper Midwestern United States. Thus, the panel that was used to estimate marker effects is diverse and related to the materials used for prediction (**Supplementary Figure 21**).

Surprisingly, the MK-TGRM-BLUP approach showed significant improvements in prediction accuracy over gBLUP and a multi-trait gBLUP model for total lipid content. Total lipid content exhibited a much more complex genetic architecture compared to the fatty acid traits; therefore, we expected the TGRM approaches to perform equally as well or slightly better than gBLUP (**Supplementary Figure 19**). Prediction accuracies were improved by 11.8 to 13.8% relative to gBLUP and 11.9 to 14.4% relative to MT-gBLUP. The MT-gBLUP approach jointly fits fatty acids and total lipid content, and should be able to use the signal contained in the fatty acid phenotypes to improve predictions for total lipid content. One explanation for the increased performance of MK-TGRM-BLUP over MT-gBLUP is that the former is a more parsimonious model. Since an unstructured covariance matrix was used for MT-gBLUP, all variances and covariances must be estimated. MK-TGRM-BLUP on the other hand does not estimate covariances between the traits, rather information on related traits is provided through the kernels. A second possibility is that the MT-gBLUP model assumes an infinitesimal architecture for all traits. While this may be the case for total lipid content and some fatty acid traits, several fatty acids showed a much simpler architecture (**Supplementary Figures 1–19**). The MT-gBLUP approach may shrink these large-effect QTL for endophenotypes with simpler genetic architectures. Nonetheless, these results demonstrate that TGRM for related endophenotypes can be leveraged to improve prediction for lower-cost traits to assess seed quality traits in breeding programs. Moreover, we show that information on a subset of fatty acids can be leveraged to significantly improve predictions for total lipid content relative to the MT-gBLUP approach. The majority of total lipid content in oat is due to triglycerides, which consist of three fatty acids bound to glycerol (Leonova et al., 2008). Leonova et al. (2008) reported that C16:0, C18:1, and C18:2 were the most predominant fatty acids found in the oat seed, which is supported by our results in both the Diversity and Elite panels (**Supplementary Figure 20**). Since these fatty acids should be most relevant to total lipid content, this prompted us to evaluate whether information on these endophenotypes was sufficient to improve prediction for total lipid content. MK-TGRM-BLUP models that included information for these fatty acids significantly outperformed MT-gBLUP for predicting total lipid content, suggesting that the most predominant fatty acids can be quantified and used to predict total lipid content. Surprisingly, prediction accuracies for these MK-TGRM-BLUP models that used kernels for the most abundant fatty acids showed equivalent prediction accuracies with MK-TGRM-BLUP approaches that used kernels for the three least abundant fatty acids. Several QTL were shared between fatty acids. For instance, a QTL was identified on chromosome 6A for C16:0, C18:2, and C16:1 (**Supplementary Figures 2, 3, 6**). A second shared QTL was identified on chromosome 3D for C18:1 and C20:0, suggesting that these loci may have pleiotropic effects on low and high abundant fatty acid traits (**Supplementary Figures 5, 8**).

One major assumption of the approaches used in this study is that the focal trait is influenced by a relatively small number of endophenotypes that are known beforehand. For some traits, such as seed lipid content, selecting which endophenotypes to include in the model is somewhat straightforward, as we know the focal trait is essentially a summary of all lipids in the tissue, and marker effects can be predicted for the important lipids. Information on these traits can be introduced using a multi-kernel prediction model, but this is not feasible when tens or hundreds of endophentoypes possibly affect the focal trait. High dimensionality would particularly be a problem for traits like yield, which are influenced by many molecular processes. Selecting a small subset of relevant endophenotypes for such traits from dense omics data can be challenging. In these cases, it

may be appropriate to use a combination of dimension-reduction and variable selection methods to select relevant phenotypes or linear combinations of phenotypes. Methods like principal component analysis or factor analysis have been used extensively to cope with high-dimensional traits (Runcie and Mukherjee, 2013; Wang and Stephens, 2018; Carlson et al., 2019; Sakamoto et al., 2019; Yu et al., 2019; Campbell et al., 2020; Rice et al., 2020; Runcie et al., 2020). These approaches can be used to create derived traits that capture (co)variance in the original data, and marker effects can be easily estimated using GWAS or whole-genome regression approaches. Thus, TGRMs can be constructed from marker effects for these derived phenotypes. A second limitation of our approach, which is shared with other BLUP methods, is that computations and storage of TGRM many be unfeasible with very large populations ($>$ 100k individuals) (Aguilar et al., 2011; Misztal et al., 2020). The storage of GRMs scale quadratically with the number of individuals, and inversion of GRMs increase cubically. Although populations of this size are rare in public plant breeding programs, genomic studies in animals and humans routinely involve genetic data for $>$ 100k individuals. In such cases indirect approaches can be used to overcome these computational issues and use BLUP frameworks for genetic evaluations in large populations (see Misztal et al., 2020 for review).

In conclusion, this study highlights the utility of TGRMs for related endophenotypes to predict complex traits in crops. Since the frameworks presented in this study do not require endophenotypes for selection candidates, these methods should be tractable to employ in breeding programs. Endophenotypes and their corresponding marker effects can be quantified in a large, diverse, discovery population, enabling them to be collectively leveraged to improve prediction accuracies for conventional traits in related populations.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**. The GitHub repository is https://github.com/malachycampbell/TGRM_frontiers.

AUTHOR CONTRIBUTIONS

Metabolomic data were generated by HH, TY, KS, LG, and MC-T. Analyses were performed by MC under the guidance of MG and J-LJ. MC wrote the manuscript with guidance from J-LJ and MG. Comments were provided by HH, LG, LB, MS, MG, and J-LJ. This study was supported by grants secured by KS, LG, MC-T, MS, MG, and J-LJ. All authors read and approved the manuscript.

FUNDING

Funding for this research was provided by USDA-NIFA-AFRI 2017-67007-26502. The USDA is an equal opportunity provider and employer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.643733/full#supplementary-material

REFERENCES

Aguilar, I., Misztal, I., Legarra, A., and Tsuruta, S. (2011). Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet.* 128, 422–428. doi: 10.1111/j.1439-0388.2010.00912.x

Blake, V. C., Birkett, C., Matthews, D. E., Hane, D. L., Bradbury, P., and Jannink, J.-L. (2016). The triticeae toolbox: combining phenotype and genotype data to advance small-grains breeding. *Plant Genome* 9, 1–10. doi: 10.3835/plantgenome2014.12.0099

Brown, A. P., Slabas, A. R., and Rafferty, J. B. (2009). "Fatty acid biosynthesis in plants-metabolic pathways, structure and organization," in *Lipids in Photosynthesis*, Vol. 30, eds H. Wada and N. Murata (Dordrecht: Springer), 11–34. doi: 10.1007/978-90-481-2863-1_2

Campbell, M. T., Hu, H., Yeats, T. H., Caffe-Treml, M., Gutiérrez, L., Smith, K. P., et al. (2020). Translating insights from the seed metabolome into improved prediction for healthful compounds in oat (*Avena sativa* L.). *Genetics*. iyaa043. doi: 10.1093/genetics/iyaa043

Carlson, M. O., Montilla-Bascon, G., Hoekenga, O. A., Tinker, N. A., Poland, J., Baseggio, M., et al. (2019). Multivariate genome-wide association analyses reveal the genetic basis of seed fatty acid composition in oat (*Avena sativa* L.). *Genes Genomes Genet.* 9, 2963–2975. doi: 10.1534/g3.119.400228

Chan, A. W., Hamblin, M. T., and Jannink, J.-L. (2016). Evaluating imputation algorithms for low-depth genotyping-by-sequencing (GBS) data. *PLoS ONE* 11:e0160733. doi: 10.1371/journal.pone.0160733

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS ONE* 11:e0156744. doi: 10.1371/journal.pone.0156744

de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 9:e1003608. doi: 10.1371/journal.pgen.1003608

Diepenbrock, C. H., and Gore, M. A. (2015). Closing the divide between human nutrition and plant breeding. *Crop Sci.* 55, 1437–1448. doi: 10.2135/cropsci2014.08.0555

Edriss, V., Gao, Y., Zhang, X., Jumbo, M. B., Makumbi, D., Olsen, M. S., et al. (2017). Genomic prediction in a large African maize population. *Crop Sci.* 57, 2361–2371. doi: 10.2135/cropsci2016.08.0715

Edwards, S. M., Sørensen, I. F., Sarup, P., Mackay, T. F., and Sørensen, P. (2016). Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in drosophila melanogaster. *Genetics* 203, 1871–1883. doi: 10.1534/genetics.116.187161

Gianola, D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753

Gianola, D., Fernando, R. L., and Schön, C.-C. (2020). Inferring trait-specific similarity among individuals from molecular markers and phenotypes with bayesian regression. *Theoret. Popul. Biol.* 132, 47–59. doi: 10.1016/j.tpb.2019.11.008

Guo, Z., Magwire, M. M., Basten, C. J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoret. Appl. Genet.* 129, 2413–2427. doi: 10.1007/s00122-016-2780-5

Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* 95, 535–552. doi: 10.1016/j.ajhg.2014.10.004

Karaman, E., Lund, M. S., Anche, M. T., Janss, L., and Su, G. (2018). Genomic prediction using multi-trait weighted gblup accounting for heterogeneous variances and covariances across the genome. *Genes Genomes Genet*. 8, 3549–3558. doi: 10.1534/g3.118.200673

Krause, M. R., González-Pérez, L., Crossa, J., Pérez-Rodríguez, P., Montesinos-López, O., Singh, R. P., et al. (2019). Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat. *Genes Genomes Genet*. 9, 1231–1247. doi: 10.1534/g3.118.200856

Kremling, K. A., Diepenbrock, C. H., Gore, M. A., Buckler, E. S., and Bandillo, N. B. (2019). Transcriptome-wide association supplements genome-wide association in Zea mays. *Genes Genomes Genet*. 9, 3023–3033. doi: 10.1534/g3.119.400549

Leonova, S., Shelenga, T., Hamberg, M., Konarev, A. V., Loskutov, I., and Carlsson, A. S. (2008). Analysis of oil composition in cultivars and wild species of oat (*Avena* sp.). *J. Agric. Food Chem*. 56, 7983–7991. doi: 10.1021/jf800761c

Li, Z., Gao, N., Martini, J. W., and Simianer, H. (2019). Integrating gene expression data into genomic prediction. *Front. Genet*. 10:126. doi: 10.3389/fgene.2019.00126

Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., et al. (2013). Acyl-lipid metabolism. *Arabidopsis Book* 11, 2–70. doi: 10.1199/tab.0161

MacLeod, I., Bowman, P., Vander Jagt, C., Haile-Mariam, M., Kemper, K., Chamberlain, A., et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* 17:144. doi: 10.1186/s12864-016-2443-6

Melchinger, A., Schmidt, G., and Geiger, H. (1986). Evaluation of near infra-red reflectance spectroscopy for predicting grain and stover quality traits in maize. *Plant Breed*. 97, 20–29. doi: 10.1111/j.1439-0523.1986.tb01297.x

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Misztal, I., Lourenco, D., and Legarra, A. (2020). Current status of genomic evaluation. *J. Anim. Sci*. 98:skaa101. doi: 10.1093/jas/skaa101

Morgante, F., Huang, W., Sørensen, P., Maltecca, C., and Mackay, T. F. (2020). Leveraging multiple layers of data to predict drosophila complex traits. *Genes Genomes Genet*. 10, 4599–4613. doi: 10.1534/g3.120.401847

Ohlrogge, J. B., and Jaworski, J. G. (1997). Regulation of fatty acid synthesis. *Annu. Rev. Plant Biol*. 48, 109–136. doi: 10.1146/annurev.arplant.48.1.109

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

Ren, D., An, L., Li, B., Qiao, L., and Liu, W. (2020). Efficient weighting methods for genomic best linear-unbiased prediction (BLUP) adapted to the genetic architectures of quantitative traits. *Heredity* 126, 320–334. doi: 10.1038/s41437-020-00372-y

Rice, B. R., Fernandes, S. B., and Lipka, A. E. (2020). Multi-trait genome-wide association studies reveal loci associated with maize inflorescence and leaf architecture. *Plant Cell Physiol*. 61, 1427–1437. doi: 10.1093/pcp/pcaa039

Rincent, R., Charpentier, J.-P., Faivre-Rampant, P., Paux, E., Le Gouis, J., Bastien, C., et al. (2018). Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. *Genes Genomes Genet*. 8, 3961–3972. doi: 10.1101/302117

Rohde, P. D., Kristensen, T. N., Sarup, P., Muñoz, J., and Malmendal, A. (2020). Prediction of complex phenotypes using the Drosophila metabolome. *bioRxiv [Preprint]*. doi: 10.1101/2020.06.11.145623

Rosales, A., Galicia, L., Oviedo, E., Islas, C., and Palacios-Rojas, N. (2011). Near-infrared reflectance spectroscopy (NIRS) for protein, tryptophan, and lysine evaluation in quality protein maize (QPM) breeding programs. *J. Agric. Food Chem*. 59, 10781–10786. doi: 10.1021/jf201468x

Runcie, D. E., Cheng, H., and Crawford, L. (2020). Mega-scale linear mixed models for genomic predictions with thousands of traits. *bioRxiv bioRxiv [Preprint]*. doi: 10.1101/2020.05.26.116814

Runcie, D. E., and Mukherjee, S. (2013). Dissecting high-dimensional phenotypes with bayesian sparse factor analysis of genetic covariance matrices. *Genetics* 194, 753–767. doi: 10.1534/genetics.113.151217

Sakamoto, L., Kajiya-Kanegae, H., Noshita, K., Takanashi, H., Kobayashi, M., Kudo, T., et al. (2019). Comparison of shape quantification methods for genomic prediction, and genome-wide association study of sorghum seed morphology. *PLoS ONE* 14:e0224695. doi: 10.1371/journal.pone.0224695

Schrag, T. A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., et al. (2018). Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics* 208, 1373–1385. doi: 10.1534/genetics.117.300374

Speed, D., and Balding, D. J. (2014). Multiblup: improved SNP-based prediction for complex traits. *Genome Res*. 24, 1550–1557. doi: 10.1101/gr.169375.113

Su, G., Christensen, O. F., Janss, L., and Lund, M. S. (2014). Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. *J. Dairy Sci*. 97, 6547–6559. doi: 10.3168/jds.2014-8210

Sun, X., Qu, L., Garrick, D. J., Dekkers, J. C., and Fernando, R. L. (2012). A fast EM algorithm for bayesa-like prediction of genomic breeding values. *PLoS ONE* 7:e49157. doi: 10.1371/journal.pone.0049157

Tiezzi, F., and Maltecca, C. (2015). Accounting for trait architecture in genomic predictions of US holstein cattle using a weighted realized relationship matrix. *Genet. Select. Evol*. 47:24. doi: 10.1186/s12711-015-0100-1

Turner-Hissong, S. D., Bird, K. A., Lipka, A. E., King, E. G., Beissinger, T. M., and Angelovici, R. (2020). Genomic prediction informed by biological processes expands our understanding of the genetic architecture underlying free amino acid traits in dry Arabidopsis seeds. *Genes Genomes Genet*. 10, 4227–4239. doi: 10.1534/g3.120.401240

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci*. 91, 4414–4423. doi: 10.3168/jds.2007-0980

Wang, W., and Stephens, M. (2018). Empirical bayes matrix factorization. *arXiv preprint arXiv:1802.06931*. Available online at: https://arxiv.org/abs/1802.06931

Westhues, M., Schrag, T. A., Heuer, C., Thaller, G., Utz, H. F., Schipprack, W., et al. (2017). Omics-based hybrid prediction in maize. *Theoret. Appl. Genet*. 130, 1927–1939. doi: 10.1007/s00122-017-2934-0

Xiang, R., Van Den Berg, I., MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., et al. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci. U.S.A*. 116, 19398–19408. doi: 10.1073/pnas.1904159116

Yu, H., Campbell, M. T., Zhang, Q., Walia, H., and Morota, G. (2019). Genomic bayesian confirmatory factor analysis and bayesian network to characterize a wide spectrum of rice phenotypes. *Genes Genomes Genet*. 9, 1975–1986. doi: 10.1534/g3.119.400154

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.-J., and Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE* 5:e12648. doi: 10.1371/journal.pone.0012648

Zhou, S., Morgante, F., Geisz, M. S., Ma, J., Anholt, R. R., and Mackay, T. F. (2020). Systems genetics of the Drosophila metabolome. *Genome Res*. 30, 392–405. doi: 10.1101/gr.243030.118

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with the authors MC and MG.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 May 2021
doi: 10.3389/fgene.2021.655287

TrainSel: An R Package for Selection of Training Populations

Deniz Akdemir[1]*, *Simon Rio*[2] *and Julio Isidro y Sánchez*[2]*

[1] Agriculture & Food Science Centre, Animal and Crop Science Division, University College Dublin, Dublin, Ireland, [2] Centro de Biotecnologia y Genómica de Plantas (CBGP, UPM-INIA), Instituto Nacional de Investigación y Tecnologia Agraria y Alimentaria (INIA), Universidad Politécnica de Madrid (UPM), Madrid, Spain

A major barrier to the wider use of supervised learning in emerging applications, such as genomic selection, is the lack of sufficient and representative labeled data to train prediction models. The amount and quality of labeled training data in many applications is usually limited and therefore careful selection of the training examples to be labeled can be useful for improving the accuracies in predictive learning tasks. In this paper, we present an R package, TrainSel, which provides flexible, efficient, and easy-to-use tools that can be used for the selection of training populations (STP). We illustrate its use, performance, and potentials in four different supervised learning applications within and outside of the plant breeding area.

Keywords: training optimization, machine learning, genomic selection, genomic prediction, image classification, multi-objective optimization, mixed models

OPEN ACCESS

Edited by:
Diego Jarquin, University of Nebraska-Lincoln, United States

Reviewed by:
Roberto Fritsche-Neto, International Rice Research Institute (IRRI), Philippines
Luc L. Janss, Aarhus University, Denmark

***Correspondence:**
Deniz Akdemir
deniz.akdemir.work@gmail.com
Julio Isidro y Sánchez
j.isidro@upm.es

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 January 2021*
Accepted: *31 March 2021*
Published: *07 May 2021*

Citation:
Akdemir D, Rio S and Isidro y Sánchez J (2021) TrainSel: An R Package for Selection of Training Populations. Front. Genet. 12:655287. doi: 10.3389/fgene.2021.655287

1. INTRODUCTION

Genomic selection (GS) uses supervised learning for predicting genetic values of phenotyped and un-phenotyped individuals by using genomewide molecular markers (Meuwissen et al., 2001). Genomic prediction (GP) models are built using a training data, i.e., genomic and phenotypic data for a set of individuals. Unfortunately, phenotyping of plants is an expensive and time-consuming process due to factors such as reliance on human input and budget time and resource constraints. Therefore, the most important current bottleneck in application of GS in plant breeding programs is phenotyping. Selection of training populations (STP) in this context refers to identification of a set of training individuals to be phenotyped.

While the usefulness of optimal training set (TRS) in GS is clearly supported by the literature (Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015; Lorenz and Smith, 2015; He et al., 2016; Cericola et al., 2017; Neyhart et al., 2017; Norman et al., 2018; Akdemir and Isidro-Sánchez, 2019; Guo et al., 2019; Mangin et al., 2019; de Bem Oliveira et al., 2020; Olatoye et al., 2020; Yu et al., 2020; Kadam et al., 2021), the flexible and efficient software tools for implementing them have been limited. Indeed, only a few software tools such as STPGA (Akdemir, 2017) and TSDFGS (Ou and Liao, 2019) are available for public use. The TSDFGS is an R package that focuses on optimization of the TRS by a genetic algorithm (GA) and can be used for STP based on three built-in design criteria. Similarly, STPGA is an R package that uses a modified GA for solving subset selection problems but also allows users to chose from many predefined or user-defined criteria. Here, we designed a TrainSel package that provides many more options, for example, the ability to select multiple sets from multiple candidate sets, specification of whether or not the resulting set needs to be ordered, or the power to perform multi-objective optimization. In addition, TrainSel can be used for searching for solutions to variety of TRS and experimental design problems, such as randomized complete

block design, lattice design, etc. TrainSel uses GA in conjunction with simulated annealing (SA) steps, and functions are written in C++ using Rcpp (Eddelbuettel et al., 2011), and therefore, improves performance and is more efficient compared to both of the above alternatives.

In addition, the TrainSel package was designed to be applied not just for genomic assisted breeding situations, it can also be utilized for STP in general supervised learning problems. Supervised learning refers to the exercise of building predictive models that allow us to predict the states of certain output variables (referred as labels) based on certain input variables. To build supervised learning models we make use of a training dataset that includes observations of both the input variables and the labels, and generally, the larger and more representative the training dataset, the greater is the statistical power for supervised learning. We use the term label throughout this article to refer to the output variables that we are trying to predict. In genomic selection, labeling a genotype refers to measurement of phenotypic values for that genotype in one or more environments.

In this paper, we demonstrated the usage of the TrainSel R package for STP on genomic assisted breeding applications, but also included other applications to illustrate that STP may also be worthwhile for other supervised learning tasks, such as image classification.

2. MATERIALS AND METHODS

2.1. Populations for Selection of Training Population (STP)

During STP, we will encounter different types of populations. The target population (Akdemir and Isidro-Sánchez, 2019) is the population that the researcher is interested in, i.e., the population we want to make inferences about. The study population is the population that is accessible to the researcher. The candidate set (CS) is a countably finite representative subset of the study population, similarly, the test set (TS) is a countably finite representative subset of the target population. We assume that we either have an idea about the topology (referring to the initial data available on CS and TS before doing the experiment) of the union of the CS and TS, or that it is relatively easy to obtain this information. Finally, the initial information about the topology of the CS and TS is used to identify a subset of the CS as the training set (TRS) for measuring the labels and additional features. These populations and the default supervised learning paradigm is illustrated in **Figure 1**.

2.2. Optimization Algorithm in TrainSel

Selection of training population involves the selection of a subset from a set of candidates and therefore is a combinatorial problem. These problems are typically exponential in terms of computational complexity and may require exploring all possible solutions. Nevertheless, many modern publications point to the effectiveness of applying metaheuristics in obtaining "good" answers to combinatorial optimization problems.

TrainSel uses a combination of GA (Holland, 1992) and simulated annealing (SA) algorithm (Haines, 1987) for solving combinatorial optimization problems. Genetic algorithm uses techniques inspired by natural evolution such as inheritance, mutation, selection, and crossover to generate better solutions through iterations (Holland, 1992). Simulated annealing moves between solutions using a perturbation and acceptance scheme. At each iteration, a new solution is generated by perturbing the current solution, and this new solution is accepted if it improves the optimization criterion. If the perturbed solution is inferior to the current solution the new solution is accepted based on an acceptance probability that is inversely proportional to the distance of the new solution to the current solution and the current temperature of the system (Haines, 1987). Temperature parameter varies during the iterations of the SA algorithm and usually is a decreasing function of the iteration number. Acceptance of inferior solutions during the SA iterations allows the algorithm to explore more of the possible space of solutions.

Algorithms such as GA and SA outperform other traditional methods in many applications, as they are flexible and easy to implement (no mathematical analysis is needed when considering a large, complex, non-smooth, poorly-understood optimization problem). There is no proof of convergence for either GA or SA, however, they are effective on a large range of classic optimization problems, and more specifically, have proved to be effective for approximating globally optimal solutions to many combinatorial optimization problems (Glover and Kochenberger, 2006; Fischetti and Lodi, 2010).

Algorithm 1 describes the main steps of the sample selection algorithm for the single optimization criteria problems. A similar algorithm is used when optimizing more than one criteria. The main difference is that the elite solutions of a population are defined as the non-dominated solutions of the current population.

The parameters of the selection algorithm in TrainSel are: "npop" which is the size of the genetic algorithm population, "nelite" which is the number of elite solutions selected in each iteration, "niterations" which is the maximum number of iterations for the genetic algorithm, "miniterbefstop" is the minimum number of iterations of "no change" before the algorithm is deemed converged, "tolconv" which is the tolerance for determining "no change" in the criteria values, "niterSANN" which is the number of iterations for the SA algorithm, "stepSANN" which controls the speed of cooling of the SA algorithm. Each of these parameters comes with default settings, most of which do not need to be changed by the user for small to medium-sized optimization problems. For larger problems increasing "niterations" and "niterbefstop" parameters will usually suffice. We have done some experimentation with the default settings of the remaining parameters (and with relatively large values for "niterations" and "minitbefstop") algorithm in several problems with different complexities where the true solution was known. The results from these convergence experiments are provided in **Supplementary Figure 1**. The user can use these figures to guess initial estimates for these two parameters for their problems. After the run of the algorithm, the best way to decide if the algorithm has worked is by checking the flattening of the objective function values during the final iterations.

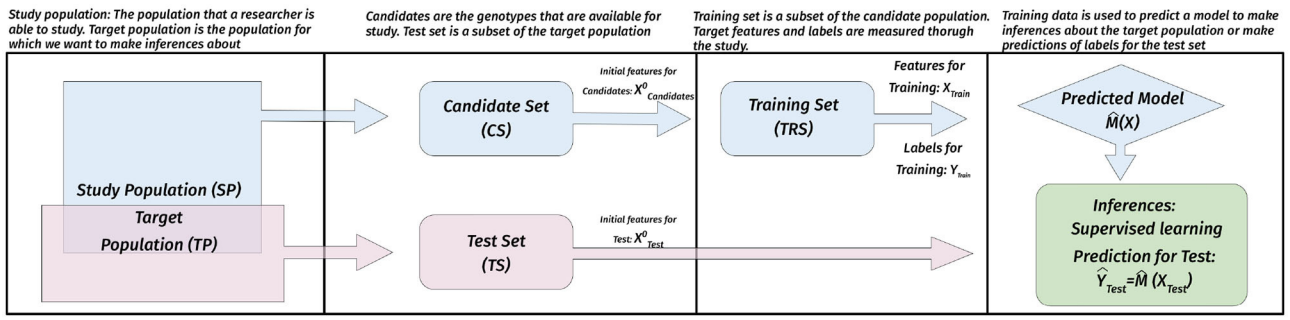

FIGURE 1 | Populations in STP and their use. The study population is the population that is accessible to the researcher. The candidate set (CS) is a countably finite representative subset of the SP, similarly, the test set (TS) is a countably finite representative subset of the target population. The initial information about the topology of the CS and TS ($X^0_{Candidate}$ and $X^0_{Candidate}$) is used to identify a subset of the CS as the training set (TRS) for measuring the labels (phenotypic values in GS) (Y_{Train}) and other related features (X_{Train}) (for instance, environmental covariates). The training data for TRS is used to build supervised learning models which are then used to make inferences and predictions.

Algorithm 1 : Combinatorial optimization algorithm in TrainSel

1: $t = 0$.
2: Initialization—Create an initial population of solutions of desired size, S_t. Parameters: npop
3: **repeat**
4: $t = t + 1$.
5: $S_t = \emptyset$.
6: Selection—Identify the best solutions in S_{t-1} by the ordering of criterion values. Let the best solutions be s_t. Parameters: nelite
7: SA—Improve elements of s_t with simulated annealing algorithm. Parameters: niterSANN, stepSANN
8: Elitism—Put s_t in S_t,
9: **repeat**
10: Crossover—Randomly pick two solutions in S_t. Obtain a recombination of these two solutions.
11: Mutation—Mutate the solution from the above step with a certain mutation probability and intensity. Parameters: mutprob, mutintensity
12: Insert this solution into S_t.
13: **until** S_t has N_{pop} solutions.
14: **until** Convergence: the achievement of the maximum number of iterations or non-improvement for a prescribed number of iterations. Parameters: niterations, miniterbefstop, tolconv **return** Best Solution.

In most applications of STP, the ordering of selected samples in the TRS will not be important and therefore only one instance of each individual is required for TRS sample; we refer to this case as an unordered set (UOS). In certain cases, the order of the sample will be important but again only one instance of each individual is required, we refer to this case as ordered set (OS). The cases where we allow more than one instance of each individual is referred to as unordered multiset (UOMS) and ordered multiset (OMS). TrainSel allows users to specify which of these types of sets the optimization problem falls into. An application of the use of finding optimal ordered sets is the design of a blocked experiment where we care about the design of the experiment, i.e., the assignment of individuals to different blocks, in addition to selecting which individuals to include in the study.

The search algorithm in TrainSel is not guaranteed to find globally optimal solutions, i.e., the solutions obtained by any run of TrainSel may be sub-optimal, and different solutions can be obtained given different starting conditions and optimization parameters. Another layer of safety can be obtained if the algorithm is started from multiple initial conditions, and the best of all the runs is selected as the final solution.

Numerous other algorithms have been proposed for the optimal subset selection problem, many of them are heuristic exchange type algorithms (Fedorov, 1972; Mitchell, 1974; Nguyen and Miller, 1992; Rincent et al., 2012; Isidro et al., 2015). In exchange type algorithms, new solutions are obtained by adding a sample unit and removing another at a time (some exchange algorithms might allow the exchange of more than one samples at once), these algorithms are greedy and are only proven to find the best subset for a certain type of design criteria.

2.3. Design Criteria

Selection of training populations is an optimal experimental design problem, and the work on the optimal experimental designs has a long and rich history (Smith, 1918; Kiefer, 1959; Fisher, 1960; Fedorov, 1972; Atkinson and Donev, 1992; Pukelsheim and Rosenberger, 1993; Fedorov and Hackl, 2012; Silvey, 2013) and it is not a surprise that many different design criteria have been proposed. These criteria can be categorized into three major groups:

- Parametric design criteria which assume that the experimenter has specified a model before the training data is obtained. These criteria depend on a scalar function of the information matrix for the model parameters that give some indication about the sampling variances and covariances of the estimated quantities by the model. The estimated quantity might be some function of the model parameters or predictions from the model for target individuals. There are many designs obtained

by optimizing such criteria are referred to as $A-$, $D-$, $E-$, $G-$, etc… optimal designs (Kiefer et al., 1985). Bayesian design criteria use priors on the parameters of the models to evaluate the utility of designs.

- Nonparametric designs include criteria that are based on distance or similarity measures. For example, the maximin-distance design is a space-filling design that chooses a training population such that the minimum distance among the TRS is maximized (Johnson et al., 1990). Another such design is the minimax design (Johnson et al., 1990) where the training population is such that the maximum of the minimum distances from the training population to the rest of the CS or the TS is minimized. Space-filling designs aim to cover the experimental region with as few gaps or holes as possible. Unlike the parametric design criteria, minimax distance presumes no underlying model and, in turn, is suitable for situations where the model is unknown.
- Multiple design. The choice of an appropriate criterion requires knowledge about the model and what is required from the model. Multiple model optimal experimental design and compound optimization criteria try to overcome the choice issue by combining more than one criteria into one via some type of averaging. Alternatively, we can compare different designs using more than one criteria based on the dominance concept and use multi-objective optimization methods to decide on a certain design from out a set of Pareto optimal designs (Markowitz, 1952, 1968; Akdemir and Sánchez, 2016; Akdemir et al., 2019).

TrainSel allows users to use optimization criteria by letting them write their optimization functions and therefore can be used to search designs based on all of the above categories. Given the multitude of design criteria, this flexibility is one key advantage of TrainSel to its alternatives such as STPGA or TSDFGS.

2.3.1. Built in Criterion: CDmin

The STP involves the selection of TS from CS using optimization criteria. TrainSel is supplemented with a predefined design criterion CDmin which is related to the CDmean criteria in Laloë (1993), Laloë and Phocas (2003), Rincent et al. (2012). The main reason for implementing this design criterion as the only built-in design criterion is due to our specific interest in applying TrainSel to the design of single and multi-environmental GP experiments.

The built-in criterion CDmin depends on the linear mixed models. The linear mixed-effects model for a n-dimensional response variable y, $n \times p$ design matrix of fixed effects, $n \times q$ design matrix of random effects is defined as:

$$y = X\beta + Zu + \varepsilon;$$

where $\varepsilon \sim N_n(0, R)$ is independent of $u \sim N_q(0; G)$, $\beta \in \mathscr{R}^p$, G is a $q \times q$ covariance matrix and R is a $n \times n$ covariance matrix. The assumptions of the linear mixed-effects model imply $E(\boldsymbol{y}|X; Z) = X\beta$, $\boldsymbol{y} \sim N_n(X\beta; ZGZ' + R) = N_n(X\beta; V)$ with V defined as $V = ZGZ' + R$. For this model, the coefficient of determination matrix (Laloë, 1993; Laloë and Phocas, 2003; Rincent et al., 2012) of $\widehat{u}$ for predicting u is given by

$$(GZ'PZG) \oslash G$$

where $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$ and $\oslash$ expresses the elementwise division. The minimum of the selected diagonal elements of this matrix is called the CDmin. The minimum of the coefficient of determination takes on values between 0 and 1, and the designs that give higher values for this criterion are preferred to designs with lower values. The CDmin criterion follows the maximin decision rule, maximizing this criterion amounts to maximizing the utility for the worst case scenario, and it is suitable for making risk averse decisions.

Most authors use the mean of the selected diagonal elements of this matrix as the criterion, this is called the CDmean criterion. We have used CDmin instead of CDmean for several reasons. Firstly, the distribution of CD values along the diagonal for a given G matrix includes both the training samples and the remaining samples. The CD values that correspond to the training samples, as expected, form a different cluster (high values of CD) than the cluster of CD values corresponding to the samples that are not selected (low values of CD) and therefore we have a bimodal distribution for the CD values. Secondly, if the aim is to improve the generalization performance of the resulting model we prefer to move the lower part of this distribution to the right, i.e., the maximin decision amounts to improving the worst case CD value in this distribution which leads to the CDmin approach. Thirdly, the purpose of this article is not to compare effect of using different selection criteria but to show that TrainSel can be easily adopted to many different selection criteria.

Alternatively, we could approach the bimodality by restricting the mean measure to be calculated only on the set difference of the CS and the TRS or on a predefined TS. It should be trivial to apply any of these modifications with TrainSel. We stress here that the choice among the many different optimization criteria require thorough analysis, but this is beyond the aims of this paper.

We use two parameterizations of the above mixed model: In the first parameterization, we assume that $G = \sigma_k^2 K$ and $R = \sigma_e^2 I$ where σ_k^2 and σ_e^2 are the variances of the random terms u and e correspondingly and K is a relationship matrix of the same dimension as G. In the second parameterization $G = K \otimes V_k$ and $R = I \otimes V_e$ where V_k and V_e are covariance matrices that relate to the effects in u and e using Kronecker structured covariances.

The first model is useful for modeling random effects u related by a relationship matrix K. The STP for this model involves the selection of a predefined size set from the levels of the random term u that also correspond to factor levels in the rows (and columns) of K for labeling.

The second model is useful for modeling factor levels that correspond to the rows (and columns) of K in several related environments. The covariance of these random effects in several environments is given by V_k and similarly, the covariance of the residual effects in these environments is given by V_e. In this case, we want to select predefined sizes of sets from the factor levels that correspond to the rows (and columns) of K to be labeled in the corresponding environments.

The purpose of the X matrix in the mixed models above is to account for fixed effects. If the rows of the X matrix

corresponding to the conditions in a given environment are heterogeneous, then, in addition to selecting the levels of the random effect in the TRS, we would like to arrange the training sample optimally to the conditions expressed in the rows of X. In these cases, we are looking to identify a TRS that is an ordered subset of the CS. If no X matrix is specified or if the rows of X are homogeneous within environments the order of the assignments will not matter. In this case, STP involves the selection of an unordered sample as TRS.

2.4. Datasets and Applications

In this section, we describe the datasets, simulations, and related analysis. We are testing Trainsel with four applications: The first application deals with STP for GP of hybrid performance, the second with a design of multi-environmental GS experiment. The third application deals with STP for an image recognition problem. Our final application on splines regression entails simultaneous selection of design points among a set of candidates and allocation of knots through the range of the explanatory variables.

2.5. Application 1: Wheat Data for Hybrid Performance Prediction

This dataset was published in Liu et al. (2016) and was used in a similar context in Guo et al. (2019). The genetic dataset included the marker data (90 k SNP array based on an Illumina Infinium genotyping platform) for 135 elite winter wheat individuals adapted to Central Europe. A total of 1,604 F1 hybrids were generated in a factorial crossing scheme with 120 inbred individuals serving as female and 15 inbred individuals serving as male parents.

All genomic data for the wheat data for hybrid performance prediction application were obtained from the Dryad Digital Repository (doi: 10.5061/dryad.461nc). All related phenotypic data were obtained from the Digital Repository (doi: 10.5447/IPK/2016/11). Marker information for the hybrids was deduced from the parental individuals.

All individuals were evaluated in up to six environments. The adjusted means over environments for each of the 1,604 F1 hybrids for 7 traits (gluten content, kernel hardness, protein content, SDS volume, starch content, test weight, 1,000-kernel weight) were treated as the labels for the traits.

After removing the hybrids that came from parents with partial phenotypic data, we were left with 795 hybrids (full factorial crosses between 15 males and 53 females with complete phenotypic data). We have complete phenotypic data for all of these 795 hybrids in this application. Nevertheless, in practice, the evaluation of each of the hybrids involves making the cross between the corresponding parents and evaluating them in phenotypic trials, which are time-consuming and expensive. It is, therefore, desirable to reduce the costs involved in the generation and phenotypic evaluation by using a subset of all possible hybrids in the experiments and to use the data generated from these experiments for training genomic prediction models to make inferences about the phenotypic performance of untested hybrids.

In this application, we examine STP for hybrid performance prediction, i.e., we would like to select a prespecified size subset (50, 75, 100, 200 hybrids) of all possible 795 hybrids for training and use the phenotypic data from the TRS to predict the performance of the remaining hybrids. The TRSs were determined either by TrainSel using the CDmin criterion or by random sampling (repeated 30 times). The remaining hybrids were used as the TS where the prediction accuracies were evaluated using the correlation or the mean squared error between the predicted genotypic values and the observed phenotypes.

We only used the additive effects when calculating the CDmin criterion values through use of an additive relationship calculated from the marker scores. It is possible to include other effects such as dominance by supplementing the additive effects matrix with a dominance relationship matrix.

2.6. Application 2: Wheat Data for Multi-Environmental GS Experiment Design

We have obtained this dataset from https://triticeaetoolbox.org/wheat. The genotypic data included 989 individuals genotyped for 24,740 markers. All of these individuals had complete phenotypic data on plant height and stripe rust severity from three environmental trials. Using this data we have performed a cross-validation experiment where we explored the potential of STP for the multi-environmental design of GS experiments. We varied the number of overlapping individuals between the environments intending to see the effect on the predictive ability for the untested individuals.

We start each replication of the experiment by randomly selecting 240 individuals as the CS and the remaining individuals as the TS. Given the candidate individuals, we assume would like to construct an experiment in tree environments each of which can accommodate a fixed number of individuals (20, 40, 60, 80). To see how the replication affects the maximum CDmin values we also restrict the total number of individuals in the whole experiment to multiples of 1.2, 1.5, 2, 2.5, 3 of the number of individuals in each environment. Note that, restricting the total number of individuals to a multiple of 1.2 of the number of individuals allowed in each of the environments correspond to almost total replication (we did not use a factor of 1 because this value corresponds to a different type of combinatorial problem), on the other hand, a multiple of 3 corresponds to no replication, the intermediate values allow some amount of replication. We have assumed that the covariance of genotypic values between all trials pairs were 0.7 and we have assumed that the residuals were independent within and between trials. Besides, we have assumed that the heritabilities of both experiments were the same and equal to 0.5. We repeated this experiment 15 times and for each replication, we record the maximum CDmin value obtained and we also check the accuracy of the model in the TSs by calculating the correlation of the trait values in the TS and corresponding predictions from models based on different TRSs.

2.7. Application 3: MINST Datasets for Image Recognition

Image classification refers to the task of predicting the kind of objects in images. To train image classification models we need labeled images as training data. In this context, the purpose of STP would be to identify a subset of images to be labeled from out of a larger set of images.

In this application, we used a standard image classification data, the MINST fashion dataset, obtained using the "tf.keras.datasets" module, which consists of 28 × 28 grayscale images of 70,000 in 10 categories. The original data is split into two parts, the training set has 60,000 images and the test set has 10,000 images. In both the training and test datasets, the different classes were equally represented.

We performed the following experiment with this dataset: We started each replication of the experiment by identifying 1,000 samples at random from the original training set of size 60,000 as candidates. The number of samples from each class in the CS were arbitrarily set as 500, 450, 400, 350, 300, 250, 200, 150, 100, and 50 to assure an unbalanced CS. We chose a TRS of 100 or 200 samples out of the CS using TrainSel with the maximin distance criterion and using the distances among the 794 image features of samples in the CS. In addition, 100 random samples of sizes 100 and 200 were taken from the same CS as random TRSs. For each TRS, we recorded the entropy for the class distributions in the TRSs, the loss, and the accuracy for the predictions in the TS. We used the same 4-layer convolutional deep neural network prediction model for all the TRSs, these models were trained using the Keras R package (Allaire and Chollet, 2018). This experiment was repeated 50 times.

2.8. Application 4: STP for Splines Regression

Spline regression is a commonly used regression technique for modeling nonlinear relationships between a continuous response and continuous explanatory variables. In this technique the ranges of the explanatory variables are divided into bins using points which are called knots and the response is modeled with a piecewise polynomial with a set of extra constraints (continuity, continuity of the first derivative, and continuity of the second derivative) at the knots.

A commonly used form of splines, namely the natural cubic splines, uses cubic segments. The model for a natural cubic spline that relates the response y to the input variable x can be expressed as

$$y = \beta_0 + \beta_1 x + \beta_2 (x-k_1)_+ + \beta_3 (x-k_2)_+ + \ldots + \beta_6 (x-k_p)_+ + \sigma_\varepsilon^2$$

where

$$(x-k)_+ = \begin{cases} 0, \text{ if } x < k \\ x-k, \text{ if } x \geq k \end{cases}$$

and $k_1, k_2, \ldots, k_p$ are the knot positions that are to be specified as hyper-parameters. Due to this dependence the model matrix for this model will be written as $X(k)$. The qubic spline is a linear model, therefore, the formula for D-optimality criteria for this model can be expressed as $D(k) = |X(k)'X(k)|$ and its value depends on the choice of the knots. A "good" design maximizes the value of this function, i.e., we need to select the design points and also find the best knots for the selected set of design points.

In this simulation exercise, we show that we can simultaneously pick a TRS of design points out of a set of candidates and set the knot positions using TrainSel, i.e., we want to select a set of x values from a set of given candidates and find values of $k_1, k_2, \ldots, k_p$ that maximizes $D(k)$. Just like in other supervised learning scenarios, we assume we have no access to the values of the response apriori, their values will be observed only in the TRS and these along with the selected optimal knots will be used to fit the cubic spline model. The model will be used in the prediction of the response and the predicted response values in the CS will be compared to the true value of the response (the function value at x) by calculating mean squared errors. The results obtained by the optimization approach will be compared to the same size random sample of x selected from the CS and with the standard approach that involves placing knots at equally spaced quantiles of the range of the x values (Ruppert, 2002) in the CS.

In each replication of the experiment, we started with a 1,000 candidate x values sampled uniformly between 0 and 1. We selected 200 (or 300) x values from these candidate values and also determine the placement of 15 knots. Following the benchmark experiments in Ruppert (2002) we generated our response variables from four different functions (namely logit, sine, bump, spahat functions). More details on these functions and the generation of the response values are given in the **Supplementary Material**. The mean squared error for the predictions from the optimized set with optimized knots and random TRSs with equally spaced quantile knots were compared. This experiment was replicated 30 times.

3. RESULTS AND DISCUSSION

3.1. Application 1: Wheat Data for Hybrid Performance Prediction

The results of the application on hybrid performance are summarized by the boxplots in **Figure 2** for two traits. The results for the remaining five traits were summarized in **Supplementary Figure 2**. Preliminary analysis with the wheat data indicated that the hybrids selected as training by maximizing the CDmin criterion, provided more accurate prediction models for predicting the remaining hybrids as compared to models based on a random sample of hybrids. The relative efficiency of the optimized samples depended on the number of hybrids selected in the TRS, and also on the trait. Nevertheless, there was a clear optimized trend overall. The relative performance of the optimized TRS to random samples is minimal when the sample size were as low as 50, and it peaked for about sample size of 100, this relative efficiency decreased as the sample size increases. These results indicated that the CDmin criterion was a useful method for selecting wheat hybrids for predictive performance. In our opinion, hybrid prediction problems provide a perfect situation to exploit the STP approaches.

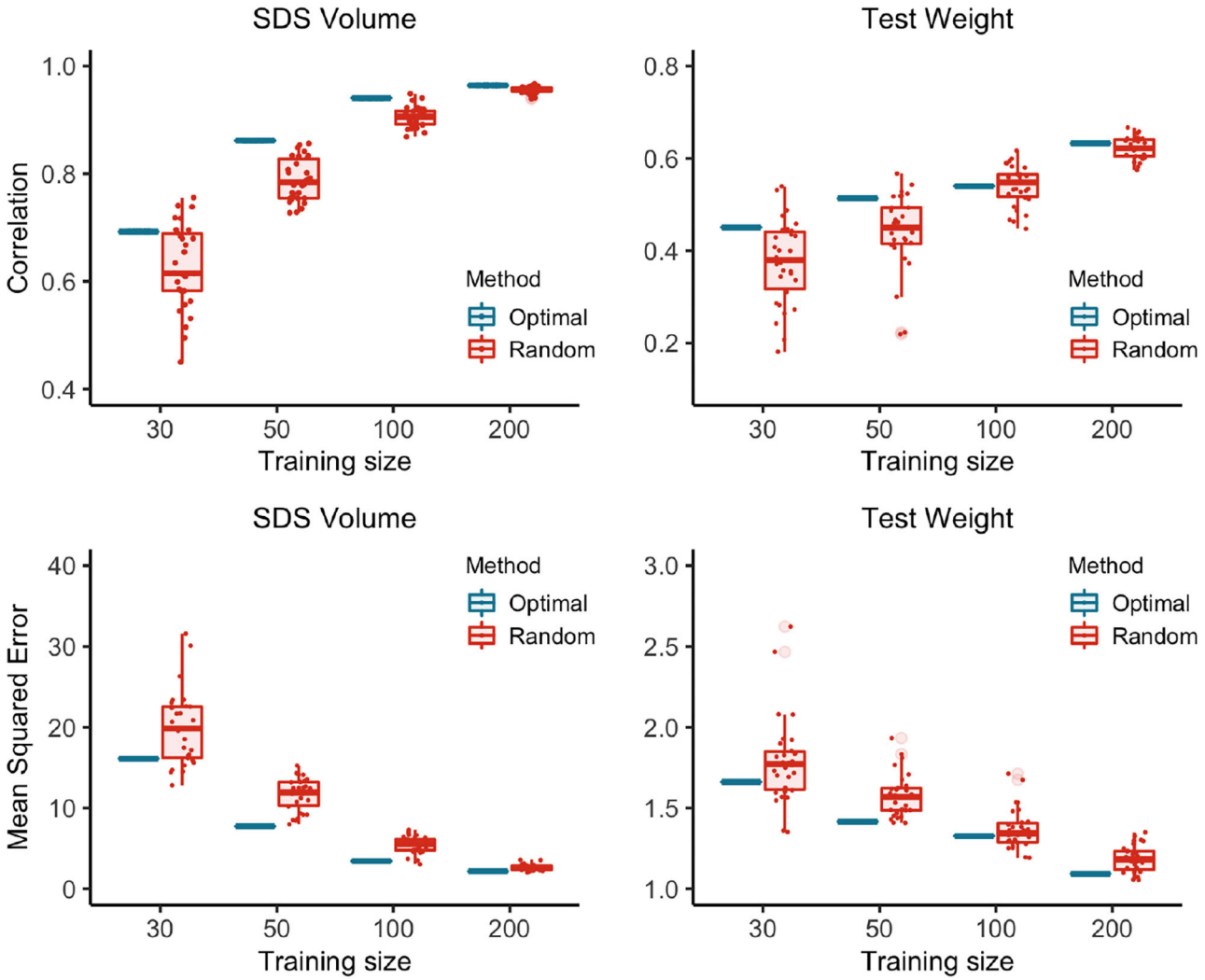

FIGURE 2 | The correlations and the mean squared errors between the predicted and observed trait values of the hybrids in the test data. There is an advantage in using optimized training samples for this dataset. The correlations and mean squared errors between the predicted and observed trait values of the hybrids in the test sets were significantly better for the optimized samples than the correlations and mean squared errors of the predicted and observed for the random samples.

3.2. Application 2: Wheat Data for Multi-Environmental GS Experiment Design

When designing a multi-environmental GS experiment, we would like to allocate individuals in environments so that we have a representative sample of individuals in each environment and, at the same time, have genetically similar individuals across environments. Genomic information is not utilized when designing experiments using classical methods such as randomized block design, and therefore, these designs are expected to perform worse than designs that make use of genomic information.

The CDmin values of the optimal samples on the first row of **Figure 3** indicate that CDmin values are maximized for intermediate amount of replication between the experiments. Since, the square root of the CD relates directly to the expected accuracy, we can use this information to decide on the size and amount of replication for a multi-environmental GS experiment.

The second and third rows of **Figure 3** showed the attained accuracy for optimal samples and random samples for plant height and stripe rust. As we can see the optimal experiments had better accuracy compared to the random experiments at all experiment sizes, levels of replication and for both of the traits. The trends in the observed accuracies for both the random samples and the optimized samples followed the trends observed in the CDmin values in the first row of the **Figure 3**.

These results demonstrated that optimally designed multi-environmental GS experiments can boost prediction accuracies as compared to randomized block designs. We note here that designing multi-environmental experiments with a large

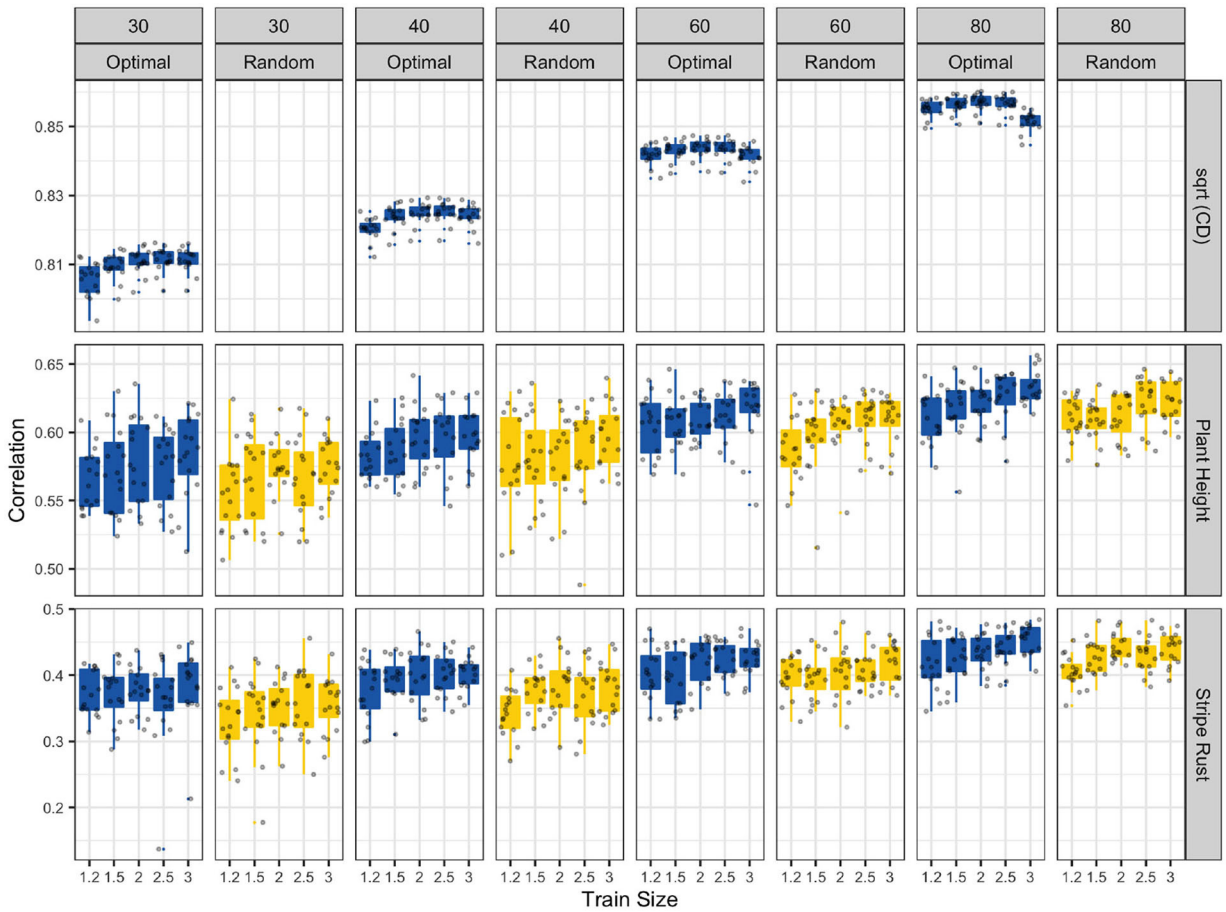

FIGURE 3 | Optimally designed multi-environmental GS experiments can boost prediction accuracies. In the first row, the CDmin values of the optimal samples show that the CDmin values are maximized for the intermediate amount of replication between the experiments. The second and third rows of figure show the attained accuracy for optimal samples and random samples for plant height and stripe rust.

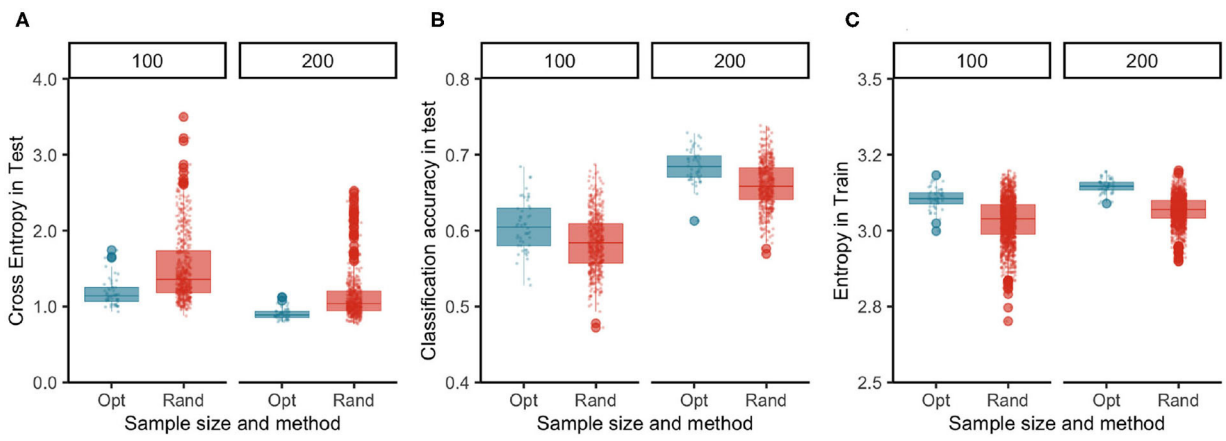

FIGURE 4 | Images selected optimally in the TRS have higher entropy in their label distributions than of the random samples **(C)** and the generalization performance of the model measured by both loss **(A)** and accuracy **(B)** functions in the test dataset indicate that optimally selected samples yield better models than the ones built on random samples.

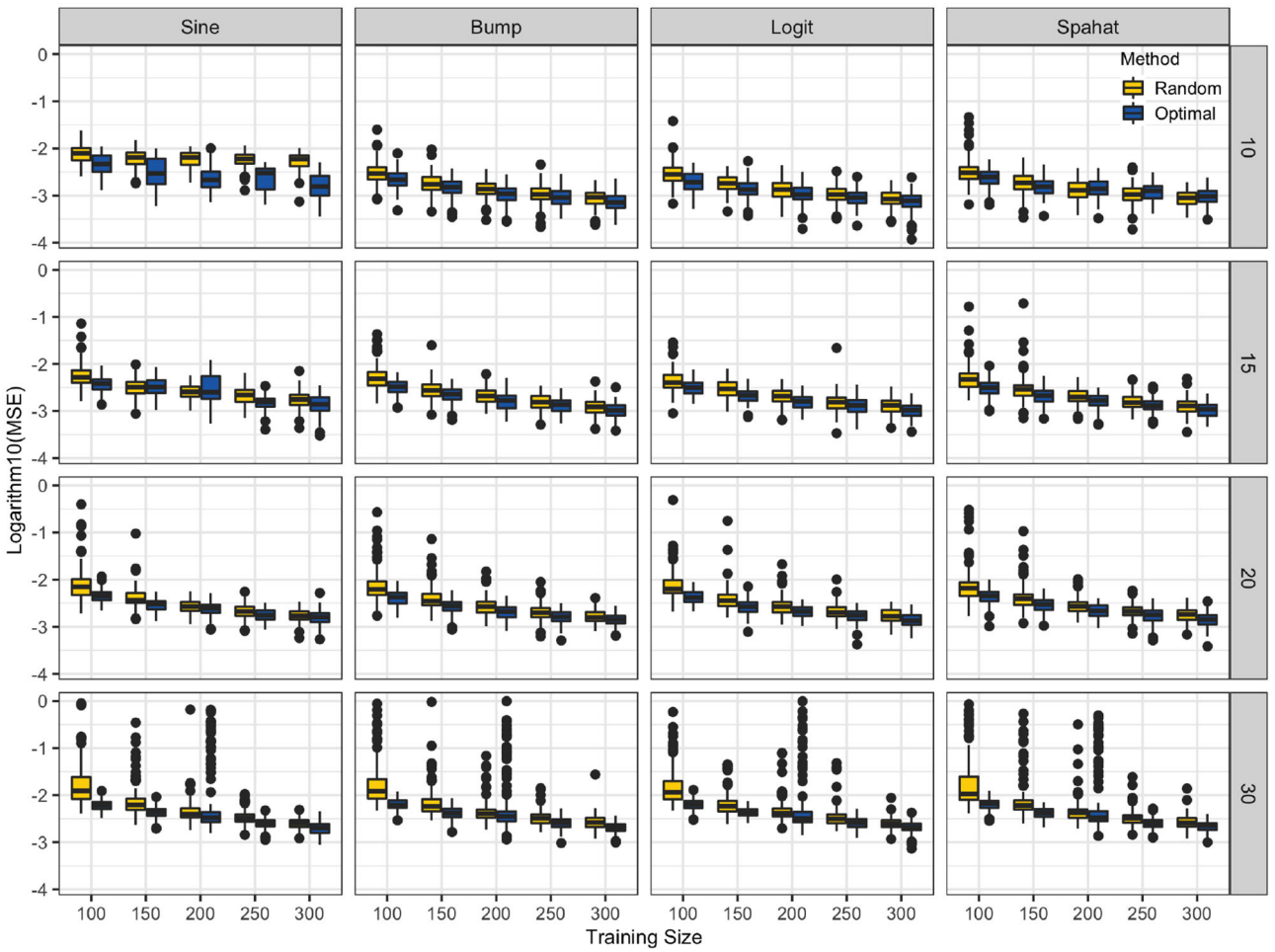

FIGURE 5 | TrainSel spline application. The logarithm of the mean square errors (y-axis) splines models trained on random samples with knots (10, 15, 20, and 30) located at equally spaced sample quantiles. Optimal training size (x-axis) and knots were selected by optimizing the D-optimality criterion for the different number of knots and different sample sizes for a set of functions. At each combination of sample size and the number of knots mean squared errors are lower for the latter approach. Although, in very few cases random sample performed slightly better than the optimized samples, the general trend is in favor of the optimized approach.

number of candidate individuals can be computationally costly. A useful strategy in these cases involves reducing the size of the candidate set to a manageable size by selecting a optimal subset from the full candidate set using suitable design criterion and using the reduced candidate set in the design of the multi-environmental experiment.

3.3. Application 3: MINST Datasets for Image Recognition

The results of this experiment are summarized in **Figure 4**. The TRS identified by TrainSel using the maximin distance criterion had higher entropy in their label distributions on average compared to those of random samples for both TRS sizes (**Figure 4**). Entropy is a widely used measure for quantifying inhomogeneity, impurity in machine learning applications. The predictions from the models trained on the optimal TRS were on average more accurate and had lower cost as measured by sparse cross-entropy.

Note that, in this application, we have started each replication of the experiment with an unbalanced CS. Entropy is a measure of balance in the label distributions, and entropy of the label distributions in the TRSs selected at random mirrors the unbalance in the CS. In addition, optimally selected samples have higher entropy values meaning that the labels for the samples were more evenly distributed, and this resulted in models with better accuracy, i.e., the percentage of correctly classified examples were higher (**Figures 4A–C**). In addition, the lower values of the loss function in the test data for optimal samples indicated that the estimates of probabilities used for the classification of observations lead to more confident decisions with more confident class probability estimates.

3.4. Application 4: STP for Splines Regression

The results of the splines experiment are summarized in **Figure 5**. For all combinations of the number of knots, the

number of TRS sizes, the optimally designed experiments where both knot placements and selected samples in the TRS were decided by optimizing the D-optimality criterion have resulted in splines models with lower mean squared error values as compared to the splines models trained on random samples with knots located at equally spaced sample quantiles. This was true for all of the four different response surfaces we have tested.

This example used TrainSel used to optimize a mixed integer optimization problem. Mixed integer programming finds many applications in plant breeding, for instance, it can be used in optimizing sequencing resources (Gonen et al., 2017; Cheng et al., 2020), estimating parental combinations to balance gains and inbreeding (Brisbane and Gibson, 1995; Jannink, 2010; Heslot et al., 2015), or genomic mating (Akdemir and Sánchez, 2016).

4. CONCLUSIONS

TrainSel provides algorithms for the optimization of mixed-integer problems. It was written with the STP problems in focus. The main use cases are given below:

1. Identifying a TRS from a larger CS for labeling especially when per sample cost of labeling is relatively high.
2. Design of experiments based on any user-defined design criteria or with built-in mixed model-based criteria.
3. Design of single or multi-environmental genomic prediction/selection experiments where the phenotyping is the major constraining factor.
4. TrainSel can also be used in other combinatorial optimization problems. Some examples of such problems include max clique, independent set, vertex cover, knapsack, set covering, set partitioning, feature subset selection (for supervised and unsupervised learning), traveling salesman, job scheduling problems.

The best feature of TrainSel is where we combine training set selection with a particular experimental design, and this option has not been implemented in any other STP software.

Reasons for using this package are as follows:

1. Most of the existing STP or statistical design software (such as TSDFGS, AlgDesign; Wheeler, 2004) will optimize only a few built-in optimization criteria. You can use TrainSel easily with your own design criteria.
2. Existing STP or statistical design software (such as STPGA, TSDFGS, AlgDesign) will optimize a single criterion at a time, but TrainSel offers an additional better possibility, i.e., we can specify multiple objectives that must be optimized simultaneously.
3. TrainSel uses a memetic evolutionary algorithm which in our experiments achieved better convergence than a simple genetic algorithm which was the basis for STPGA and TSDFGS.
4. The ability to handle ordered or unordered samples, with or without replication, along with several numerical variables to optimize user-defined functions makes this package a flexible general optimization tool.

We have illustrated with several applications that the benefits of using TrainSel in STP problems. These applications were mostly related to GP and GS, however, one of the major claims of this article is that the same techniques can be used for any supervised learning problem where labeling samples is the main bottleneck for obtaining the training data. We have exemplified this with two applications, one in image classification and another one related to spline regression.

5. IMPLEMENTATION AND USAGE

TrainSel is implemented in R with most of the code written in Rcpp. Sample usage is illustrated in the Supplementary and also in the help files within the package documentation. The source code and installation details are provided at https://github.com/TheRocinante-lab/TrainSel.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: referenced in the article.

AUTHOR CONTRIBUTIONS

DA: conception and design of the work, R and Rcpp programs, drafting the article, and critical revision of the article. JI: drafting the article and critical revision of the article. SR: critical revision of the article. All authors contributed to the article and approved the submitted version.

FUNDING

Results have been achieved within the framework of the first transnational joint call for research projects in the SusCrop ERA-Net Cofound on Sustainable Crop production, with funding from Department of Agriculture, Food and the Marine grant No.2017EN104. This project has also received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 818144, and also the Severo Ochoa Program for Centres of Excellence in R&D. JI was supported by the Beatriz Galindo Program (BEAGAL18/00115) from the Ministerio de Educación y Formación Professional of Spain and the Severo Ochoa Program for Centres of Excellence in R&D from the Agencia Estatal de Investigación of Spain, grant SEV-2016-0672 (2017-2021) to the CBGP.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.655287/full#supplementary-material

REFERENCES

Akdemir, D. (2017). *STPGA: Selection of Training Populations by Genetic Algorithm. R package version 5.2.1)*. doi: 10.1101/111989

Akdemir, D., Beavis, W., Fritsche-Neto, R., Singh, A. K., and Isidro-Sánchez, J. (2019). Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity* 122, 672–683. doi: 10.1038/s41437-018-0147-1

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci. Rep*. 9, 1–15. doi: 10.1038/s41598-018-38081-6

Akdemir, D., and Sánchez, J. I. (2016). Efficient breeding by genomic mating. *Front. Genet*. 7:210. doi: 10.3389/fgene.2016.00210

Akdemir, D., Sanchez, J. I., and Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel. Evol*. 47:38. doi: 10.1186/s12711-015-0116-6

Allaire, J., and Chollet, F. (2018). *keras: R Interface to'keras'. R Package Version 2.2. 0*.

Atkinson, A., and Donev, A. (1992). *Optimum Experimental Designs*. Oxford: Clarendon.

Brisbane, J., and Gibson, J. (1995). Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions. *Theor. Appl. Genet*. 91, 421–431. doi: 10.1007/BF00222969

Cericola, F., Jahoor, A., Orabi, J., Andersen, J. R., Janss, L. L., and Jensen, J. (2017). Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. a case of study in advanced wheat breeding lines. *PLoS ONE* 12:e0169606. doi: 10.1371/journal.pone.0169606

Cheng, H., Xu, K., and Abraham, K. J. (2020). Optimizing sequencing resources in genotyped livestock populations using linear programming. *BioRxiv [Preprint]*. doi: 10.1101/2020.06.29.179093

de Bem Oliveira, I., Amadeu, R. R., Ferr ao, L. F. V., and Mu noz, P. R. (2020). Optimizing whole-genomic prediction for autotetraploid blueberry breeding. *Heredity* 125, 437–448. doi: 10.1038/s41437-020-00357-x

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., et al. (2011). RCPP: seamless R and C++ integration. *J. Stat. Softw*. 40, 1–18. doi: 10.18637/jss.v040.i08

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Elsevier.

Fedorov, V. V., and Hackl, P. (2012). *Model-Oriented Design of Experiments, Vol. 125*. Springer Science & Business Media.

Fischetti, M., and Lodi, A. (2010). Heuristic in mixed integer programming. *Wiley Encyclop. Oper. Res. Manage. Sci*. doi: 10.1002/9780470400531.eorms0376

Fisher, R. A. (1960). *The Design of Experiments*. New York, NY: Hafner.

Glover, F. W., and Kochenberger, G. A. (2006). *Handbook of Metaheuristics, Vol. 57*. Springer Science & Business Media.

Gonen, S., Ros-Freixedes, R., Battagin, M., Gorjanc, G., and Hickey, J. M. (2017). A method for the allocation of sequencing resources in genotyped livestock populations. *Genet. Select. Evol*. 49:47. doi: 10.1186/s12711-017-0322-5

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., et al. (2019). Optimal designs for genomic selection in hybrid crops. *Mol. Plant* 12, 390–401. doi: 10.1016/j.molp.2018.12.022

Haines, L. M. (1987). The application of the annealing algorithm to the construction of exact optimal designs for linear-regression models. *Technometrics* 29, 439–447. doi: 10.1080/00401706.1987.10488272

He, S., Schulthess, A. W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., et al. (2016). Genomic selection in a commercial winter wheat population. *Theor. Appl. Genet*. 129, 641–651. doi: 10.1007/s00122-015-2655-1

Heslot, N., Jannink, J.-L., and Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Sci*. 55, 1–12. doi: 10.2135/cropsci2014.03.0249

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence*. MIT Press. doi: 10.7551/mitpress/1090.001.0001

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theor. Appl. Genet*. 128, 145–158. doi: 10.1007/s00122-014-2418-4

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genet. Select. Evol*. 42:35. doi: 10.1186/1297-9686-42-35

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *J. Stat. Plann. Infer*. 26, 131–148. doi: 10.1016/0378-3758(90)90122-B

Kadam, D. C., Rodriguez, O. R., and Lorenz, A. J. (2021). Optimization of training sets for genomic prediction of early-stage single crosses in maize. *Theor. Appl. Genet*. 134, 687–699. doi: 10.1007/s00122-020-03722-w

Kiefer, J. (1959). Optimum experimental designs. *J. R. Stat. Soc. Ser. B* 21, 272–319. doi: 10.1111/j.2517-6161.1959.tb00338.x

Kiefer, J. C., Brown, L., Olkin, I., and Sacks, J. (1985). *Jack Carl Kiefer Collected Papers: Design of Experiments*. Springer. doi: 10.1007/978-1-4613-8505-9

Laloë, D. (1993). Precision and information in linear models of genetic evaluation. *Genet. Select. Evol*. 25, 557–576. doi: 10.1186/1297-9686-25-6-557

Laloë, D., and Phocas, F. (2003). A proposal of criteria of robustness analysis in genetic evaluation. *Livest. Prod. Sci*. 80, 241–256. doi: 10.1016/S0301-6226(02)00092-1

Liu, G., Zhao, Y., Gowda, M., Longin, C. F. H., Reif, J. C., and Mette, M. F. (2016). Predicting hybrid performances for quality traits through genomic-assisted approaches in central European wheat. *PLoS ONE* 11:e0158635. doi: 10.1371/journal.pone.0158635

Lorenz, A. J., and Smith, K. P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Sci*. 55, 2657–2667. doi: 10.2135/cropsci2014.12.0827

Mangin, B., Rincent, R., Rabier, C.-E., Moreau, L., and Goudemand-Dugue, E. (2019). Training set optimization of genomic prediction by means of ethacc. *PLoS ONE* 14:e0205629. doi: 10.1371/journal.pone.0205629

Markowitz, H. (1952). Portfolio selection. *J. Fin*. 7, 77–91. doi: 10.1111/j.1540-6261.1952.tb01525.x

Markowitz, H. M. (1968). *Portfolio Selection: Efficient Diversification of Investments, Vol. 16*. Yale University Press.

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Mitchell, T. (1974). An algorithm for the construction of "d-optimal" experimental designs. *Technometrics* 16, 203–210. doi: 10.1080/00401706.1974.10489175

Neyhart, J. L., Tiede, T., Lorenz, A. J., and Smith, K. P. (2017). Evaluating methods of updating training data in long-term genomewide selection. *G3* 7, 1499–1510. doi: 10.1534/g3.117.040550

Nguyen, N., and Miller, A. (1992). A review of some exchange algorithms for constructing discrete d-optimal designs. *Comput. Stat. Data Anal*. 14, 489–498. doi: 10.1016/0167-9473(92)90064-M

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3* 8, 2889–2899. doi: 10.1534/g3.118.0200311

Olatoye, M. O., Clark, L. V., Labonte, N. R., Dong, H., Dwiyanti, M. S., Anzoua, K. G., et al. (2020). Training population optimization for genomic selection in miscanthus. *G3* 10, 2465–2476. doi: 10.1534/g3.120.401402

Ou, J.-H., and Liao, C.-T. (2019). Training set determination for genomic selection. *Theor. Appl. Genet*. 132, 2781–2792. doi: 10.1007/s00122-019-003387-0

Pukelsheim, F., and Rosenberger, J. (1993). Experimental designs for model discrimination. *J. Am. Stat. Assoc*. 88, 642–649. doi: 10.1080/01621459.1993.10476317

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., et al. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* 192, 715–728. doi: 10.1534/genetics.112.141473

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Stat*. 11, 735–757. doi: 10.1198/106186002853

Silvey, S. (2013). *Optimal Design: An Introduction to the Theory for Parameter Estimation, Vol. 1*. Springer Science & Business Media.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika* 12, 1–85. doi: 10.1093/biomet/12.1-2.1

Wheeler, B. (2004). *Algdesign. The R Project for Statistical Computing*.

Yu, X., Leiboff, S., Li, X., Guo, T., Ronning, N., Zhang, X., et al. (2020). Genomic prediction of maize microphenotypes provides insights for optimizing selection and mining diversity. *Plant Biotechnol. J.* 18, 2456–2465. doi: 10.1111/pbi.13420

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer RF-N declared a past co-authorship with one of the authors, DA, to the handling editor.

frontiers in Genetics

MINI REVIEW
published: 24 May 2021
doi: 10.3389/fgene.2021.667358

Improving Genomic Prediction Using High-Dimensional Secondary Phenotypes

Bader Arouisse[1], *Tom P. J. M. Theeuwen*[2], *Fred A. van Eeuwijk*[1] *and Willem Kruijer*[1]*

[1] *Biometris, Wageningen University and Research, Wageningen, Netherlands,* [2] *Laboratory of Genetics, Wageningen University and Research, Wageningen, Netherlands*

In the past decades, genomic prediction has had a large impact on plant breeding. Given the current advances of high-throughput phenotyping and sequencing technologies, it is increasingly common to observe a large number of traits, in addition to the target trait of interest. This raises the important question whether these additional or "secondary" traits can be used to improve genomic prediction for the target trait. With only a small number of secondary traits, this is known to be the case, given sufficiently high heritabilities and genetic correlations. Here we focus on the more challenging situation with a large number of secondary traits, which is increasingly common since the arrival of high-throughput phenotyping. In this case, secondary traits are usually incorporated through additional relatedness matrices. This approach is however infeasible when secondary traits are not measured on the test set, and cannot distinguish between genetic and non-genetic correlations. An alternative direction is to extend the classical selection indices using penalized regression. So far, penalized selection indices have not been applied in a genomic prediction setting, and require plot-level data in order to reliably estimate genetic correlations. Here we aim to overcome these limitations, using two novel approaches. Our first approach relies on a dimension reduction of the secondary traits, using either penalized regression or random forests (LS-BLUP/RF-BLUP). We then compute the bivariate GBLUP with the dimension reduction as secondary trait. For simulated data (with available plot-level data), we also use bivariate GBLUP with the penalized selection index as secondary trait (SI-BLUP). In our second approach (GM-BLUP), we follow existing multi-kernel methods but replace secondary traits by their genomic predictions, with the advantage that genomic prediction is also possible when secondary traits are only measured on the training set. For most of our simulated data, SI-BLUP was most accurate, often closely followed by RF-BLUP or LS-BLUP. In real datasets, involving metabolites in Arabidopsis and transcriptomics in maize, no method could substantially improve over univariate prediction when secondary traits were only available on the training set. LS-BLUP and RF-BLUP were most accurate when secondary traits were available also for the test set.

Keywords: GBLUP, genomic prediction, secondary traits, selection indices, penalized regression, random forest

OPEN ACCESS

Edited by:
Diego Jarquin, University of Nebraska-Lincoln, United States

Reviewed by:
Roberto Fritsche-Neto, International Rice Research Institute (IRRI), Philippines
Paulino Pérez-Rodríguez, Colegio de Postgraduados (COLPOS), Mexico

***Correspondence:**
Willem Kruijer
willem.kruijer@wur.nl

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *12 April 2021*
Accepted: *14 April 2021*
Published: *24 May 2021*

Citation:
Arouisse B, Theeuwen TPJM, van Eeuwijk FA and Kruijer W (2021) Improving Genomic Prediction Using High-Dimensional Secondary Phenotypes. Front. Genet. 12:667358. doi: 10.3389/fgene.2021.667358

1. INTRODUCTION

Genomic prediction is increasingly applied as standard tool in many animal and plant breeding programs. Since it was first introduced by Meuwissen et al. (2001), the main objective of genomic prediction was to estimate the breeding values for unphenotyped (test) genotypes with only molecular markers, using a training population for which both phenotypic and genotypic data are available. Applications of genomic prediction facilitate the rapid selection of superior genotypes (genomic selection) and accelerate genetic progress in crop breeding.

At the same time, advances in high-throughput phenotyping and cell biology technologies provide increasing amounts of phenotypic data, in addition to the "primary" or "target" traits of interest, such as yield or disease resistance. Such additional traits are typically high-dimensional, and collected using various types of technology, e.g., remote-sensing (Araus et al., 2018), machine vision (Yang et al., 2020), and automation technology (Sun et al., 2019). Common situations are that secondary traits are measured (1) in the field, on the same plant as the target trait, but much earlier in the growing season (2) on entirely different plants, in controlled environments in phenotyping platforms. In both cases, the secondary traits are either observed only for the training set of genotypes, or also for the test set. In all cases however, the question is whether some of the secondary traits are associated with the target traits of interest, and whether these correlations are genetic. In a genomic prediction context, the question becomes when and how secondary traits can improve prediction for the target trait. This is well understood if there is only one secondary trait: accuracy for the target trait then improves when the heritability of the target trait is lower than the heritability of the secondary trait times the squared genetic correlation (Schulthess et al., 2016; Velazco et al., 2019). Here we focus on the more challenging situation with a large numbers of secondary traits, which is increasingly common since the arrival of high-throughput phenotyping.

The two main approaches to incorporate high-dimensional secondary traits in genomic prediction are the use of multiple relatedness matrices, and penalized selection indices. In the former approach, the target trait is modeled as the sum of genetic effects and effects from secondary traits. Both type of effects are random, and the relative importance of these contributions is estimated either using REML-estimates for variance components or cross-validation. Predictions for the test set are the sum of the BLUPs for the different effects. Examples of this approach are Fu et al. (2012), who obtained a high level of accuracy for predicting hybrid yield performance using gene expression data from the hybrid parents. Similarly, Riedelsheimer et al. (2012) reported moderate to high accuracies for yield-related traits using 120 metabolites in maize. Schrag et al. (2018) and Xiang et al. (2019) used different relatedness matrices corresponding to different types of -omics data. Two major limitations of multiple random-effects models are that (1) they cannot be used when secondary traits are only available on the training set; (2) they cannot distinguish between genetic and residual correlations among the target and secondary traits.

The second approach was recently proposed by Lopez-Cruz et al. (2020), who extended classical selection indices by imposing a LASSO or ridge penalty on the coefficients. This achieves a dimension reduction, replacing the secondary traits by a single selection index S, which is a linear combination of the original traits. The coefficients are chosen to maximize $h^2(S)\rho_G^2(Y,S)$, i.e., the heritability of S times the squared genetic correlation between S and the target trait (Y). Lopez-Cruz et al. (2020) found that on new data, this quantity was indeed much higher than for the classical (unpenalized) selection index. Despite this promising result, penalized selection indices have not yet been applied in a genomic prediction context. One possible reason may be that accurate estimates of genetic correlations between Y and each of the secondary traits are required, for which the availability of plant/plot-level observations is assumed.

In the present paper, we propose two new approaches to deal with large numbers of secondary traits, and compare these to the approaches described above, using simulated and real data. First, we define genomic prediction using alternative dimension reductions (LS-BLUP/RF-BLUP), relying on penalized regression (or random forest regression) of the target on the secondary traits. We then compute the bivariate GBLUP with the dimension reduction as secondary trait. Second, we extend existing multi-kernel methods by replacing the secondary traits by their genomic predictions, the main advantage being that genomic prediction for the test set is always possible, also when secondary traits are only measured on the training set. For simulated data (with available plot-level data), we will also use bivariate GBLUP with the penalized selection index as secondary trait (SI-BLUP).

2. MATERIALS AND METHODS

2.1. Distributional Assumptions

To a large extent we follow the notation of Runcie and Cheng (2019), assuming observations on traits $Y_1,\ldots,Y_{p+1}$, where each Y_j is a column vector. The first one ($Y_1 = Y_f$) is the focal or target trait, for which genomic predictions are required; $Y_2,\ldots,Y_{p+1}$ are the secondary traits. $Y_s = (Y_2^t,\ldots,Y_{p+1}^t)^t$ is the column vector containing all secondary traits; similarly, $Y = (Y_1^t,\ldots,Y_{p+1}^t)^t$ is the column vector containing all traits. We have in total $n = n_t + n_o$ genotypes, including n_o genotypes for which the target trait is observed (the training set), and n_t for which it is to be predicted (the t referring to test set). We will use subscripts t and o to indicate that we take the subset of values on the test, respectively training set, for example Y_o and $Y_{f,o}$.

The secondary phenotypes are either observed only on the training set (the CV1-scenario, using the terminology of Runcie and Cheng, 2019), or also for the test genotypes (CV2). Since our focus here is on variable selection and dimension reduction (rather than different cross-validation schemes), we will refer to these simply with scenarios 1 and 2, respectively. The $n \times n$ genetic relatedness matrix K is partitioned as:

$$K = \begin{pmatrix} K_{tt} & K_{to} \\ K_{ot} & K_{oo} \end{pmatrix},$$

where the $n_t \times n_o$ matrix K_{to} defines the relatedness between new (test) and observed (training) genotypes. We will also write $K_{t\cdot} = [K_{tt}\, K_{to}]$ and $K_{o\cdot} = [K_{ot}\, K_{oo}]$. Similarly, we can decompose the genetic and residual covariance matrices Σ^u and Σ^e as

$$\Sigma^u = \begin{pmatrix} \Sigma^u_{ff} & \Sigma^u_{fs} \\ \Sigma^u_{sf} & \Sigma^u_{ss} \end{pmatrix} = \begin{pmatrix} \Sigma^u_{f\cdot} \\ \Sigma^u_{s\cdot} \end{pmatrix},$$

$$\Sigma^e = \begin{pmatrix} \Sigma^e_{ff} & \Sigma^e_{fs} \\ \Sigma^e_{sf} & \Sigma^e_{ss} \end{pmatrix} = \begin{pmatrix} \Sigma^e_{f\cdot} \\ \Sigma^e_{s\cdot} \end{pmatrix},$$

where the scalars Σ^u_{ff} and Σ^e_{ff} are respectively the genetic and residual variance of the focal trait, and the matrices Σ^u_{ss} and Σ^e_{ss} contain the genetic and residual (co)variances of the secondary traits. The row-vectors Σ^u_{fs} and Σ^e_{fs} contain the genetic and residual covariance between the focal and the secondary traits.

The joint distribution of $Y = (Y_1, \ldots, Y_{p+1})$ is assumed to be

$$\begin{aligned} Y &= X\beta + U + E \\ &= \begin{bmatrix} Y_1 \\ \vdots \\ Y_{p+1} \end{bmatrix} = \begin{bmatrix} X_1\beta_1 \\ \vdots \\ X_{p+1}\beta_{p+1} \end{bmatrix} + \begin{bmatrix} U_1 \\ \vdots \\ U_{p+1} \end{bmatrix} + \begin{bmatrix} E_1 \\ \vdots \\ E_{p+1} \end{bmatrix} \\ &= \begin{bmatrix} Y_f \\ Y_s \end{bmatrix} = \begin{bmatrix} X_f\beta_f \\ X_s\beta_s \end{bmatrix} + \begin{bmatrix} U_f \\ U_s \end{bmatrix} + \begin{bmatrix} E_f \\ E_s \end{bmatrix}, \end{aligned} \quad (1)$$

where

$$U \sim N(0, \Sigma^u \otimes K), \quad E \sim N(0, \Sigma^e \otimes I_n). \quad (2)$$

The genetic covariances (Σ^u_{fs}) quantify the degree of overlap among genetic signals, based on which multivariate methods can potentially improve genomic prediction. The residual covariances (Σ^e_{fs}) are important when traits are measured on the same individuals; if measured on different individuals (typically, in a different experiment), Σ^e can assumed to be diagonal. Σ^u and Σ^e are usually unknown, and need to be estimated from the data. For p larger than $5 - 10$, this usually requires approximations. Below we describe several dimension reduction approaches, which reduce the dimensionality of the secondary phenotypes to 1, and exact REML-estimates of Σ^u and Σ^e can be obtained with standard software.

2.2. Genomic Prediction

The main objective is the prediction of the genetic effect $U_1 = U_f$, i.e., the breeding values for the focal trait, in particular for the test set ($U_{f,t}$). In our simulations we assess prediction accuracy in terms of the Pearson correlation (r) between the simulated and predicted genetic effects, on the test set. For real data, we consider the correlation between the predicted genetic effects and the trait values observed on the test sets. Although it is well-known that this is a biased estimator of the true accuracy (i.e., the correlation with the unknown genetic effect), the bias is likely to be constant among methods, as long as the target and secondary traits are observed on different plants (Runcie and Cheng, 2019).

2.3. Univariate GBLUP

The univariate GBLUP for $U_{f,t}$ is defined by

$$\begin{aligned} \hat{U}^{(\text{uni})}_{f,t} &= E(U_{f,t}|Y_{f,o}) = \hat{\Sigma}^u_{ff} K_{to} \hat{V}^{-1}(Y_{f,o} - X_{f,o}\hat{\beta}_f) \\ &= K_{to}K_{oo}^{-1}\hat{U}^{(\text{uni})}_{f,o}, \\ \hat{U}^{(\text{uni})}_{f,o} &= \hat{\Sigma}^u_{ff} K_{oo} \hat{V}^{-1}(Y_{f,o} - X_{f,o}\hat{\beta}_f), \\ \hat{V} &= \hat{\Sigma}^u_{ff} K_{oo} + \hat{\Sigma}^e_{ff} I_{n_o}, \end{aligned} \quad (3)$$

where $\hat{U}^{(\text{uni})}_{f,o}$ is the GBLUP for the training set, and REML-estimates of β_f and the variance components Σ^u_{ff} and Σ^e_{ff} are obtained from a univariate mixed model for Y_f. This is the best (univariate) linear unbiased predictor, at least given the true values of the variance components.

2.4. Multivariate GBLUP in Scenarios 1 and 2

The multivariate GBLUP in scenario 1 is

$$\begin{aligned} \hat{U}^{(m1)}_{f,t} &= E(U_{f,t}|Y_o) = (\hat{\Sigma}^u_{f\cdot} \otimes K_{to}) \hat{V}^{-1}(Y_o - X_o\hat{\beta}) \\ &= K_{to}K_{oo}^{-1}\hat{U}^{(m1)}_{f,o}, \\ \hat{U}^{(m1)}_{f,o} &= (\hat{\Sigma}^u_{f\cdot} \otimes K_{oo}) \hat{V}^{-1}(Y_o - X_o\hat{\beta}), \\ \hat{V} &= \hat{\Sigma}^u \otimes K_{oo} + \hat{\Sigma}^e \otimes I_{n_o}, \end{aligned} \quad (4)$$

where $\hat{U}^{(m1)}_{f,o}$ is the GBLUP for the training set, and REML-estimates of β and the variance components (matrices) Σ^u and Σ^e are obtained from the multivariate mixed model for Y_f and Y_s. As pointed out by Runcie and Cheng (2019), $\hat{U}^{(m1)}_{f,t}$ and $\hat{U}^{(\text{uni})}_{f,t}$ have the same form, but the "input" $\hat{U}_{f,o}$ differs.

The multivariate GBLUP in scenario 2 is

$$\begin{aligned} \hat{U}^{(m2)}_{f,t} &= E(U_{f,t}|Y_{f,o}, Y_s) \\ &= \begin{pmatrix} \hat{\Sigma}^u_{ff} \otimes K_{to} & \hat{\Sigma}^u_{fs} \otimes K_{t\cdot} \end{pmatrix} \hat{V}^{-1} \begin{pmatrix} Y_{f,o} - X_{f,o}\hat{\beta}_f \\ Y_s - X_s\hat{\beta}_s \end{pmatrix}, \\ \hat{V} &= \begin{pmatrix} \hat{\Sigma}^u_{ff} K_{oo} & \hat{\Sigma}^u_{fs} \otimes K_{o\cdot} \\ \hat{\Sigma}^u_{sf} \otimes K^t_{o\cdot} & \hat{\Sigma}^u_{ss} \otimes K \end{pmatrix} \\ &\quad + \begin{pmatrix} \hat{\Sigma}^e_{ff} I_{n_o} & \hat{\Sigma}^e_{fs} \otimes \begin{pmatrix} 0 & I_{n_o} \end{pmatrix} \\ \hat{\Sigma}^e_{sf} \otimes \begin{pmatrix} 0^t \\ I_{n_o} \end{pmatrix} & \hat{\Sigma}^e_{ss} \otimes I_n \end{pmatrix} \end{aligned} \quad (5)$$

where 0 denotes a $n_t \times n_o$ matrix of zeros. This differs from the CV2 prediction in Runcie and Cheng (2019), who described a two-step approach.

2.5. Dimension Reduction Using LASSO or Random Forests

Expressions (4) and (5) are valid regardless whether there is just a single secondary phenotype, or multiple ones. However, when the dimension of the secondary phenotype (p) is larger than $5 - 10$, estimation of the required

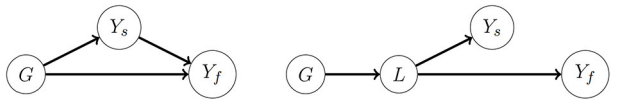

FIGURE 1 | Causal diagrams showing different assumptions about the mechanisms underlying genetic correlations between a high-dimensional secondary phenotype Y_s and a target (focal) trait Y_f. For ease of presentation, Y_s is represented by a single node; causal relationships among some of the secondary traits might exist. Outgoing arrows from the node G to a trait represent the genetic effect of all loci combined. The arrow $Y_s \rightarrow Y_f$ represents a causal effect from at least one of the secondary traits on the target trait. **(Left)** Some of the genetic correlations between Y_s and Y_f are the result of the causal effect $Y_s \rightarrow Y_f$; to some extent they may also be a consequence from correlation between the direct genetic effects $G \rightarrow Y_f$ and $G \rightarrow Y_s$ (see Kruijer et al., 2020 for more mathematical details). **(Right)** There is no causal effect $Y_s \rightarrow Y_f$, and genetic correlations between them may be induced by genetic effects on a latent trait L that is affecting both Y_s and Y_f. The **LS-BLUP** and **RF-BLUP** methods assume the left diagram, and reduce the dimension of Y_s first making a prediction $\hat{Y}_f$ using Y_s within the training set. Also the **GM-BLUP** method implicitly assumes the left diagram.

genetic covariances quickly becomes challenging and often infeasible (Zhou and Stephens, 2014; Zwiernik et al., 2017). Moreover, even if estimates of genetic covariance are available, the resulting predictions may be prone to overfitting. Reducing the dimension of the secondary phenotype appears to be a relevant strategy to deal with these issues.

Here we propose the dimension reduction $S = \hat{h}(Y_s)$, where $\hat{h}(Y_s)$ is a prediction of Y_f based on Y_s, obtained either with LASSO or random forests. Genomic prediction in scenarios 1 and 2 is then performed using (4) and (5), with $S = \hat{h}(Y_s)$ as secondary trait. We will refer to the resulting genomic predictions using LS-BLUP and RF-BLUP, depending on whether the dimension reduction was achieved by respectively LASSO or random forests. In a GWAS context, such dimension reductions have been used by van Heerwaarden et al. (2015) and Melandri (2019). The intuition behind this dimension reduction is that some of the secondary traits may have a causal effect on Y_f (**Figure 1**, left). Genomic prediction with LS-BLUP and RF-BLUP may then work well if $\hat{Y}_f$ captures most of the relevant genetic correlations. In our simulations described below, we also consider the situation where genetic correlations are not the result of a causal effect of Y_s on Y_f (for example, as in **Figure 1**, right panel). Because of the relatively small size of the populations considered here, the dimension reduction is computed on the same training set that is used for genomic prediction. This is of course not essential for this approach, and various sample splitting techniques may be of interest for larger populations; see the discussion section below.

When using RF-BLUP in the simulations described below, we used the R-package randomForest, with the default settings. Often however, a more accurate dimension reduction can be achieved by tuning various hyperparameters (like the number of trees), which we explore for the real data.

2.6. Dimension Reduction Using Selection Indices

In addition to the notation Y_s for the column vector containing all secondary traits, we will now also use $Y_s(j)$ for the column-vector containing the jth secondary trait, the dimension being either $n_o \times 1$ (scenario 1) or $n \times 1$ (scenario 2). We will use $Y_s^{(i)}$ for the row-vector containing all secondary traits for genotype i. Recall that the individual secondary traits are still labeled $Y_2, \ldots, Y_{p+1}$, Y_1 being the target trait.

A well-known alternative dimension reduction approach is to use a selection index $S = \sum_{j=1}^{p} \gamma_j Y_s(j)$, which is a linear combination of secondary traits, with coefficients such that the resulting index best predicts the genetic effect of the target trait (Falconer and Mackay, 1996). Assuming independent genetic effects (i.e., ignoring population structure), the $p \times 1$ vector γ of coefficients is obtained by minimizing, for each individual i, the expectation of $(U_f[i] - Y_s^{(i)}\gamma)^2$. The minimizing γ then equals the inverse variance-covariance of Y_s times the vector of genetic covariances between Y_s and Y_f, i.e., $\gamma^{SI} = {\Sigma_s}^{-1}\Sigma_{sf}^u$. To estimate γ^{SI} one could plug in estimates $\hat{\Sigma}_s$ and $\hat{\Sigma}_{sf}^u$, where $\hat{\Sigma}_s = \hat{\Sigma}_{ss}^u \otimes K_{oo} + \hat{\Sigma}_{ss}^e \otimes I_{n_o}$ is the estimated variance-covariance matrix of the secondary traits on the training population, and $\hat{\Sigma}_{sf}^u$ contains estimates of genetic covariances with the target trait. However, when the dimension (p) is large, Σ_{ss}^u and Σ_{ss}^e are difficult to estimate, and the selection index is likely to overfit, as some elements in Σ_{sf}^u may be large by chance, and receive too much weight.

To address these issues, Lopez-Cruz et al. (2020) proposed penalized selection indices, minimizing instead $E(U_f[i] - Y_s^{(i)}\gamma)^2 + \lambda J(\gamma)$, where $\lambda > 0$ is the penalty and $J(\gamma)$ is either $\sum_{j=1}^{p} \gamma_j^2$ (ridge penalty) or $\sum_{j=1}^{p} |\gamma_j|$ (LASSO penalty). $\lambda = 0$ gives the classical (unpenalized) SI. In case of a ridge penalty, the penalized SI is given by

$$\hat{\gamma}^{SI}(\lambda) = (\hat{\Sigma}_s + \lambda I_p)^{-1}\hat{\Sigma}_{sf}^u. \tag{6}$$

We will follow the implementation by Lopez-Cruz et al. (2020) in their R-package SFSI, where Σ_{sf}^u is estimated with MANOVA on the individual plant or plot-level data, and Σ_{ss}^u is estimated using the sample covariance matrix of the secondary traits. We emphasize that no multi-trait mixed-model of the form (1)–(2) is fitted. Moreover, the regularization only controls how $\hat{\Sigma}_s$ affects $\hat{\Sigma}_{sf}^u$; the estimates $\hat{\Sigma}_s$ and $\hat{\Sigma}_{sf}^u$ themselves are not regularized.

Following again (Lopez-Cruz et al., 2020), we use internal cross-validation within the training set to choose an appropriate value of λ, maximizing $h(S)\rho_G(S, Y_f)$. After selecting a value for λ, genomic prediction in scenarios 1 and 2 is performed using (4) and (5), with a single secondary trait, i.e., the selection index $\sum_{j=1}^{p} \gamma_j^{(\lambda)} Y_s[j]$. We will use SI-BLUP to refer to the genomic prediction obtained this way.

2.7. Genomic Prediction Using Multiple Relatedness Matrices

Another alternative to selection indices is to model the secondary traits using random effects (see e.g., Riedelsheimer et al., 2012;

Van De Wiel et al., 2016; Xu et al., 2016; Schrag et al., 2018; Xiang et al., 2019; Azodi et al., 2020). In addition to the genetic relatedness matrix K, these models use an additional relatedness matrix M derived from the secondary phenotypes, and assume that

$$Y_f = X_f\beta_f + U_f^{(\text{gen})} + V_f^{(\text{sec})} + E_f = X_f\beta_f + U_f^{(\text{gen})} + Y_s b_s + E_f, \quad (7)$$

where $U_f^{(\text{gen})} \sim N(0, \sigma_K^2 K)$ and $V_f^{(\text{sec})} \sim N(0, \sigma_M^2 M)$. We will call this the Multi-BLUP model (not to be confused with Speed and Balding, 2014, where the same type of model is used, but where genomic regions are represented by different relatedness matrices). The variance components σ_K^2, σ_M^2, and σ_E^2 can be estimated with REML or with cross-validation. For simplicity we consider only one type of secondary phenotypes. Similar to the equivalence between GBLUP and SNP-BLUP, the effects $V_f^{(\text{sec})}$ can be written as $Y_s b_s$, for a vector b_s of independent random effects with $N(0, p^{-1}\sigma_M^2)$ distribution. Hence, similar to the LS-BLUP and RF-BLUP, the Multi-BLUP approach implicitly assumes a causal effect of Y_s on Y_f (**Figure 1**, left), which is assumed to be linear, with random coefficients. The usual "genomic" prediction based on model (7) is

$$\hat{U}_{\text{Multi}} = \hat{U}_f^{(\text{gen})} + \hat{V}_f^{(\text{sec})}, \quad (8)$$

i.e., the sum of the BLUPs for the genetic and secondary trait effects. We put genomic between quotes because (8) is partly a phenotypic prediction: instead of the genetic component of the secondary traits, it directly relies on these traits themselves, which are assumed to be available on the test set. As a consequence, the use of (8) is limited to scenario 2.

To overcome these limitations we propose the GM-BLUP:

$$\hat{U}_{GM} = \hat{U}_f^{(\text{gen})} + \hat{U}_s^{(\text{gen})}\hat{b}_s, \quad (9)$$

where $\hat{b}_s$ is the vector of predicted random coefficients obtained from the Multi-BLUP model, and $\hat{U}_s^{(\text{gen})}$ is the matrix of GBLUPs for the secondary traits (either univariate or multivariate). These GBLUPs can of course also be computed in scenario 1. Apart from being the "genomic analogue" of (8), (9) can also be motivated by a causal model of the form

$$Y_f = X_f\beta_f + U_f + E_f + h(U_s), \quad (10)$$

as considered by Töpner et al. (2017) and Grotzinger et al. (2019). In contrast to the Multi-BLUP, GM-BLUP only depends on the genetic components of the secondary traits.

Finally, following many other authors (e.g., Riedelsheimer et al., 2012; Xu et al., 2016) we will also compute a prediction based on the secondary traits alone, using the model

$$Y_f = X_f\beta_f + V_f^{(\text{sec})} + E_f = X_f\beta_f + Y_s b_s + E_f, \quad (11)$$

and define the MBLUP

$$\hat{U}_M = \hat{V}_f^{(\text{sec})} = Y_s\hat{b}_s. \quad (12)$$

Again, this is to some degree a phenotypic prediction, and since the direct effects of the SNPs are ignored, the estimated effects $\hat{b}_s$ will differ from those obtained from model (7).

2.8. Simulations

We first compare the different methods on simulated data, with $p = 300$ secondary traits. We used existing genotypic data, from the Arabidopsis RegMap, containing 1,307 accessions genotyped with 214,051 SNPs (Horton et al., 2012). For each data-set we randomly selected 500 accessions, from which we randomly sampled a test set of 100 accessions. We randomly selected 1,500 SNPs with a minor allele frequency of at least 0.3. For each data-set we first simulated direct genetic effects (g_i) and residuals (r_i) for each accession i, and the final trait values were obtained using a structural equation model, describing functional relations between traits. More specifically, for each individual i, the $(p+1)\times 1$ vector of trait values is defined by $y_i = y_i\Lambda + g_i + r_i$, Λ being the $(p+1)\times(p+1)$ matrix of structural coefficients. The (k,l)th entry of Λ contains the effect of trait k on trait l, and the vectors g_i and r_i have zero mean Gaussian distributions with covariance matrices Σ^g and Σ^r, respectively. The joint distribution of all $n(p+1)$ trait values is then as in (1), with $\Sigma^u = \Gamma^t\Sigma^g\Gamma$ and $\Sigma^e = \Gamma^t\Sigma^r\Gamma$, where $\Gamma = (I-\Lambda)^{-1}$ (Gianola and Sorensen, 2004; Töpner et al., 2017; Kruijer et al., 2020).

The target trait is defined as $Y_f = Y_1 = \lambda(Y_2 + Y_3 + Y_4) + G_1 + R_1$, and we do not assume any functional relations among the secondary traits. Hence, if $\lambda \neq 0$, there is a causal effect from Y_2, Y_3, and Y_4 on Y_1, but the algorithms under consideration do not know which of the 300 secondary traits are the actual causal ones. We consider λ values on the grid $\{-1, -0.5, 0, 0.5, 1\}$. Σ^g has diagonal elements $(0.2, 0.7, \ldots, 0.7)$, i.e., the variances of the direct genetic effects are 0.2 for Y_f and 0.7 for each of the secondary traits. The off-diagonal elements corresponding to Y_1 vs. (Y_2, Y_3, Y_4) are $\rho_G\sqrt{0.2\cdot 0.7}$, where we choose $\rho_G \in \{-0.5, 0, 0.5\}$. Similarly, Σ^r has diagonal elements 0.8 for Y_f and 0.3 for the secondary traits, and the off-diagonal elements between Y_1 and (Y_2, Y_3, Y_4) are $\rho_E\sqrt{0.8\cdot 0.3}$, with $\rho_E \in \{-0.5, 0, 0.5\}$. The other off-diagonal elements in Σ^g and Σ^r are zero.

For the special case $\lambda = 0$ we have $\Gamma = I$, $\Sigma^u = \Sigma^g$ and $\Sigma^e = \Sigma^r$, and Y_f will have a heritability of 0.2. The secondary traits will have heritability 0.7, and there is no causal effect of (Y_2, Y_3, Y_4) on Y_1. Genomic prediction for Y_1 can however still benefit from the genetic correlation between these traits (which is present when $\rho_G \neq 0$). When $\lambda \neq 0$, the causal effect of $(Y_2 + Y_3 + Y_4)$ on Y_1 will introduce additional genetic and residual covariance in Σ^u and Σ^e.

For each of the 125 combinations of λ, ρ_G and ρ_E we simulate 50 data-sets; for each of them we predicted the simulated genetic effects for the test set, with the different methods.

2.8.1. Benchmark

In addition to the methods described above, we evaluate a benchmark prediction, by computing (4) and (5) for the four-dimensional mixed model with $Y_1 - Y_4$, using the true (simulated) variance components.

2.9. Data

To test the methods on real data, we consider four data-sets with various target and secondary phenotypes. To assess accuracy, each data set was randomly split into training (70%) and a test genotypes (30%). This was repeated 160 times, and we report accuracy averaged over the 160 test sets. Because of the required computing time, only 50 test sets were analyzed for RF-BLUP with hyper-parameter-optimization (for the Arabidopsis data-sets), and 30 test-sets for the maize data (for all methods). With one exception (mentioned below), the target and secondary phenotypes were measured on different plants; therefore, all bivariate mixed models were fitted with diagonal residual covariance (i.e., diagonal Σ^e in Equations 4 and 5).

The first two data sets were measured on the *A. thaliana* HapMap population, where 36 metabolites from Fusari et al. (2017) were used as secondary phenotypes and the kinship matrix was estimated based on one million imputed SNPs (Arouisse et al., 2020). Dataset 1 contains three target traits related to biotic and abiotic stress, from Thoen et al. (2017). In dataset 2, the target is the rosette fresh weight, measured in of the experiments of Fusari et al. (2017). This is the only dataset for which the residual covariance is non-diagonal.

In the third data set, we predicted the grain yield, plant height (PH) and flowering time (FT) of 388 inbred maize lines (*Z. mays*), using 5,760 transcripts (Azodi et al., 2020) as secondary traits. In this case, we selected for each data-set a subset of transcripts using the LASSO on the training set, following Azodi et al. (2020). In other words, the transcripts selected by LS-BLUP were also used for the other methods.

2.10. Data Availability

The data that support the findings of this study are available at:
https://doi.org/10.1105/tpc.19.00332 (Maize data)
https://doi.org/10.1105/tpc.17.00232 (*A. thaliana* Metabolite data)
https://doi.org/10.1111/nph.14220 (*A. thaliana* Phenotypes)
https://doi.org/10.1111/tpj.14659 (*A. thaliana* SNP data)

All data-sets (except the maize transcriptomics) are included in an Rdata file available at: https://figshare.com/s/5d01062711ce33bb327e.

2.11. Software and Computing Time

The required computing time is mainly driven by the complexity of fitting either a bivariate mixed model with a single relatedness matrix, or univariate mixed models with either one or two relatedness matrices. For the datasets considered here, each bivariate mixed model took between 20 and 50 s to fit, the univariate mixed models taking at most a few seconds. For complexity as function of n and p we refer to Zhou and Stephens (2014).

R-code for all methods is available at https://figshare.com/s/5d01062711ce33bb327e, where we mostly relied on asreml-R (Butler et al., 2009). Several open source alternatives are however available; in particular sommer (Covarrubias-Pazaran, 2016) for bivariate mixed models, and gaston for univariate mixed models. Using gaston's lmm.diago.likelihood function, the (univariate) GBLUP for large numbers of traits can be computed in only a few seconds, which is useful for the GM-BLUP method. For the dimension reduction in LS- and RF-BLUP we used the R-packages glmnet (Friedman et al., 2010), caret (https://cran.r-project.org/package=caret), and randomForest (Liaw and Wiener, 2002). For the maize data, LASSO and random-forest regression were performed in python, using the scikit-learn packages.

3. RESULTS

3.1. Simulations

Figures 2, **3** show the estimated accuracy as function of λ, i.e., the size of the causal effects of Y_2, Y_3, and Y_4 on the target trait Y_f (i.e., Y_1). We focus on three cases, with different values for the correlations between the direct genetic effects on $Y_1, \ldots, Y_4$, as well as the corresponding residuals (see section 2): (A) $\rho_G = 0.5$ and $\rho_E = -0.5$, (B) $\rho_G = \rho_E = 0$, and (C) $\rho_G = 0.5$ and $\rho_E = 0.5$. In scenario 1 (**Figure 2**) as well as scenario 2 (**Figure 3**), accuracies are generally higher when λ moves away from zero. This is expected, as the total genetic variance and heritability increase due to the causal effect, especially when ρ_G and λ have the same sign. When they have opposite sign, the lowest accuracy can occur at an intermediate value of λ [e.g., at $\lambda = -0.5$ in case of (A)].

The multi-trait benchmark with perfect information on the genetic and residual covariance between the target trait Y_f and secondary traits Y_2, Y_3, and Y_4 always outperforms univariate GBLUP, except when $\rho_G = \lambda = 0$, in which case accuracies are equal. When $\rho_G \neq 0$, the benchmark always benefits from the genetic correlations between the target trait and the secondary traits, even if the latter do not have a causal effect on Y_f.

The accuracy of univariate GBLUP varied between $r = 0.44$ and $r = 0.70$, while the benchmark had accuracy between $0.50 - 0.70$ (scenario 1) and $0.50 - 0.92$ (scenario 2). The difference between scenario 2 (secondary traits observed on the test set) and scenario 1 (secondary traits only observed on the training set) was bigger for large values of $|\lambda|$. This is because for large $|\lambda|$, the total genetic correlation (which is also a function of ρ_G) between Y_f and the causal secondary traits (Y_2, Y_3, and Y_4) is larger.

In absence of a causal effect $Y_s \rightarrow Y_f$ ($\lambda = 0$) and residual genetic and residual correlations having opposite sign (case A), our simulation setup appeared to be too challenging, and none of the methods performed better than univariate GBLUP. Something similar occurred in case C, for $\lambda = -0.5$. On the positive side, for large values of $|\lambda|$, both SI-BLUP and LS-BLUP have near-benchmark accuracy, where the latter did not rely on plot-level observations. In scenario 2, RF-BLUP appeared to be an interesting alternative, with somewhat lower accuracy on the extreme sides, but relatively good performance at unfavorable values of λ.

Prediction based on the secondary traits only (M-BLUP; only available in scenario 2) is generally one of the least successful. The multi-kernel methods (Multi-BLUP and GM-BLUP) are somewhere in between, GM-BLUP often having an accuracy similar to that of RF-BLUP. GM-BLUP appears to be slightly better than Multi-BLUP, but in most cases the difference is smaller than the standard errors of the accuracy estimates.

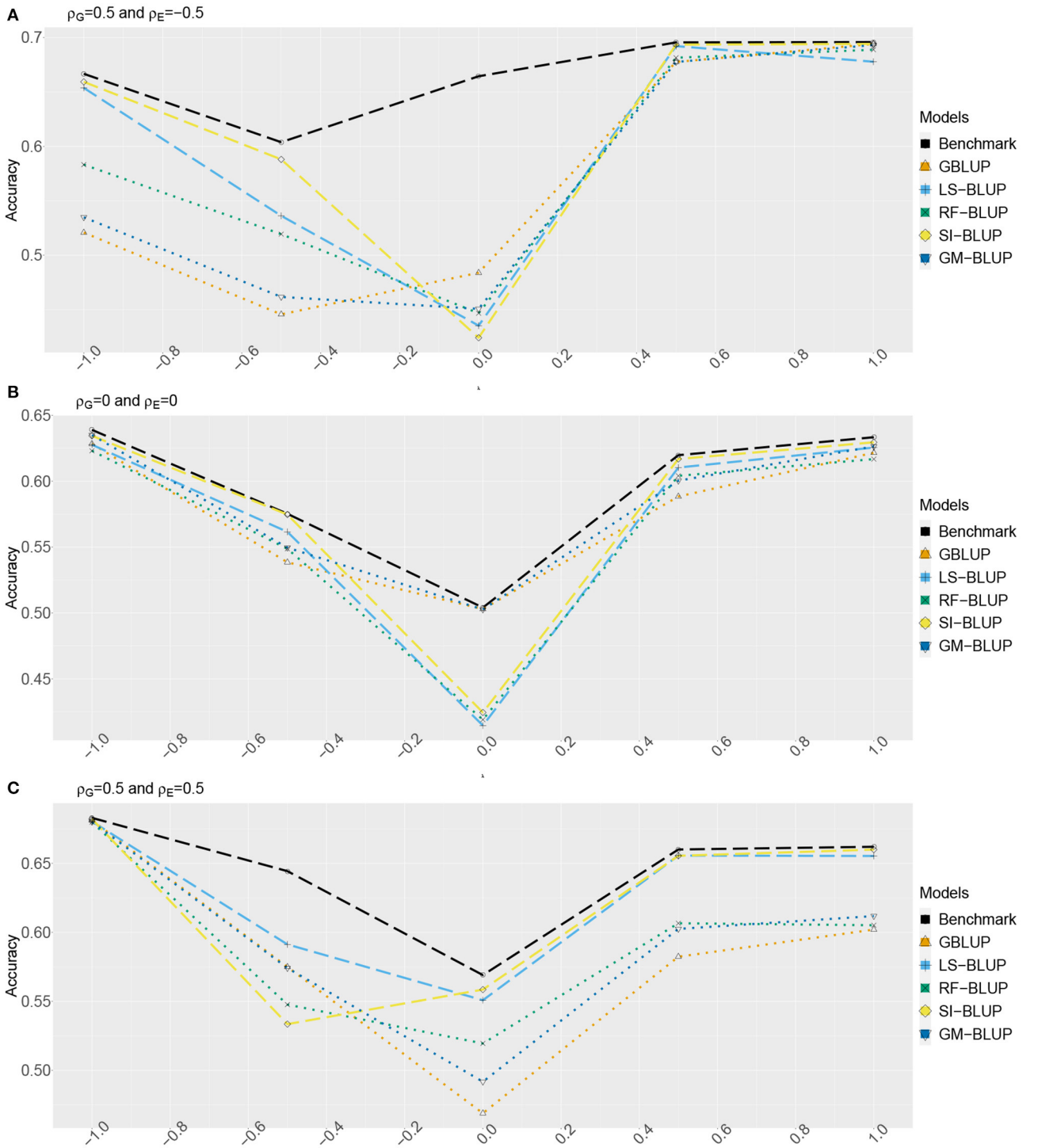

FIGURE 2 | Accuracy of genomic prediction methods in scenario 1, which for each value of λ is estimated from 50 simulated data-sets (standard errors between 0.011 and 0.042). "GBLUP" is the univariate GBLUP, and the benchmark is the multivariate GBLUP based on $Y_1, \ldots, Y_4$, using the true (simulated) values of the variance components (see section 2.8.1). Acronyms of the other methods are given in section 2; they use all secondary traits ($Y_2, \ldots, Y_{301}$), without knowledge of (Y_2, Y_3, Y_4) being causal. λ is the size of the causal effect of (Y_2, Y_3, Y_4) on Y_1. ρ_G is the correlation between the direct genetic effects on $Y_1, \ldots, Y_4$; similarly, ρ_E is the correlation between the non-genetic effects. The total genetic correlation is function of λ and ρ_G. **(A)** $\rho_G = 0.5$, $\rho_E = -0.5$, **(B)** $\rho_G = 0$, $\rho_E = 0$, and **(C)** $\rho_G = 0.5$, $\rho_E = 0.5$.

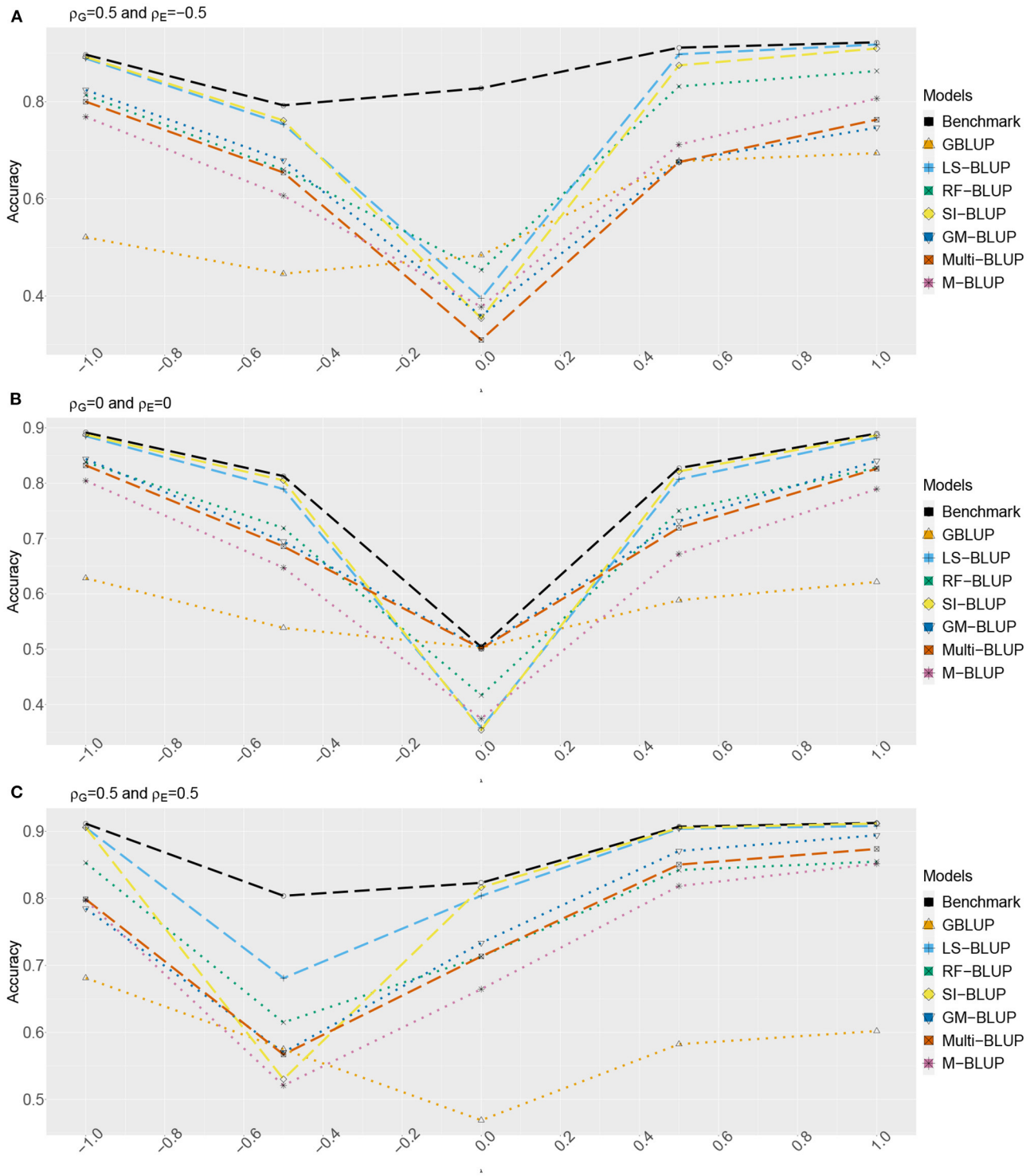

FIGURE 3 | Accuracy of genomic prediction methods in scenario 2, which for each value of λ is estimated from 50 simulated data-sets (standard errors between 0.014 and 0.051). "GBLUP" is the univariate GBLUP, and the benchmark is the multivariate GBLUP based on $Y_1, \ldots, Y_4$, using the true (simulated) values of the variance components (see section 2.8.1). Acronyms of the other methods are given in section 2; they use all secondary traits ($Y_2, \ldots, Y_{301}$), without knowledge of (Y_2, Y_3, Y_4) being causal. λ is the size of the causal effect of (Y_2, Y_3, Y_4) on Y_1. ρ_G is the correlation between the direct genetic effects on $Y_1, \ldots, Y_4$; similarly, ρ_E is the correlation between the non-genetic effects. The total genetic correlation is function of λ and ρ_G. **(A)** $\rho_G = 0.5$, $\rho_E = -0.5$, **(B)** $\rho_G = 0$, $\rho_E = 0$, and **(C)** $\rho_G = 0.5$, $\rho_E = 0.5$.

3.2. Arabidopsis and Maize Data

Tables 1, **2** contain the accuracies for datasets 1–4 described above, averaged over randomly sampled test sets (see section 2). Because the original individual plant (or plot) data were not available, we could not compute the SI-BLUP here.

In scenario 1 (**Table 1**), none of the multi-trait methods performed consistently better than univariate GBLUP. For the second trait in data-set 1 (Salt5), RF-BLUP had accuracy 0.09, vs. 0.03 for univariate GBLUP; the latter had highest accuracy for the first and third trait in dataset 1 (fungus, and drought and fungus stress combined).

The remainder of this section we focus on scenario 2 (**Table 2**), in which there were more substantial differences among methods. For all datasets, methods based on multiple relatedness matrices (Multi-BLUP and GM-BLUP) had accuracies similar to single-trait GBLUP. As in the simulations, GM-BLUP gave only a minor (if any) improvement over Multi-BLUP. The approaches based on dimension reduction of the secondary traits (LS-BLUP and RF-BLUP) appeared to give a substantial improvement over univariate GBLUP, e.g., from $r = 0.03$ to $r = 0.23$ (LS-BLUP) for the Salt5 trait in data-set 1, or from $r = 0.55$ to $r = 0.65$ (RF-BLUP) for Maize yield in data-set 3, with transcriptomics as secondary traits.

LS-BLUP had the highest accuracy in all Arabidopsis datasets, with a small but consistent improvement over RF-BLUP (0.02–0.03 higher), also when optimized with the caret/scikit-learn packages. This hyperparameter optimization appeared to be rather important for the Maize data; using the default settings from the randomForest package (as in the simulations), accuracy was considerably lower (for yield and the transcripts for example, $r = 0.65$ vs. $r = 0.51$).

For the maize data, RF/LS-BLUP improved accuracy for yield from around $0.64 - 0.65$ to $0.71 - 72$ when plant height and flowering time were included as secondary phenotypes, together with the transcriptome data. None of the other methods could exploit the additional data, and accuracies were similar to those obtained with the transcripts alone. Prediction based on the secondary traits alone (M-BLUP) had around zero accuracy in all Arabidopsis data-sets, but $r = 0.49 - 0.54$ for the maize data, similar to GBLUP and multi-BLUP.

4. DISCUSSION

Given the importance of genomic selection in plant breeding and the rapid development of phenotyping technology, it becomes increasingly important to know if and how the availability of additional phenotypic traits can improve prediction accuracy for a target trait. Here we proposed new methods to incorporate large numbers of such additional traits in genomic prediction, and compared these to existing methods, in simulated and real data. In many of the simulated data-sets, some of our methods indeed greatly improved univariate genomic prediction. In these cases, the accuracy was often close to that of penalized selection indices, without requiring plot-level data. In other cases, none of the methods did very much better than univariate prediction, while the multi-trait benchmark indicated that there is in fact scope for improvement. This happens especially when genetic and residual correlation have opposite sign. Moreover, our study indicates that current methods do not perform well when the secondary traits are available only on the training set (i.e., in scenario 1): while there was often some improvement in many of the simulations, accuracy in scenario 1 was hardly improved for any of the real data-sets.

While scenario 1 is probably most common, scenario 2 (secondary traits being also observed for the test set) may arise in a number of applications. In particular, it has become increasingly common to screen large collections for metabolites or other types of -omics data, and scenario 2 may also arise in a biomedical context when biomarkers could be used to predict disease. Our results for various stress traits in Arabidopsis showed that metabolites can indeed improve accuracy, even if they were measured in a different study. While Multi-BLUP and the LS- and RF-BLUP require balanced data, the GM-BLUP is more flexible, and can also handle an intermediate scenario where only some of the secondary traits are measured for all (or some of) the test genotypes.

Except SI-BLUP, all methods implicitly assume a causal relationship between the secondary traits and the target trait. In our simulations, accuracy was indeed suboptimal when this relationship was weak or absent. However, in these cases the SI-BLUP often performed poorly as well. The accuracy of LS-BLUP and RF-BLUP may be improved if one could successfully address the following two artifacts. First, the dimension reduction and genomic prediction should ideally be carried out on different subsets of the training set. In the populations we considered here, this however led to poor estimation of variance components and lower accuracies, because of the relatively small population size. We therefore used the whole training set for both dimension reduction and genomic prediction. The advantage of a larger training set seems to outweigh the incurred overfitting, but this may be different for larger populations, in which case sub-sampling strategies like bootstrap aggregation (bagging) might be useful. Second, specifically for LS-BLUP, the cross-validation in the first (dimension reduction) step appears to select too many variables. Often, this may still result in an accurate prediction $\hat{Y}_s$ on the training set, but for the prediction of breeding values on the test set that leads to overfitting. The methodology implemented in the hdi-package (Dezeure et al., 2015) might resolve this issue, by first assessing significance of secondary traits. Such improvements should at least guarantee an accuracy that is never (much) below that of univariate GBLUP. Finally, a remaining limitation of RF-BLUP and LS-BLUP is that the dimension reduction relies on phenotypic rather than genetic values, which is likely to stay sub-optimal in case genetic and residual correlations have opposite sign.

We attempted to improve existing multi-kernel methods with our GM-BLUP approach, replacing secondary traits by their genomic predictions. Unfortunately, this led to only minor improvements. In case secondary traits have high heritability, there is little shrinkage and genomic predictions and trait values are highly correlated, leading to similar accuracies. In case secondary traits have lower heritabilities, the methods may potentially differ more, but at the same time, in such a scenario

TABLE 1 | Prediction accuracy in scenario 1, for various target and secondary traits in Maize and Arabidopsis.

Data sets	Target trait	Secondary phenotypes	GBLUP	GM-BLUP	LS-BLUP	RF-BLUP	RF-BLUP*
1	Number of spreading lesions under fungus stress	Metabolites	0.23	0.22	0.20	0.21	0.21
	Fresh weight of the rosette under Salt_5 stress	Metabolites	0.03	0.00	0.07	0.09	0.09
	Number of spreading lesions under Drought_and_fungus stress	Metabolites	0.19	0.18	0.16	0.16	0.15
	Number of damaged leaves and feeding sites under Caterpillar_3 stress	Metabolites	0.10	0.09	0.06	0.10	0.10
2	Fresh weight	Metabolites	0.30	0.30	0.29	0.30	0.30
3	Flowering time (FT) [4]	Transcripts	0.54	0.55	0.55	0.53	0.55
	Plant height (PH)	Transcripts	0.54	0.55	0.55	0.53	0.51
	Yield	Transcripts + FT+PH	0.53	0.53	0.54	0.52	0.52
	Yield	Transcripts	0.55	0.55	0.55	0.55	0.55

*Acronyms of the methods are as in **Figures 2**, **3**. For RF-BLUP*, we used the randomForest package with the default settings; for RF-BLUP, hyper-parameters were optimized using the caret package (data-sets 1 and 2) or scikit-learn (data-set 3). For data-sets 1 and 2, reported accuracies are averages over* 160 *test sets (standard errors between 0.006 and 0.007), except for RF-BLUP, where 50 sets were used (SE between 0.010 and 0.014). In dataset 3, 30 test sets were used for all methods (SE between 0.006 and 0.03).*

TABLE 2 | Prediction accuracy in scenario 2, for various target and secondary traits in Maize and Arabidopsis.

Data sets	Target trait	Secondary phenotypes	GBLUP	M-BLUP	Multi-BLUP	GM-BLUP	LS-BLUP	RF-BLUP	RF-BLUP*
1	Number of spreading lesions under fungus stress	Metabolites	0.23	−0.04	0.21	0.22	0.31	0.28	0.28
	Fresh weight of the rosette under Salt_5 stress	Metabolites	0.03	0.09	0.08	0.07	0.23	0.20	0.19
	Number of spreading lesions under Drought_and_fungus stress	Metabolites	0.19	−0.02	0.16	0.17	0.27	0.25	0.23
	Number of damaged leaves and feeding sites under Caterpillar_3 stress	Metabolites	0.10	0.05	0.06	0.07	0.14	0.12	0.11
2	Fresh weight	Metabolites	0.30	0.00	0.29	0.30	0.32	0.30	0.28
3	Flowering time (FT) [4]	Transcripts	0.55	0.54	0.55	0.55	0.66	0.65	0.54
	Plant height (PH)	Transcripts	0.54	0.53	0.54	0.55	0.66	0.64	0.53
	Yield	Transcripts + FT+PH	0.53	0.49	0.50	0.52	0.72	0.71	0.49
	Yield	Transcripts	0.55	0.52	0.53	0.54	0.64	0.65	0.51

*Acronyms of the methods are as in **Figures 2**, **3**. For RF-BLUP*, we used the randomForest package with the default settings; for RF-BLUP, hyper-parameters were optimized using the caret package (data-sets 1 and 2) or scikit-learn (data-set 3). For data-sets 1 and 2, reported accuracies are averages over* 160 *test sets (standard errors between 0.006 and 0.012), except for RF-BLUP, where 50 sets were used (SE between 0.010 and 0.014). In dataset 3, 30 test sets were used for all methods (SE between 0.006 and 0.03).*

there is much less scope for improvement with multi-trait methods in the first place. Both Multi-BLUP and GM-BLUP were often less accurate than competing methods. To some extent this may be explained by the absence of variable selection, or, compared to RF-BLUP, the assumed linearity. Nonetheless, GM-BLUP extended the use of Multi-BLUP to scenario 1, without ever being less accurate.

For the case of a single secondary trait, Runcie and Cheng (2019) studied the bias in accuracy estimates, when these are based on the correlation with the observed phenotype, rather than with the (unobserved) genetic effect. This can become problematic when traits are measured on the same plants, in which case the amount of bias is likely to vary among methods, in particular when residual correlations between the target and secondary traits are large. For the Arabisopsis and maize data considered here, the bias should be constant, as all target and secondary traits were measured on different plants. No bias occurred for the simulated data, where we used the true genetic values to assess accuracy. Nevertheless, further work is needed to extend the methods presented here with reliable estimates of accuracy, also in the case of traits measured on the same plants. For the LS-BLUP, RF-BLUP and SI-BLUP, the parametric and semi-parametric accuracy estimates of Runcie and Cheng (2019) can in principle be computed, since all these methods reduce the dimension of the secondary traits to one. This would however require the sample-splitting or bagging schemes mentioned above, and it is an open question how the different accuracy estimates should be aggregated.

Statistical methods for high-dimensional data often benefit from initial screening, for example by removing variables with very low marginal correlation (see e.g., Fan and Lv, 2008). In the present context, screening should be based on heritability and genetic correlation with the target trait. This is however difficult for several reasons. First, as pointed out before, reliable estimates of these correlations require plot-level data, at least for the population sizes considered here. Moreover, bivariate mixed models need to be fitted for each secondary trait, increasing computation time. A more fundamental problem is that even if accurate estimates were available, it would be difficult to formulate an appropriate criterion and threshold. The well-known criterion for a single secondary trait (whose heritability times the squared genetic correlation with the target trait should exceed the heritability of the latter) cannot directly be generalized. For example, in one of our simulation settings (i.e., with $\lambda = 0$ and $\rho_G = 0.5$), each of the three relevant secondary traits (Y_2, Y_3, Y_4) has heritability 0.7, the heritability of the target trait being 0.2. Consequently, we have $0.7 \times \rho_G^2 < 0.2$ for each secondary trait individually, while at the same time genomic prediction using a mixed model for $Y_1 - Y_4$ is more accurate than with a mixed model for Y_1 alone.

More generally, the methods presented here could be extended in several ways. First, for all of them, prediction relies on the GBLUP: either bivariate GBLUP, or univariate GBLUP extended with additional relatedness matrices. This corresponds to a Gaussian prior on the marker effects, which could be generalized to a mixture of Gaussians and a point mass at 0, as for example in Bayes-R (Moser et al., 2015). Another extension would be the prediction of sensitivities to environmental covariates, which could then be used to predict new environments, as in Millet et al. (2019). In the LS- and RF-BLUP methods, a wider range of prediction methods could be considered to achieve the dimension reduction, such as elastic nets or gradient tree boosting. Ideally, this reduction is driven by genetic rather than phenotypic effects, and the dimension should not necessarily be reduced to one (like we did here), but to a data-driven number. Finally, it would be of interest to relax the linearity assumption on which most methods (except RF-BLUP) rely. Deep learning with feedforward or convolutional neural networks seems of particular interest here, especially for the relationship between target and secondary traits.

AUTHOR CONTRIBUTIONS

BA performed the research. WK, BA, and FE designed the research. BA and WK wrote the paper, with input from TT and FE. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Netherlands Scientific Organization for Research NWO-STW project 11145 Learning from Nature, and the EU project H2020 731013 (EPPN2020).

REFERENCES

Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., and Cairns, J. E. (2018). Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci.* 23, 451–466. doi: 10.1016/j.tplants.2018.02.001

Arouisse, B., Korte, A., van Eeuwijk, F., and Kruijer, W. (2020). Imputation of 3 million snps in the arabidopsis regional mapping population. *Plant J.* 102, 872–882. doi: 10.1111/tpj.14659

Azodi, C. B., Pardo, J., VanBuren, R., de los Campos, G., and Shiu, S.-H. (2020). Transcriptome-based prediction of complex traits in maize. *Plant Cell* 32, 139–151. doi: 10.1105/tpc.19.00332

Butler, D. G., Cullis, B. R., Gilmour, A. R., and Gogel, B. J. (2009). ASReml-R reference manual. *Release 3.0. Technical Report*, Queensland Department of Primary Industries, Australia.

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the r package sommer. *PLoS ONE* 11:e156744. doi: 10.1371/journal.pone.0156744

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, *p*-values and R-software HDI. *Stat. Sci.* 30, 533–558. doi: 10.1214/15-STS527

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics, 4th Edn.* Harlow: Prentice Hall. Available online at: https://www.worldcat.org/title/introduction-to-quantitative-genetics/oclc/422852955

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* 70, 849–911. doi: 10.1111/j.1467-9868.2008.00674.x

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Fu, J., Falke, K. C., Thiemann, A., Schrag, T. A., Melchinger, A. E., Scholten, S., et al. (2012). Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. *Theor. Appl. Genet.* 124, 825–833. doi: 10.1007/s00122-011-1747-9

Fusari, C. M., Kooke, R., Lauxmann, M. A., Annunziata, M. G., Enke, B., Hoehne, M., et al. (2017). Genome-wide association mapping reveals that specific and pleiotropic regulatory mechanisms fine-tune central metabolism and growth in arabidopsis. *Plant Cell* 29, 2349–2373. doi: 10.1105/tpc.17.00232

Gianola, D., and Sorensen, D. (2004). Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics* 167, 1407–1424. doi: 10.1534/genetics.103.025734

Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., et al. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* 3, 513–525. doi: 10.1038/s41562-019-0566-x

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., et al. (2012). Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nat. Genet.* 44, 212–216. doi: 10.1038/ng.1042

Kruijer, W., Behrouzi, P., Bustos-Korts, D., Rodríguez-Álvarez, M. X., Mahmoudi, S. M., Yandell, B., et al. (2020). Reconstruction of networks with direct and indirect genetic effects. *Genetics* 214, 781–807. doi: 10.1534/genetics.119.302949

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News* 2, 18–22. Available online at: https://www.researchgate.net/publication/228451484_Classification_and_Regression_by_RandomForest

Lopez-Cruz, M., Olson, E., Rovere, G., Crossa, J., Dreisigacker, S., Mondal, S., et al. (2020). Regularized selection indices for breeding value prediction using hyper-spectral image data. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-020-65011-2

Melandri, G. (2019). *Understanding drought tolerance in rice by the dissection and genetic analysis of leaf metabolism, oxidative stress status and stomatal behavior* (Ph.D. thesis). Wageningen University, Wageningen, Netherlands.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., et al. (2019). Genomic prediction of maize yield across European environmental conditions. *Nat. Genet*. 51, 952–956. doi: 10.1038/s41588-019-0414-y

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet*. 11:e1004969. doi: 10.1371/journal.pgen.1004969

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., et al. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet*. 44, 217–220. doi: 10.1038/ng.1033

Runcie, D., and Cheng, H. (2019). Pitfalls and remedies for cross validation with multi-trait genomic prediction methods. *G3* 9, 3727–3741. doi: 10.1534/g3.119.400598

Schrag, T. A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., et al. (2018). Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics* 208, 1373–1385. doi: 10.1534/genetics.117.300374

Schulthess, A. W., Wang, Y., Miedaner, T., Wilde, P., Reif, J. C., and Zhao, Y. (2016). Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. TAG. Theoretical and applied genetics. *Theor. Angew. Genet*. 129, 273–287. doi: 10.1007/s00122-015-2626-6

Speed, D., and Balding, D. J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res*. 24, 1550–1557. doi: 10.1101/gr.169375.113

Sun, J., Poland, J. A., Mondal, S., Crossa, J., Juliana, P., Singh, R. P., et al. (2019). High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. *Theor. Appl. Genet*. 132, 1705–1720. doi: 10.1007/s00122-019-03309-0

Thoen, M. P. M., Davila Olivas, N. H., Kloth, K. J., Coolen, S., Huang, P.-P., Aarts, M. G. M., et al. (2017). Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. *New Phytol*. 213, 1346–1362. doi: 10.1111/nph.14220

Töpner, K., Rosa, G. J. M., Gianola, D., and Schön, C.-C. (2017). Bayesian networks illustrate genomic and residual trait connections in maize (*Zea mays* l.). *G3* 7, 2779–2789. doi: 10.1534/g3.117.044263

Van De Wiel, M. A., Lien, T. G., Verlaat, W., van Wieringen, W. N., and Wilting, S. M. (2016). Better prediction by use of co-data: adaptive group-regularized ridge regression. *Stat. Med*. 35, 368–381. doi: 10.1002/sim.6732

van Heerwaarden, J., van Zanten, M., and Kruijer, W. (2015). Genome-wide association analysis of adaptation using environmentally predicted traits. *PLoS Genet*. 11:e1005594. doi: 10.1371/journal.pgen.1005594

Velazco, J. G., Jordan, D. R., Mace, E. S., Hunt, C. H., Malosetti, M., and van Eeuwijk, F. A. (2019). Genomic prediction of grain yield and drought-adaptation capacity in sorghum is enhanced by multi-trait analysis. *Front. Plant Sci*. 10:997. doi: 10.3389/fpls.2019.00997

Xiang, R., Berg, I. v. d., MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., et al. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci. U.S.A*. 116, 19398–19408. doi: 10.1073/pnas.1904159116

Xu, S., Xu, Y., Gong, L., and Zhang, Q. (2016). Metabolomic prediction of yield in hybrid rice. *Plant J*. 88, 219–227. doi: 10.1111/tpj.13242

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., et al. (2020). Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant* 13, 187–214. doi: 10.1016/j.molp.2020.01.008

Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* 11, 407–409. doi: 10.1038/nmeth.2848

Zwiernik, P., Uhler, C., and Richards, D. (2017). Maximum likelihood estimation for linear gaussian covariance models. *J. R. Stat. Soc. Ser. B* 79, 1269–1292. doi: 10.1111/rssb.12217

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 18 June 2021
doi: 10.3389/fgene.2021.680569

lme4GS: An R-Package for Genomic Selection

Diana Caamal-Pat[1], Paulino Pérez-Rodríguez[1], José Crossa[1,2]*, Ciro Velasco-Cruz[1], Sergio Pérez-Elizalde[1] and Mario Vázquez-Peña[3]*

[1] Department of Socioeconomics, Statistics, and Informatics, Colegio de Postgraduados, Texcoco, Mexico, [2] Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, [3] Department of Irrigation, Universidad Autónoma Chapingo, Texcoco, Mexico

Genomic selection (GS) is a technology used for genetic improvement, and it has many advantages over phenotype-based selection. There are several statistical models that adequately approach the statistical challenges in GS, such as in linear mixed models (LMMs). An active area of research is the development of software for fitting LMMs mainly used to make genome-based predictions. The lme4 is the standard package for fitting linear and generalized LMMs in the R-package, but its use for genetic analysis is limited because it does not allow the correlation between individuals or groups of individuals to be defined. This article describes the new lme4GS package for R, which is focused on fitting LMMs with covariance structures defined by the user, bandwidth selection, and genomic prediction. The new package is focused on genomic prediction of the models used in GS and can fit LMMs using different variance–covariance matrices. Several examples of GS models are presented using this package as well as the analysis using real data.

Keywords: genomic selection, genomic prediction, linear mixed model, lme4, kernel

OPEN ACCESS

Edited by:
Waseem Hussain, International Rice Research Institute (IRRI), Philippines

Reviewed by:
Jiabo Wang, Southwest Minzu University, China
Alexandre Bureau, Laval University, Canada

****Correspondence:***
Paulino Pérez-Rodríguez
perpdgo@gmail.com
José Crossa
jcrossa@cgiar.org

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *14 March 2021*
Accepted: *19 May 2021*
Published: *18 June 2021*

Citation:
Caamal-Pat D, Pérez-Rodríguez P, Crossa J, Velasco-Cruz C, Pérez-Elizalde S and Vázquez-Peña M (2021) lme4GS: An R-Package for Genomic Selection. Front. Genet. 12:680569. doi: 10.3389/fgene.2021.680569

INTRODUCTION

With the new, low-cost, high-throughput genotyping technologies of the last decade, a breeding selection paradigm called genomic selection (GS) has emerged (Meuwissen et al., 2001). GS combines molecular and phenotypic data to obtain the genomic estimated breeding values (GEBVs) of individuals that have been genotyped but not phenotyped (Bernardo and Yu, 2007; de los Campos et al., 2009; Hayes et al., 2009; VanRaden et al., 2009; Crossa et al., 2010). The main advantages of GS over family-based selection in breeding are that it reduces the cost per cycle and the time required for variety development. However, several factors could impact the accuracy of prediction; they occur at different levels and are influenced by several genetic, environmental, and statistical factors.

Complications arise in GS when determining (i) the size and diversity of the training population, (ii) the relationship between the training and testing sets, (iii) genetic complexity, and (iv) the heritability of the traits to be predicted. Challenges in GS are related to the high dimensionality of marker data, where, the number of markers is much larger than the number of observations, the multi-collinearity among markers, the cryptic interaction between

Abbreviations: BGLR, Bayesian generalized linear regression; BLR, Bayesian linear regression; BLUP, best linear unbiased prediction; GEBV, genomic estimated breeding value; GLMM, generalized linear mixed model; GS, genomic selection; LMM, linear mixed model; REML, restricted maximum likelihood.

markers, the complexity of the trait, sample size, correlation among markers, and the ever-present genotype × environment interaction. These complexities require parametric and semi-parametric statistical models, especially mixed models, Bayesian estimations, and, recently, deep machine learning methods that can deal appropriately with the usually large datasets (Crossa et al., 2017). This has led to computational challenges due to the data size and statistical challenges that include model fitting and parameter optimization. Therefore, the development of complete and simple computer packages to estimate the GEBV of the individuals to be selected under these complex scenarios is crucial for an efficient application of GS.

The first R software (R Core Team, 2021) developed for genome-based prediction was presented by de los Campos et al. (2009). Shortly afterward, Pérez et al. (2010) formally described the Bayesian linear regression (BLR) that allows fitting high-dimensional linear regression models including dense molecular markers, pedigree information, and several other covariates other than markers. The BLR R-package described by Pérez et al. (2010) allows including not only markers but also pedigree data jointly. Furthermore, Pérez et al. (2010) explained the challenges that arise when evaluating genomic-enabled prediction accuracy through random cross-validation (CV), as well as how to select the best choice of hyperparameters for the Bayesian models.

Linear mixed models play a fundamental role in GS and genomic-enabled predictions. This kind of models is widely used for predictions, although other models, such as nonlinear models, neural networks, and other machine learning models, could be used for this purpose. The standard linear mixed model of the form $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$, where, $\mathbf{y}$ is a response vector of dimension $n \times 1$; $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices for the fixed (β) and genotypic random ($\mathbf{u}$) effects, respectively; and two variance components are estimated $\mathbf{u} \sim MN(\mathbf{0}, \sigma_u^2\mathbf{K})$, with $\mathbf{K}$ being a known semidefinite variance–covariance matrix and $\mathbf{e} \sim MN(\mathbf{0}, \sigma_e^2\mathbf{I})$. In the context of GS, $\mathbf{K}$ could be the additive relationship matrix derived from the coefficient of co-ancestry (numerator relationship matrix $\mathbf{A}$), or it could be the genomic relationship matrix obtained from markers ($\mathbf{G}$). As shown below, there are several alternative ways of expressing the incidence matrix $\mathbf{Z}$ and the vector of random effects $\mathbf{u}$ when using the numerical relationship matrix ($\mathbf{A}$). Bayesian versions of linear regression models have been extensively developed, and their companion software largely distributed and used for research and extended to more complicated cases, for example, the introduction of genotype × environment interaction incorporating pedigree and environmental covariables (Jarquín et al., 2014).

Endelman (2011) developed the rrBLUP R-package, which is able to fit the basic linear mixed model with two variance components (σ_u^2 and σ_e^2) described before with the maximum likelihood or restricted maximum likelihood (REML) methods. As an extra facility, the rrBLUP computes the Gaussian kernel and the exponential kernel that usually account for small cryptic epistatic effects among the markers. The rrBLUP has a CV algorithm to measure the prediction accuracy of the models and shows rapid solutions of the mixed model equations for moderate-to-intermediate data sizes. More specialized computer software, such as the synbreed of Wimmer et al. (2012) and GEMMA of Zhou and Stephens (2012), were later developed.

Although the previously mentioned genomic software programs solve important genomic prediction problems (e.g., prediction in training and testing sets, CV, and estimation of variance parameters), they are separate software pieces without a unified statistical and computing framework. So from the user's perspective, having a single package implementing all the models to be fitted will save data preparation time and data analysis time. Thus, Pérez and de los Campos (2014) extended the original BLR R-package developed by Pérez et al. (2010) to a more general R-package, the Bayesian generalized linear regression (BGLR) that offers users a great variety of genomic models and methods in a unified computing software for data analysis. The BGLR is available at CRAN. The BGLR package includes several Bayesian regression models, including parametric variable selection and shrinkage methods, and semi-parametric procedures [Bayesian reproducing kernel Hilbert space (RKHS) regressions]. Many non-genomic applications are implemented as well, and response traits can be continuous or categorical (binary or ordinal). The Bayesian algorithm is based on a Gibbs sampler with scalar updates implemented in efficient routines written in C programming language. Furthermore, the BGLR is the main machinery for adapting other more complex genomic models, for example, the complex phenomenon of genotype × environment interaction including pedigree and environmental covariables (Jarquín et al., 2014). The BGLR is also used for assessing the marker effect × environment interaction of Lopez-Cruz et al. (2015) and for fitting Bayesian ridge regression and the Bayes B, as shown by Crossa et al. (2017), or for using the threshold model for ordinal data as did Montesinos-López et al. (2016), and for running all the Bayesian alphabet models.

Although linear mixed models are important tools for fitting GS models, Covarrubias-Pazaran (2016) mentioned like that current GS software includes only one random effect; and therefore, using genomic prediction for more complicated situations hybrid prediction using additive, dominance, and epistatic effects is not possible under the available models. The authors proposed likelihood-based software for fitting mixed models with multiple random effects that allow the user to specify the variance–covariance structure of random effects. Covarrubias-Pazaran (2016) presented an R-package called sommer for genomic prediction with three algorithms for estimating variance components: average information, expectation–maximization, and efficient mixed model association. Results from sommer were comparable with those of other software, and sommer was faster than its Bayesian counterparts.

The development of software for fitting linear mixed models is an active area of research. The use of pedigree and genomic-enabled prediction linear mixed models is crucial for advancing the application of genomic-assisted breeding. The lme4 package (Bates et al., 2015) for R (R Core Team, 2021) has efficient functions for analyzing linear mixed models and generalized linear mixed models (GLMMs). Some of the main features of lme4 are that (i) it is efficient for large dataset problems; (ii)

it handles any number of grouping factors, nested or cross-classified; and (iii) it can use a combination of sparse and dense matrix representations to facilitate the processing of large datasets at high computational speed.

However, the use of lme4 for genetic analysis has been limited because it does not allow using the correlation between individuals or groups of individuals. When individual lines or animals are related, the marginal likelihood must allow using this covariance between relatives. Vazquez et al. (2010) developed a package called pedigreemm that uses the lme4 but allows for correlations between levels of random effects, such as those due to genetic relationships between relatives expressed as pedigree relationships. The methodology of Vazquez et al. (2010) uses the numerator relationship matrix **A** (a positive-definite matrix) and subjects it to the Cholesky decomposition, where, the Cholesky factor (**L**) can be obtained from the pedigree information.

Based on the above considerations and some limitations in terms of the computing efficiency of some existing genomic-enabled prediction models, in this research, we describe the new lme4GS R-Package that is based on the lme4 software of Bates et al. (2015) that is available in CRAN. The lme4GS is focused on genomic-based prediction of GS and can fit mixed models with several different variance–covariance matrices. The lme4GS introduces fixed and random effects, and associated variance–covariance matrices, from which matrices for fixed and random effects ($\mathbf{X}$, $\mathbf{Z}_1, ..., \mathbf{Z}_q$, respectively) are obtained. The original variance–covariance matrices are introduced and transformed by using the Cholesky factorization or the eigenvalue decomposition of variance–covariance matrices and later used for defining the objective function (deviance function). Once the objective function has been defined, the optimization module optimizes the objective function and provides REML estimates of the parameters of interest.

MATERIALS AND METHODS

Consider the linear mixed model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (1)$$

where, $\mathbf{y}$ is a response vector of dimensions $n \times 1$, $\mathbf{X}$ is a matrix of fixed effects of dimensions $n \times p$, β is a vector of fixed effects of dimensions $p \times 1$, $\mathbf{Z}$ is an incidence matrix of dimensions $n \times r$, and $\mathbf{u}$ is a vector of random effects. We assume $\mathbf{u} \sim MN(\mathbf{0}, \sigma_a^2\mathbf{K})$ and $\mathbf{e} \sim MN(\mathbf{0}, \sigma_e^2\mathbf{I})$, with $\mathbf{K}$ a known variance–covariance matrix, and σ_a^2 and σ_e^2 are variance parameters associated with $\mathbf{u}$ and $\mathbf{e}$, respectively; furthermore, we assume that $\mathbf{u}$ and $\mathbf{e}$ are independently distributed. In the case of GS, the variance–covariance matrix can be derived from markers or from pedigree.

The linear mixed model (1) can be rewritten as;

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}^*\mathbf{u}^* + \mathbf{e}, \quad (2)$$

where, $\mathbf{Z}^* = \mathbf{Z}\mathbf{L}$, with $\mathbf{L}$ obtained from the Cholesky factorization of $\mathbf{K}$; alternatively, $\mathbf{Z}^* = \mathbf{Z}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}$ with $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ the matrices of eigenvectors and eigenvalues, respectively, obtained from the eigenvalue decomposition of $\mathbf{K}$, and $\mathbf{u}^* \sim MN(\mathbf{0}, \sigma_u^2\mathbf{I})$. Note that $\mathbf{Z}^*\mathbf{u}^*$ has the same distribution as $\mathbf{Z}\mathbf{u}$; that is, $\mathbf{Z}^*\mathbf{u}^* \overset{d}{=} \mathbf{Z}\mathbf{u} \sim \text{MN}(\mathbf{0}, \sigma_a^2\mathbf{Z}\mathbf{K}\mathbf{Z}')$.

Best Linear Unbiased Predictions

Once mixed model (2) is fitted, the conditional means of the random effects can be obtained, that is, $\widehat{\mathbf{u}}^*$.The best linear unbiased predictions (BLUPs) for $\mathbf{u}^*$ are obtained as follows: $\widehat{\mathbf{u}}^* = \widehat{\sigma}_u^2\mathbf{Z}^{*'}\widehat{\mathbf{V}}^{*-1}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})$ where, $\widehat{\mathbf{V}}^* = \widehat{\sigma}_u^2\mathbf{Z}^*\mathbf{Z}^{*'} + \widehat{\sigma}_e^2\mathbf{I}$, with $\widehat{\sigma}_e^2$, $\widehat{\sigma}_u^2$ and $\widehat{\beta}$ REML estimates of variance parameters and vector of fixed effects, respectively. The conditional means of random effects for the model in equation (1) are obtained as follows: $\widehat{\mathbf{u}} = \mathbf{L}\widehat{\mathbf{u}}^*$ if the Cholesky factorization is used, or alternatively, $\widehat{\mathbf{u}} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\widehat{\mathbf{u}}^*$ if the eigenvalue is used.

Prediction of New Observations

The main goal of GS is to predict new observations (phenotypic values) or simply obtain the BLUPs for random effects not present in the observed data but drawn from the same population as $\mathbf{u}$ and $\mathbf{e}$ (Gilmour et al., 2004). Assume that the random vector $\mathbf{u}$ and matrix $\mathbf{K}$ are partitioned as follows:

$$\mathbf{u} = \begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix}, \mathbf{K} = \begin{bmatrix}\mathbf{K}_{11} & \mathbf{K}_{12}\\ \mathbf{K}_{21} & \mathbf{K}_{22}\end{bmatrix},$$

the BLUPs for $\mathbf{u}_2$ are obtained as:

$$E\left(\mathbf{u}_2 \mid \mathbf{y}_1\right) = \mathbf{K}_{21}\mathbf{K}_{11}^{-1}\mathbf{u}_1. \quad (3)$$

In a more general case, model (1) can be extended to include more random effects, that is:

$$\mathbf{y} = \mathbf{X}\beta + \sum_{j=1}^{q}\mathbf{Z}_j\mathbf{u}_j + \mathbf{e}, \quad (4)$$

where, $\mathbf{Z}_j$ is a design matrix of random effects, and $\mathbf{u}_j$ is a vector of random effects, $j = 1, ..., q$, where, q corresponds to the number of random terms included in the model. We assume that $\mathbf{u}_j \sim MN(\mathbf{0}, \sigma_j^2\mathbf{K}_j)$ is independently distributed. Note that model (1) is a special case of model (4) obtained by setting $q = 1$, $\mathbf{Z} = \mathbf{Z}_1$, $\mathbf{u} = \mathbf{u}_1, \mathbf{K} = \mathbf{K}_1$, $\sigma_a^2 = \sigma_1^2$. Based on the same computational strategy used to rewrite model (1) as the model in (2), model (4) can be rewritten as:

$$\mathbf{y} = \mathbf{X}\beta + \sum_{j=1}^{q}\mathbf{Z}_j^*u_j^* + \mathbf{e}. \quad (5)$$

Implementation

The lme4GS package is an extension of the lme4 R-package (Bates et al., 2015); lme4GS development was inspired by existing R-packages, pedigreemm (Vazquez et al., 2010) and lme4qtl (Ziyatdinov et al., 2018), which are focused on quantitative trait locus (QTL) mapping association and linkage studies, whereas, lme4GS is focused on the problem of prediction in GS (Meuwissen et al., 2001) with GBLUP-type models, although the models can be applied in other research areas.

lme4GS uses the computational engine provided by the well-tested and widely used lme4 package to fit mixed models with a variance–covariance matrix provided by the user. lme4GS can be considered a generalization of existing package rrBLUP (Endelman, 2011) because it is able to fit model (4), whereas, rrBLUP is able to fit model (1). The package also implements some of the models in the sommer package of Covarrubias-Pazaran (2016). lme4GS uses the high-level modular structure of lmer (formula module, objective function module, optimization module, and output module) to fit the models with variance–covariance matrices provided by the user. The formula module allows the specification of fixed and random effects and associated variance–covariance matrices, from which matrices for fixed and random effects ($\mathbf{X}$, $\mathbf{Z}_1, ..., \mathbf{Z}_q$, respectively) are obtained. After that, the variance–covariance matrices are introduced by computing transformed incidence matrices ($\mathbf{Z}_j^*, j = 1, ..., q$) using the Cholesky or eigenvalue decomposition of variance–covariance matrices provided by the user, which are taken as inputs to define the objective function (deviance function). Once the objective function has been defined, the optimization module is used to optimize the objective function and provide REML estimates of the parameters of interest. Finally, the output module is used to provide an output that can be interpreted by the end user. We developed three main R functions:

- **lmerUvcov:** Fits a linear mixed model with a variance–covariance matrix provided by the user. This function takes as input a formula to specify the response **y**, the fixed effects (fixed) and the random effects (random), a data.frame, and a list (Uvcov) to specify the variance–covariance matrix for random effects. Once the model is fitted, the routine returns an object of class merMod for which many methods are available in R for further processing (e.g., summary, print, predict, and VarCorr).
- **ranefUvcov:** Extracts the conditional means of random effects. This function takes as input an object returned by the lmerUvcov function. If the ranef function in the lme4 package is used taking as input the object provided by the lmerUvcov function, it will extract the conditional means for the random effects in model (6); the conditional means for random effects in model (5) are obtained as explained in the BLUPs section. The ranef function in lme4 is overwritten with ranefUvcov, so the user can call either of these two routines and obtain the same results.
- **ranefUvcovNew:** Obtains BLUPs for new levels of random effects with user-specified variance–covariance matrices. The function takes as input an object provided by the lmerUvcov function and a two-level list with variance–covariance matrices that contains information of the genotype identifiers (GIDs) to be predicted and those that were included when fitting the model. The BLUPs are obtained using partitions similar to those used to derive equation (4).

The software is available in the github repository[1].

BOX 1 | Loading wheat data.

```
library(lme4GS)
library(pedigreemm)
data(wheat599)
ls()   #list objects
```

EXAMPLES

In this section, we illustrate the use of the R-package lme4GS with several examples using sample data included in the package. In our examples, we consider only the prediction of random effects and the estimation of variance parameters, although the package is also able to estimate fixed effects.

Example 1: Genome-Wide Prediction Using Markers and Pedigree

In this example, we analyze a set of 599 wheat lines developed by the CIMMYT Global Wheat Breeding Program. The dataset has

[1]https://github.com/perpdgo/lme4GS

BOX 2 | Computing A and G matrices.

```
## Complete and sort incomplete Pedigree using editPed
PedEdit< editPed(sire = wheat.Pedigree$gpid1, dam=wheat.Pedigree$gpid2,
                 label = wheat.Pedigree$progenie,
                verbose = TRUE)

## Converted the data frame PedEdit into an S4 object of formal
## class 'Pedigree'
PedFinal<-with(PedEdit,pedigree(label=label, sire=sire,dam=dam))

#A
AFull<-getA(PedFinal)
GID<-unique(wheat.Pheno$GID)
selected<-rownames(AFull)%in%GID
A<-AFull[selected,selected]
A<-matrix(A,599,599)
rownames(A)<-colnames(A)<-rownames(AFull[selected,selected])

W<-scale(wheat.X,center=TRUE,scale=TRUE)
G<-tcrossprod(W)/ncol(W)

#Environment 1
e1<-which(wheat.Pheno$Env==1)
y<-wheat.Pheno[e1,]$Yield
GID<-as.character(wheat.Pheno[e1,]$GID)

wheat<-data.frame(y = y,mrk=GID,ped=GID)
random<-list(mrk=list(K=G),ped=list(K=A))
fmGA<-lmerUvcov(y~(1| mrk)+(1| ped),data = wheat,Uvcov = random)
summary(fmGA)

#BLUPs
ranefUvcov(fmGA)

#or equivalently
ranef(fmGA)
```

been analyzed several times in the literature (e.g., de los Campos et al., 2009; Crossa et al., 2010; Pérez et al., 2010). The dataset includes grain yield information, a pedigree, and 1,477 markers generated by Triticarte Pty., Ltd. (Canberra, Australia[2]). Here, we present the raw phenotypic data, including the replicates in each environment and the pedigree information, in order to show how to use R tools to obtain the additive relationship matrix that is later used as input for fitting the models. The dataset is loaded into the R environment with the commands shown in **Box 1**.

Once the commands are executed, the following objects are available:

- wheat.Pheno: A data.frame with four columns: Env for environments, Rep for replicates, GID for genotype identifiers, and Yield for grain yield.
- wheat.Pedigree: A data.frame with three columns: gpid1 and gpid2, which correspond to the GID of parents 1 and 2, respectively, and progeny, which correspond to the GIDs of progeny.
- wheat.X: A matrix of dimensions 599 × 1,279, which corresponds to Diversity Array Technology (DArT) markers coded as 0 and 1.

A linear model to predict grain yield in one of the environments using markers and pedigree is given by:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}, \quad (6)$$

where, $\mathbf{y}$ is the response vector in one environment, $\mathbf{1}$ is a vector of ones, μ is an intercept, $\mathbf{u}_1 \sim MN(\mathbf{0}, \sigma^2_m\mathbf{G})$, $\mathbf{G} = \mathbf{WW}'/p$ (see Lopez-Cruz et al., 2015) is a genomic relationship matrix, $\mathbf{W}$ is the matrix of markers centered and standardized, p is the number of markers, σ^2_m is a variance parameter associated with markers, $\mathbf{u_2} \sim MN(\mathbf{0}, \sigma^2_a\mathbf{A})$, $\mathbf{A}$ is an additive relationship matrix derived from pedigree, σ^2_a is its associated variance parameter, $\mathbf{Z}_1$, $\mathbf{Z}_2$ are matrices that connect phenotypes with genotypes, and $\mathbf{e}$ is a random term distributed as in model (1). The additive relationship matrix $\mathbf{A}$ can be easily computed in R using the pedigreemm package (Vazquez et al., 2010); the corresponding Cholesky decomposition can be computed very efficiently, and the package is able to store the result as a sparse matrix. The code in **Box 2** computes the $\mathbf{A}$ and $\mathbf{G}$ matrices and then fits the mixed model using the lmerUvcov function. After that, it extracts the BLUPs using the ranefUvcov function.

The model fitting time is about 81 s on a computer with a 2.8-GHz Intel Core i7 processor. After the model is fitted, the summary function can be used to show some of the results. The estimates of variance parameters are $\widehat{\sigma}^2_m = 0.2218$,

[2]https://www.diversityarrays.com

BOX 3 | Partial output from **Box 2**.

```
Linear mixed model fit by REML ['lmerUvcov']
Formula: y ~ (1 | mrk) + (1 | ped)
   Data: wheat

REML criterion at convergence: 1103.5

Scaled residuals:
      Min       1Q       Median       3Q       Max
 -2.51852  -0.44430    0.00982  0.42390   2.60030

Random effects:
   Groups   Name             Variance    Std.Dev.
   mrk      (Intercept)      0.22189     0.4711
   ped      (Intercept)      0.21138     0.4598
   Residual                  0.03496     0.1870
Number of obs: 1198, groups: mrk, 599; ped, 599

Fixed effects:
               Estimate Std. Error t value
(Intercept)    4.81719   0.08757     55.01
```

BOX 4A | Single training and testing partition.

```
set.seed(456)
trn<-sample(unique(GID),size=as.integer(0.80*599))
tst<-setdiff(unique(GID),trn)

#Phenotypes in training and testing
y_trn<-y[GID%in%trn]
y_tst<-y[GID%in%tst]

A_trn<-A[rownames(A)%in%trn,colnames(A)%in%trn]
G_trn<-G[rownames(G)%in%trn,colnames(G)%in%trn]
GID_trn<-GID[GID%in%trn]
GID_tst<-GID[!(GID%in%trn)]

pheno_trn<-data.frame(y_trn=y_trn,mrk=GID_trn,
                      ped = GID_trn)

random<-list(mrk=list(K=G_trn),ped=list(K=A_trn))

fmGA_trn<-lmerUvcov(y_trn~(1| mrk)+(1| ped),
data=pheno_trn,
                    Uvcov = random)

plot(pheno_trn$y_trn, predict(fmGA_trn),
     xlab="Observed phenotype",ylab="Predicted
     phenotype")

#Predict for new levels
blup_tst<-ranefUvcovNew(fmGA_trn,
                        Uvcov=list(mrk=list(K=G),
                        ped=list(K=A)))
i1<-match(GID_tst,rownames(blup_tst$mrk))
i2<-match(GID_tst,rownames(blup_tst$ped))
blup_mrk<-blup_tst$mrk[i1,1]
blup_ped<-blup_tst$ped[i2,1]
yHat_tst<-fixef(fmGA_trn)[1] + blup_mrk + blup_ped

points(y_tst,yHat_tst,col="red",pch=19)
legend("topleft",legend=c("Training","Testing"),
       pch=c(1,19),col=c("black","red"),bty="n")

#Correlation in testing set
cor(y_tst,yHat_tst)

#MSE
var(y_tst-yHat_tst)

#Data frame with prediction for further processing
predictions<-data.frame(GID=GID_tst,y=y_tst,
yHat=yHat_tst)
```

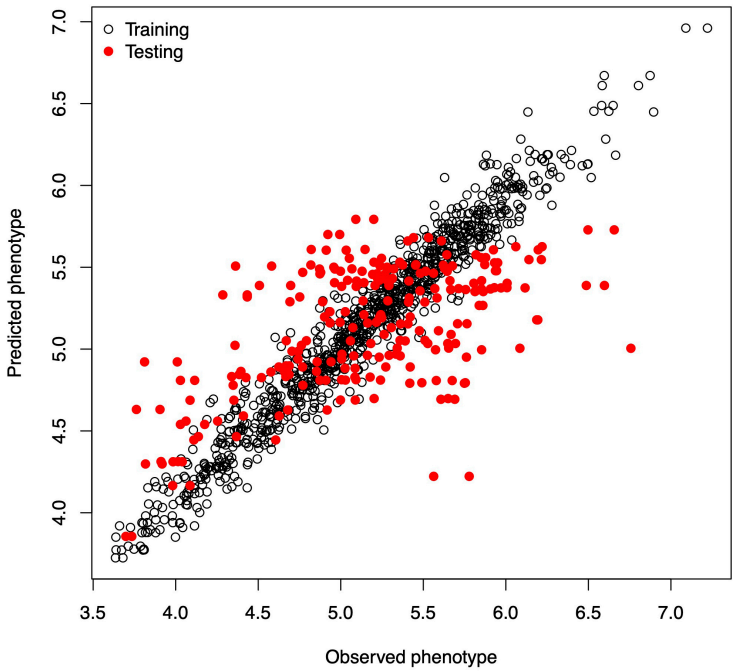

FIGURE 1 | Observed vs. predicted phenotypic values in the training and testing sets.

$\widehat{\sigma}_a^2 = 0.2113$, and $\widehat{\sigma}_e^2 = 0.0349$ (see **Box 3**). The functions predict, residuals, etc., that are routinely used after fitting the model with the lmer function; they can also be used with the resulting object.

Example 2: Training and Testing Sets

In this example, we mimic the GS problem faced by breeders; we evaluate the predictive ability of model (6) by CV, which requires randomly partitioning the data into two disjoint sets, assigning 80% of the lines to the training set and the remaining 20% to the testing set. The code in **Box 4A** partitions the data into the training and testing sets and defines two vectors, y_trn and y_tst, with the phenotypic values of both sets. Next, it creates a list object with the random effects for the linear mixed model. The linear mixed model is fitted using the training set of the data, with the lmerUvcov function. In the next step, we define a list of random effects including the variance–covariance matrices **G** and **A** and the GIDs of the lines to be predicted; the row and column names of the covariance matrices correspond to the GIDs. The ranefUvcovNew function is used for prediction and provides a list of BLUPs for each of the random terms as a result. Finally, the predictions for individuals in the testing set are obtained by simply adding up the intercepts to the BLUPs. Observed and predicted values are stored in a data.frame with three columns: GID, y (observed phenotypic values), and yHat (predicted phenotypic values) used for graphical displays. **Figure 1** shows a scatter plot with observed and predicted phenotypic values in both the training and testing sets. Pearson's correlation coefficient between the observed and predicted values is 0.5638, and the mean squared error (MSE) is 0.2581.

Box 4B shows the R code to perform a five-fold CV that is widely used to study prediction accuracy (e.g., Crossa et al., 2010). We randomly divided the data into five disjoint sets based on the GID, $\{S_1, ..., S_5\}$. Each set is used to measure prediction accuracy. With the use of these sets, the data are divided into the training and testing populations; for example, the data in $\{S_2, ..., S_5\}$ are the training data, and S_1 are the testing data. The model is fitted using the training data, then phenotypes for S_1 are predicted, and prediction accuracy is measured. The same exercise can be carried out taking S_f as the testing data, $f = 2, ..., 5$. **Table 1** shows the results of CV, column 1 corresponds to fold, column 2 shows Pearson's correlation coefficient between observed and predicted values for individuals in the training set, column 3 corresponds to the MSE in the training set, and columns 4 and 5 show the correlations and MSE for individuals in the testing set. The average correlation in the training set is 0.9768, whereas, the correlation in the testing set is 0.5192. The average MSE in the training set is 0.0187, and that in the testing set is 0.2897. The results are as expected: the correlation in the training set is higher than in the testing set, and the MSE is higher in the testing set than in the training set.

Example 3: Hybrid Prediction

The prediction of hybrid performance is very important in agricultural breeding programs. Technow et al. (2014) and Acosta-Pech et al. (2017) employed G-BLUP type models to predict the performance of maize hybrids. The linear model used to that end is given by:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{W}\theta + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{Z}_3\mathbf{u}_3 + \mathbf{e}, \qquad (7)$$

BOX 4B | Cross-validation.

```
set.seed(789)
uGID<-unique(GID)
nFolds<-5
sets<-sample(1:nFolds,size=length(uGID),replace=TRUE)
resultsCV<-matrix(NA,nrow=nFolds,ncol=4)
colnames(resultsCV)=c("r_trn","MSE_trn",
  "r_tst","MSE_tst")

for(f in 1:nFolds)
{
          #Training and testing
          trn<-(uGID[sets!=f])
          tst<-(uGID[sets==f])

          #Phenotypes in training and testing
          y_trn<-y[GID%in%trn]
          y_tst<-y[GID%in%tst]

          A_trn<-A[rownames(A)%in%trn,
          colnames(A)%in%trn]
          G_trn<-G[rownames(G)%in%trn,
          colnames(G)%in%trn]
          GID_trn<-GID[GID%in%trn]
          GID_tst<-GID[!(GID%in%trn)]

          pheno_trn<-data.frame(y_trn=y_trn,
          mrk=GID_trn,
                           ped=GID_trn)

          random<-list(mrk=list(K=G_trn),
          ped=list(K=A_trn))

          fmGA_trn<-lmerUvcov(y_trn~(1| mrk)+(1|
          ped), data=pheno_trn,
                       Uvcov=random)

          yHat_trn<-predict(fmGA_trn)

          #Correlation in training set
          resultsCV[f,1]<-cor(y_trn,yHat_trn)

          #MSE in training set
          resultsCV[f,2]<-var(y_trn-yHat_trn)


          #Predict for new levels
          blup_tst<-ranefUvcovNew(fmGA_trn,
          Uvcov=list(mrk=list(K=G),
          ped=list(K=A)))
          i1<-match(GID_tst,rownames(blup_tst$mrk))
          i2<-match(GID_tst,rownames(blup_tst$ped))
          blup_mrk<-blup_tst$mrk[i1,1]
          blup_ped<-blup_tst$ped[i2,1]
          yHat_tst<-fixef(fmGA_trn)[1] + blup_mrk +
          blup_ped
           #Correlation in testing set
          resultsCV[f,3]<-cor(y_tst,yHat_tst)
          #MSE
          resultsCV[f,4]<-var(y_tst-yHat_tst)
}
resultsCV
```

TABLE 1 | Results from five-fold cross-validation.

Fold	Training		Testing	
	r	MSE	*r*	MSE
1	0.9752	0.0201	0.5290	0.2778
2	0.9775	0.0181	0.5680	0.2729
3	0.9755	0.0197	0.5096	0.3035
4	0.9786	0.0173	0.4179	0.3280
5	0.9775	0.0182	0.5714	0.2663
avg	0.9769	0.0187	0.5192	0.2897
sd	0.0015	0.0012	0.0624	0.0256

MSE, mean squared error.

BOX 5 | Loading maize data.

```
library(lme4GS)
data(cornHybrids)
ls()  #List objects
```

BOX 6 | Fitting model for hybrid prediction.

```
 maize.Pheno$GCA1<-as.character(maize.Pheno$GCA1)
 maize.Pheno$GCA2<-as.character(maize.Pheno$GCA2)
 maize.Pheno$SCA<-as.character(maize.Pheno$SCA)

 #Genomic relationship matrix for parent 1
 GCA1<-unique(maize.Pheno$GCA1)
 selected<-rownames(maize.G)%in%GCA1
 K1<-maize.G[selected,selected]

#Genomic relationship matrix for parent 2
GCA2<-unique(maize.Pheno$GCA2)
selected<-rownames(maize.G)%in%GCA2
K2<-maize.G[selected,selected]

#kronecker, make.dimmanes is necessary to identify
the hybrids
#with the label Parent 1:Parent 2
K3<-kronecker(K1,K2,make.dimnames=TRUE)

#Training set
trn<-which(!is.na(maize.Pheno$PlantHeight))

hybrid<-data.frame(y=maize.Pheno$PlantHeight[trn],
                   loc=maize.Pheno$Location[trn],
                   P1=maize.Pheno$GCA1[trn],
                   P2=maize.Pheno$GCA2[trn],
                   H=maize.Pheno$SCA[trn])

random<-list(P1=list(K=K1),
             P2=list(K=K2),
             H=list(K=K3))

#Fit the model
fm<-lmerUvcov(y~loc+(1|P1)+(1|P2)+(1|H),
data=hybrid, Uvcov=random)

summary(fm)
```

where, $\mathbf{y}$ is the response vector; $\mathbf{1}$ is a vector of ones; μ is an intercept; $\mathbf{W}$ is the design matrix for environments; θ is the vector of environmental effects (fixed); $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ are incidence matrices for paternal, maternal, and hybrids, respectively; $\mathbf{u}_1$ and $\mathbf{u}_2$ are vectors of general combining abilities for parental and maternal lines, respectively; $\mathbf{u}_1 \sim MN(\mathbf{0}, \sigma_1^2\mathbf{K}_1)$, $\mathbf{u}_2 \sim MN(\mathbf{0}, \sigma_2^2\mathbf{K}_2)$ with $\mathbf{K}_1$ and $\mathbf{K}_2$ relationship matrices for

BOX 7 | Output from **Box 6**.

```
#...
Random effects:
 Groups  Name           Variance     Std.Dev.
 H       (Intercept)     0.016385     0.12800
 P2      (Intercept)     0.000841     0.02900
 P1      (Intercept)     0.002047     0.04525
 Residual                0.001182     0.03438
Number of obs: 400, groups: H, 100; P2, 20; P1, 20
#...
```

BOX 8 | Predicting hybrid's performance.

```
#Unobserved hybrid performance
blup_tst<-ranefUvcovNew(fm,Uvcov=list(H=list(K=K3)))
blup_tst$H

#variance parameters
vc<-VarCorr(fm)
print(vc,comp=c("Variance","Std.Dev."),digits=4)
variances<-as.data.frame(vc)$vcov
variances

#Heritability
h2<-sum(variances[2:3])/sum(variances[2:4])
h2
```

paternal and maternal lines σ_1^2, σ_2^2 associated variance parameters, $\mathbf{u}_3 \sim MN(\mathbf{0}, \sigma_3^2\mathbf{K}_3)$, with $\mathbf{K}_3 = \mathbf{K}_1 \otimes \mathbf{K}_2$, σ_3^2 variance parameter associated with hybrids, and $\mathbf{e} \sim MN(\mathbf{0}, \sigma_e^2\mathbf{I})$. Note that model (7) can be rewritten as $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{Z}_3\mathbf{u}_3 + \mathbf{e}$, where, $\mathbf{X} = [\mathbf{1W}]$ and $\beta = (\mu, \boldsymbol{\theta}')'$, which corresponds to model (4) discussed before. To exemplify how to fit this model in the lme4GS package, we used the DT_cornHybrids dataset included in the R-package sommer (Covarrubias-Pazaran, 2016), and we included a copy of the original data in the package (cornHybrids). The dataset contains phenotypic data for grain yield and plant height for 100 out of 400 possible crosses that originated from 40 inbred lines belonging to two heterotic groups, with 20 lines in each. Only 100 hybrids were evaluated in four locations, and then the problem was to estimate their general combining abilities and specific combining abilities and to predict the performance of untested hybrids at each location. The dataset can be loaded in R using the commands shown in **Box 5**:

The dataset contains the following R objects:

- maize.Pheno: A data.frame with six columns: Location, GCA1 (Parent 1), GCA2 (Parent 2), SCA (hybrid), Yield, and PlantHeight. Records with missing values in the last two columns correspond to hybrids (identified with the Parent 1:Parent 2 label) that were not evaluated in the field and that we need to predict.
- maize.G: A matrix with relationships between individuals for parents of both heterotic groups ($\mathbf{K}_1$ and $\mathbf{K}_2$). The matrix was computed using 511 single-nucleotide polymorphisms (SNPs) using the A.mat function included in the rrBLUP package (Endelman, 2011). The row names and column names of this matrix correspond to the GIDs for Parent 1 and Parent 2.

BOX 9 | Gaussian and exponential kernel.

```
#Box 9: Gaussian and exponential kernel
library(lme4GS)
library(pedigreemm)

#Load data
data(wheat599)

## Complete and sort incomplete Pedigree using editPed
PedEdit<-editPed(sire=wheat.Pedigree$gpid1, dam=wheat.Pedigree$gpid2,
                 label=wheat.Pedigree$progenie,
                 verbose=TRUE)

## Converted the data frame PedEdit into an S4 object of formal
## class 'Pedigree'
PedFinal<-with(PedEdit,pedigree(label=label, sire=sire,dam=dam))

#A
AFull<-getA(PedFinal)
GID<-unique(wheat.Pheno$GID)
selected<-rownames(AFull)%in%GID
A<-AFull[selected,selected]
A<-matrix(A,599,599)
rownames(A)<-colnames(A)<-rownames(AFull[selected,selected])

#X (markers)
X<-scale(wheat.X,center=TRUE,scale=TRUE)

#Phenotypes environment 1
e1<-which(wheat.Pheno$Env==1)
y<-wheat.Pheno[e1,]$Yield
GID<-as.character(wheat.Pheno[e1,]$GID)

wheat<- data.frame(y=y, ped=GID,k_id=GID)

fm1<-theta_optim(y~(1| k_id)+(1| ped), Uvcov=list(ped=list(K=A)),
                 kernel=list(kernel_type="gaussian",MRK=X),
                 data=wheat)

fm2<-theta_optim(y~(1| k_id)+(1| ped), Uvcov=list(ped=list(K=A)),
            kernel=list(kernel_type="exponential",MRK=X),
            data=wheat)

par(mfrow=c(1,2))
plot(fm1$theta,fm1$LL,xlab=expression(theta), ylab="Log-Likelihood",
     type="b",pch=19,main="a)")
plot(fm2$theta,fm2$LL,xlab=expression(theta), ylab="Log-Likelihood",
     type="b",pch=19,main="b)")
```

The code in **Box 6** computes matrices $\mathbf{K}_1$, $\mathbf{K}_2$, and $\mathbf{K}_3$ in model (7) and fits the model using the lmerUvcov function using only observed phenotypic values for plant height.

The model fitting takes about 1 s to complete on a computer with a 2.8-GHz Intel Core i7 processor. Once the model is

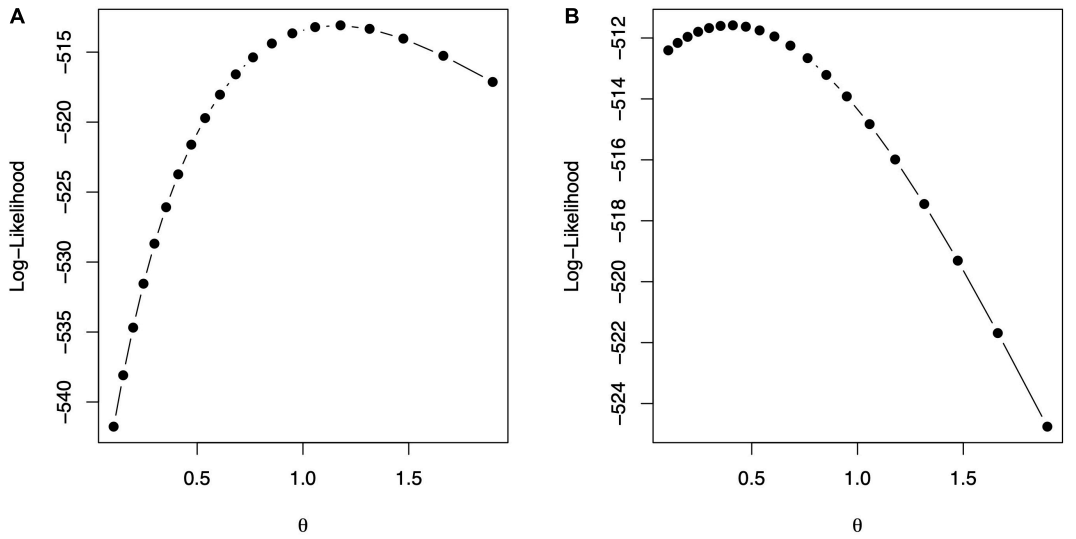

FIGURE 2 | The values of the bandwidth parameter vs. the log-likelihood. **(A)** Gaussian kernel. **(B)** Exponential kernel.

BOX 10 | Summary of fitted models.

```
summary(fm1$fm)
#Output (edited)
Random effects:
 Groups   Name          Variance    Std.Dev.
 k_id     (Intercept)    0.29043    0.5389
 ped      (Intercept)    0.07751    0.2784
 Residual                0.03434    0.1853
Number of obs: 1198, groups: k_id, 599; ped, 599

Fixed effects:
             Estimate Std. Error t value
(Intercept)   4.6510    0.1314     35.4

summary(fm2$fm)
#Output (edited)
Random effects:
 Groups  Name          Variance    Std.Dev.
 ped     (Intercept)    0.05154    0.2270
 k_id    (Intercept)    0.62952    0.7934
 Residual               0.03408    0.1846
Number of obs: 1198, groups: k_id, 599; ped, 599
Fixed effects:
             Estimate Std. Error t value
(Intercept)   4.5311    0.5599     8.093
```

fitted, the summary function can be used to display some relevant information. The summary output is displayed in **Box 7**, which shows estimates for general combining ability, and specific combining ability and the variance parameter associated with residuals, $\widehat{\sigma}_1^2 = 0.016385, \widehat{\sigma}_2^2 = 0.000841, \widehat{\sigma}_3^2 = 0.002047$, and $\widehat{\sigma}_e^2 = 0.001182$.

The expected hybrid performance of individuals not evaluated in field can be obtained by combining the outputs from the ranefUvcov and ranefUvcovNew functions. **Box 8** shows the instructions to compute the BLUPs for the specific combining ability of hybrids. The ranefUVcov function is called internally in ranefUvcovNew. **Box 8** also shows how to extract variance parameters using the VarCorr function and then compute heritability using the results. Following Covarrubias-Pazaran (2016), $h^2 = (\sigma_1^2 + \sigma_2^2)/(\sigma_1^2 + \sigma_2^2 + \sigma_e^2)$, which leads to an estimated heritability of 0.70.

Example 4: Selection of the Bandwidth Parameter With a Gaussian Kernel

Gianola et al. (2006) introduced the Gaussian kernel into quantitative genetics with the idea of capturing the total genetic effects in the problem of genomic prediction. The Gaussian kernel is defined as (e.g., Morota and Gianola, 2014; Pérez and de los Campos, 2014)

$$\mathbf{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left\{-\theta \frac{d_{ij}^2}{m}\right\} = \exp\left\{-\theta \frac{\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2}{m}\right\} \quad (8)$$

where, θ is a positive bandwidth parameter; d_{ij} is the Euclidean distance; and $\mathbf{x}_{ik}$ $(i, j = 1, ..., n,\ k = 1, ..., m)$ is the marker genotype code for individual i at marker k, and m is the number of markers. The bandwidth parameter may be chosen by CV, REML, or maximum likelihood or with Bayesian methods. The Gaussian kernel has been used by many authors for genomic prediction (e.g., de los Campos et al., 2010; Endelman, 2011; Pérez-Elizalde et al., 2015). The selection of the bandwidth is not an easy problem due to high computational cost; de los Campos et al. (2010) and Endelman (2011) proposed evaluating the performance of the model, which includes the Gaussian kernel over a grid of values of θ. Given that $\theta > 0$, if we set $\rho = \exp(-\theta)$, then $\rho \in (0, 1)$, so we can define a grid of values for ρ and then, using these values, set the values for θ, that is, $\theta = -\log\rho$, so that equation (8) can be rewritten as $\mathbf{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left\{\log \rho d_{ij}^2/m\right\}$. Another kernel that is also used in

TABLE 2 | Time comparison (seconds) among different software for models fitted in the work.

Software	Version	Examples			
		Model (6)	Model (7)	Model (10) Gaussian	Model (10) exponential
lme4GS	0.1	81.5	1.3	1,608.8	1,701.1
BGLR 0.8	0.8	143.0	20.2	–	–
sommer 4.1.3	4.1.3	46.0	2.7	–	–

genomic prediction is the exponential kernel (e.g., Piepho, 2009; Endelman, 2011):

$$\mathbf{K}\left(\mathbf{x}_i,\mathbf{x}_j\right) \;=\; \exp\left\{-\theta d_{ij}/\sqrt{m}\right\}, \tag{9}$$

where, all the terms have been described previously. Similar to the case of the Gaussian kernel, the model can be reparametrized in terms of parameter $\rho \in (0, 1)$.

We developed the function theta_optim that fits model (5) when one of the random terms ($\mathbf{u}_j, j = 1, ..., q$) includes as the variance–covariance matrix a Gaussian or exponential kernel. This function takes as input the same objects as the lmerUvcov function and a list (kernel) containing (i) a matrix with distances ($\{d_{ij}/\sqrt{m}\}, i, j = 1, ..., n$) or the marker matrix ($\left\{x_{ij}\right\}, i = 1, ..., n, j = 1, ..., m$), (ii) the kernel type (either "gaussian" or "exponential"), and (iii) a sequence of values for θ; the IDs for the individuals are taken directly from the row names of matrices that provide the distances or the markers. If the sequence of values for θ is not provided, then it is generated automatically. The software then fits the mixed model in (5) using the lmerUvcov function for each of the distinct values of θ. The value of θ that maximizes the log-likelihood is chosen as the optimum. The function returns a list with the following elements: a vector of values of the log-likelihood, the maximum value of the log-likelihood, the values of θ used for fitting the model, the optimum value of θ, the fitted model, and the kernel computed with the optimum value of θ .

In the following example, we show how to predict grain yield using a relationship matrix derived from a Gaussian or exponential kernel and a relationship matrix derived from a pedigree. A linear model to predict grain yield for environment one is analogous to model (1):

$$\mathbf{y} \;=\; \mathbf{1}\mu + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}, \tag{10}$$

where, $\mathbf{y}$ is the grain yield; $\mathbf{1}$ is a vector of ones; μ is an intercept, $\mathbf{u}_1 \sim MN(\mathbf{0}, \sigma^2_m\mathbf{K})$, with $\mathbf{K}$ a kernel, which can be either Gaussian or exponential, and σ^2_m is a variance parameter associated with markers; $\mathbf{u_2} \sim MN(\mathbf{0}, \sigma^2_a\mathbf{A})$, where, $\mathbf{A}$ is an additive relationship matrix derived from pedigree, and σ^2_a its associated variance parameter; $\mathbf{Z}_1$, $\mathbf{Z}_2$ are matrices that connect phenotypes with genotypes; and $\mathbf{e}$ is a random term distributed as in model (1).

The code in **Box 9** is used to fit model (10) for Gaussian and exponential kernels. **Figure 2** shows the profile of the log-likelihood for different θ values. For the Gaussian kernel, the maximum of the log-likelihood is equal to -513.0865, attained at $\widehat{\theta} = 1.1779$, whereas, for the exponential kernel, the maximum of the log-likelihood is equal to -511.585, attained at $\widehat{\theta} = 0.4107$. The code in **Box 10** shows how to summarize parameter estimates for the fitted model with the optimum value of the bandwidths from, where, estimates of the variance parameters can be obtained. The model fitting time is about 1,608 s for the model with Gaussian kernel and 1,701 s for the model with exponential kernel using the same processor described before. Note that the selection of bandwidth parameter is a very computer intensive task, but several authors (e.g., Endelman, 2011; Pérez-Elizalde et al., 2015) have reported that the prediction accuracy with nonadditive kernels is higher than the prediction accuracy of ridge regression (or equivalently GBLUP).

Computational Times and Comparison With Other Software

We fitted models (6) and (7) in sommer (Covarrubias-Pazaran, 2016) and BGLR (Pérez and de los Campos, 2014). In the case of BGLR, the number of iterations for the Gibbs sampler was set to 30,000. We were unable to fit the models in rrBLUP (Endelman, 2011) because it is not possible to include more than one covariance matrix in the software; that is also the reason that we were unable to fit model (10) with this software. The predictions from the different software programs were about the same. Here, we present a small comparison of running times for model (6) fitted in **Box 2**, model (7) fitted in **Box 6**, and model (10) fitted in **Box 9** with Gaussian and exponential kernels. Covarrubias-Pazaran (2016) also included a benchmark of sommer against other packages. Models were fitted using a 2.8-GHz Intel Core i7 processor in R-4.0.5 (R Core Team, 2021). **Table 2** presents the resulting time (in seconds) it takes to fit the different models. Some entries in the **Table 2** are empty because the corresponding models cannot be fitted in the corresponding software package. From this **Table 2**, we conclude that sommer is the fastest software, followed by lme4GS and BGLR.

CONCLUSION

We developed an R software package that can be used to fit mixed models with user-defined covariance structures for random effects. The software was developed with applications of GS in mind, mainly for applications in plant breeding with small to moderately sized datasets. However, given the omnipresence of mixed models, the package can be used in other research areas. The software fits the model using well-known and widely tested computational routines available in the lme4 package. The software provides a user-friendly and intuitive interface that allows users to fit a wide variety of classic linear mixed models.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/perpdgo/lme4GS.

AUTHOR CONTRIBUTIONS

DC-P, PP-R, and JC conceived the work and wrote the manuscript. DC-P and PP-R wrote the software. CV-C, SP-E, and MV-P assisted in drafting the manuscript, discussed the analysis, and provided useful comments. All authors contributed to the article and approved the submitted version.

FUNDING

Open-access fees were received from the Bill and Melinda Gates Foundation. We acknowledge the financial support provided by the Bill and Melinda Gates Foundation [INV-003439 BMGF/FCDO Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods (AG2MW)] as well as USAID projects [Amend. No. 9 MTO 069033, USAID-CIMMYT Wheat/AGGMW, AGG-Maize Supplementary Project, AGG (Stress Tolerant Maize for Africa)]. We are also thankful for the financial support provided by the Foundations for Research Levy on Agricultural Products (FFJ) and the Agricultural Agreement Research Fund (JA) in Norway through NFR grant 267806 as well as the CIMMYT CRP (wheat and maize).

ACKNOWLEDGMENTS

We thank all scientists, field workers, and lab assistants from the National Programs and CIMMYT who collected the data used in this study.

REFERENCES

Acosta-Pech, R., Crossa, J., de los Campos, G., Teyssèdre, S., Claustres, B., Pérez-Elizalde, S., et al. (2017). Genomic models with genotype x environment interaction for predicting hybrid performance: an application in maize hybrids. *Theoretical and Applied Genetics* 130, 1431–1440. doi: 10.1007/s00122-017-2898-0

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48. doi: 10.18637/jss.v067.i01

Bernardo, R., and Yu, J. (2007). Prospects for genome wide selection for quantitative traits in maize. *Crop Science* 47, 1082–1090. doi: 10.2135/cropsci2006.11.0690

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PloS one* 11:6. doi: 10.1371/journal.pone.0156744

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in plant science* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

de los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research* 92, 295–308. doi: 10.1017/S0016672310000285

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., et al. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385. doi: 10.1534/genetics.109.101501

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776. doi: 10.1534/genetics.105.049510

Gilmour, A., Cullis, B., Welham, S., Gogel, B., and Thompson, R. (2004). An efficient computing strategy for prediction in mixed linear models. *Computational statistics & data analysis* 44, 571–586. doi: 10.1016/S0167-9473(02)00258-X

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science* 92, 433–443. doi: 10.3168/jds.2008-1646

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and applied genetics* 127, 595–607. doi: 10.1007/s00122-013-2243-1

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J. L., et al. (2015). Increased prediction accuracy in wheat breeding trials using a marker× environment interaction genomic selection model. *G3: Genes, Genomes, Genetics* 5, 569–582. doi: 10.1534/g3.114.016097

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

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Pérez-Hernández, O., Eskridge, K. M., et al. (2016). A genomic Bayesian multi-trait and multi-environment model. *G3: Genes, Genomes, Genetics* 6, 2725–2744. doi: 10.1534/g3.116.032359

Morota, G., and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in genetics* 5:363. doi: 10.3389/fgene.2014.00363

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

Pérez, P., de los Campos, G., Crossa, J., and Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *The plant genome* 3, 106–116. doi: 10.3835/plantgenome2010.04.0005

Pérez-Elizalde, S., Cuevas, J., Pérez-Rodríguez, P., and Crossa, J. (2015). Selection of the bandwidth parameter in a Bayesian kernel regression model for genomic-enabled prediction. *Journal of agricultural, biological, and environmental statistics* 20, 512–532. doi: 10.1007/s13253-015-0229-y

Piepho, H. P. (2009). Ridge regression and extensions for genomewide selection in maize. *Crop Sci.* 49, 1165–1176. doi: 10.2135/cropsci2008.10.0595

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Technow, F., Schrag, T. A., Schipprack, W., Bauer, E., Simianer, H., and Melchinger, A. E. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* 197, 1343–1355. doi: 10.1534/genetics.114.165860

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., et al. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of dairy science* 92, 16–24. doi: 10.3168/jds.2008-1514

Vazquez, A. I., Bates, D. M., Rosa, G. J. M., Gianola, D., and Weigel, K. A. (2010). an R package for fitting generalized linear mixed models in animal breeding. *Journal of animal science* 88, 497–504. doi: 10.2527/jas.2009-1952

Wimmer, V., Albrecht, T., Auinger, H. J., and Schön, C. C. (2012). synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* 28, 2086–2087. doi: 10.1093/bioinformatics/bts335

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* 44, 821–824. doi: 10.1038/ng.2310

Ziyatdinov, A., Vázquez-Santiago, M., Brunel, H., Martinez-Perez, A., Aschard, H., and Soria, J. M. (2018). lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC bioinformatics* 19:68. doi: 10.1186/s12859-018-2057-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 June 2021
doi: 10.3389/fgene.2021.692870

Genomic Prediction of Yield Traits in Single-Cross Hybrid Rice (*Oryza sativa* L.)

Marlee R. Labroo[1], Jauhar Ali[2], M. Umair Aslam[2], Erik Jon de Asis[2], Madonna A. dela Paz[2], M. Anna Sevilla[2], Alexander E. Lipka[1], Anthony J. Studer[1] and Jessica E. Rutkoski[1]*

[1] Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [2] Rice Breeding Platform, International Rice Research Institute, Los Baños, Philippines

Hybrid rice varieties can outyield the best inbred varieties by 15 – 30% with appropriate management. However, hybrid rice requires more inputs and management than inbred rice to realize a yield advantage in high-yielding environments. The development of stress-tolerant hybrid rice with lowered input requirements could increase hybrid rice yield relative to production costs. We used genomic prediction to evaluate the combining abilities of 564 stress-tolerant lines used to develop Green Super Rice with 13 male sterile lines of the International Rice Research Institute for yield-related traits. We also evaluated the performance of their F_1 hybrids. We identified male sterile lines with good combining ability as well as F_1 hybrids with potential further use in product development. For yield per plant, accuracies of genomic predictions of hybrid genetic values ranged from 0.490 to 0.822 in cross-validation if neither parent or up to both parents were included in the training set, and both general and specific combining abilities were modeled. The accuracy of phenotypic selection for hybrid yield per plant was 0.682. The accuracy of genomic predictions of male GCA for yield per plant was 0.241, while the accuracy of phenotypic selection was 0.562. At the observed accuracies, genomic prediction of hybrid genetic value could allow improved identification of high-performing single crosses. In a reciprocal recurrent genomic selection program with an accelerated breeding cycle, observed male GCA genomic prediction accuracies would lead to similar rates of genetic gain as phenotypic selection. It is likely that prediction accuracies of male GCA could be improved further by targeted expansion of the training set. Additionally, we tested the correlation of parental genetic distance with mid-parent heterosis in the phenotyped hybrids. We found the average mid-parent heterosis for yield per plant to be consistent with existing literature values at 32.0%. In the overall population of study, parental genetic distance was significantly negatively correlated with mid-parent heterosis for yield per plant ($r = -0.131$) and potential yield ($r = -0.092$), but within female families the correlations were non-significant and near zero. As such, positive parental genetic distance was not reliably associated with positive mid-parent heterosis.

Keywords: hybrid rice, genomic selection, general combining ability, breeding, best linear unbiased prediction

OPEN ACCESS

Edited by:
Shiori Yabe, Institute of Crop Science (NARO), Japan

Reviewed by:
Motoyuki Ishimori, The University of Tokyo, Japan
Hiromi Kajiya-Kanegae, National Agriculture and Food Research Organization (NARO), Japan

***Correspondence:**
Jauhar Ali
j.ali@irri.org

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *09 April 2021*
Accepted: *09 June 2021*
Published: *30 June 2021*

Citation:
Labroo MR, Ali J, Aslam MU, de Asis EJ, dela Paz MA, Sevilla MA, Lipka AE, Studer AJ and Rutkoski JE (2021) Genomic Prediction of Yield Traits in Single-Cross Hybrid Rice (Oryza sativa L.). Front. Genet. 12:692870. doi: 10.3389/fgene.2021.692870

INTRODUCTION

Hybrid crop varieties are economically valued for increased vigor, yield, yield stability, and uniformity in species including maize, sugar beet, and cotton (Hochholdinger and Baldauf, 2018). Rice (*Oryza sativa L.*) is a self-pollinated crop that has traditionally been cultivated as an inbred, but the introduction of male sterility into cultivated germplasm in the 1970s enabled hybrid breeding (Virmani and Wan, 1988; Yuan et al., 1989; Nalley et al., 2017). Public hybrid rice varietal development to date has resulted primarily from identification of superior single crosses rather than the systematic breeding of heterotic pools (Lu and Xu, 2010; Spielman et al., 2013). Developing heterotic pools for rice by reciprocal recurrent selection methods may increase the rate of genetic gain for hybrid rice breeding compared to evaluating random crosses, because reciprocal recurrent selection can concurrently improve the additive value of the populations while exploiting heterosis due to dominance (Comstock et al., 1949).

Existing hybrid rice varieties may outyield the best inbred varieties by 10 to 30% with appropriate management (Spielman et al., 2013). However, adoption of hybrid rice varieties is low outside of China, in part because the hybrid yield advantage of temperate *japonica* varieties used in China is much greater than that observed in tropical *indica* varieties to date (Janaiah and Xie, 2010; Longin et al., 2012; Spielman et al., 2013). In some countries, socioeconomic factors such as lack of irrigation systems, paved roads, certified seed suppliers, seed marketing, farmer education, and available credit to purchase seed and fertilizer have limited hybrid rice adoption (Mottaleb et al., 2015; Abebrese et al., 2019). Of the agronomic factors that influence hybrid rice adoption, poor grain quality has been a longstanding challenge, but breeding progress since the early 2000s has produced some acceptable varieties (Spielman et al., 2013). Surveys of farmers suggest that poor quality is not the primary determinant of hybrid rice rejection (Spielman et al., 2013; Feng et al., 2017). Farmers choose not to grow hybrid rice for many reasons, including the high cost of seed, poor seed quality, and lack of hybrid seed availability (Spielman et al., 2013). However, the key agronomic reason for limited adoption is that hybrid rice varieties require more intensive management of irrigation, fertilizer, weeds, and other biotic stressors to provide a yield advantage over inbred varieties in otherwise high-yielding environments (Spielman et al., 2013; Mottaleb et al., 2015; Nalley et al., 2016). Therefore, the development of stress-tolerant hybrids with lowered input requirements could spur hybrid adoption and unlock hybrid yield advantages.

In this study, we evaluated the general combining abilities (GCAs) of the existing male sterile lines of the International Rice Research Institute (IRRI) with stress-tolerant germplasm used in the development of Green Super Rice varieties, as well as the performance, or genetic value, of their F_1 hybrids (Sprague and Tatum, 1942; Ali et al., 2018; Yu et al., 2020). The founders of the Green Super Rice program were selected for multiple stress tolerances, including salinity, submergence, tungro disease, anaerobic germination conditions, and low water and nitrogen inputs (Ali et al., 2018). We sought to identify any outstanding F_1 hybrids—which may be as stress-tolerant and yet higher-yielding than existing Green Super Rice lines—to advance for further testing for varietal release. We also sought to identify male and female lines with superior GCA which could be used to initiate the development of heterotic pools from IRRI germplasm, presumably stacked with alleles conferring stress tolerance. In addition to phenotypic evaluation, we used genomic prediction to evaluate non-phenotyped parental lines and hybrid crosses.

We also tested whether parental genetic distance was correlated with mid-parent heterosis using a large sample of hybrids and genome-wide molecular markers. Mid-parent heterosis due to dominance is expected to be positively correlated with parental squared difference in allele frequency (SDAF) by quantitative genetic theory (Falconer and Mackay, 1996; Amuzu-Aweh et al., 2013). It has also been posited that genetic divergence in founders of heterotic pools may lead to improved gain in reciprocal recurrent selection programs, even though in practice heterotic pools have been developed from closely related germplasm in species such as maize (Melchinger, 1999; Tracy and Chandler, 2006; Rembe et al., 2019). A previous study of rice which used > 100k genome wide markers and six parental lines found a curvilinear relationship of genetic distance and mid-parent heterosis, with mid-parent heterosis increasing with genetic distance to a point and then declining (Waters et al., 2015). Due to past lack of availability of molecular markers, other studies used fewer than 500 markers and found positive correlations of genetic distance and heterosis using 10 or 22 parents (Xiao et al., 1996; Kwon et al., 2002). However, another study using 319 markers found no correlation of genetic distance and heterosis in progeny of 13 parents (Boeven et al., 2020). In other species, such as wheat, whether parental genetic distance is correlated with heterosis varies, with different findings among studies and populations (Melchinger, 1999; Boeven et al., 2020). We wished to test whether parents of hybrids with high SDAF tended to produce hybrids with high mid-parent heterosis in our rice population of study.

MATERIALS AND METHODS

Plant Materials and Population Design

The plant materials for prediction comprised 13 female lines, 564 male lines, and their 10,716 possible F_1 hybrids (**Supplementary File 1**). Twelve of the female lines were wild-abortive cytoplasmic male sterile (CMS), and one female line, A07, was thermosensitive genic male sterile (TGMS). The 564 male tester lines were backcross introgression lines (BILs) from 11 families. Each family of BILs was generated by crossing one of the 11 diverse males to a common female, Weed Tolerant Rice 1 (WTR-1), and advancing the backcrosses to the BC_1F_5 generation under stringent selection for multiple stress tolerances as described in Ali et al. (2018). The recurrent parent, WTR-1, was a restorer line, but the male lines likely segregated for fertility restoration.

Of the 10,716 possible F_1 hybrids, a random subset of 938 were made to comprise the genomic prediction model training set by crossing six female lines to 137 males. To avoid unintentional

selection for synchronous flowering, the female parents had two planting dates. None of the 137 male lines were completely crossed to all six females. However, in pairwise comparisons of females, all females had some overlap with each other female in males crossed (**Supplementary Table 1**). In total, 85 of the males were crossed to a single female, 108 of the males were crossed to 2 females, 124 of the males were crossed to 3 females, 60 of the males were crossed to 4 females, and 5 of the males were crossed to 5 females. All female lines were manually emasculated to prevent contamination by selfing and to expose the stigma.

Two groups of lines were used to estimate mid-parent heterosis and commercial relative performance but were not included for prediction. The five maintainer (B) lines of the five CMS female parents were used to estimate mid-parent heterosis. Six inbred lines were used as commercial checks to estimate commercial relative performance: five of the donor parents, Y 134 (DP 6), Khazar (DP 8), OM 997 (DP 10), M 401 (DP17), and X 21 (DP19), and the recurrent parent, WTR-1.

Field Experimental Design

The F_1 hybrids, their inbred parents, the female maintainer lines, and the commercial checks were phenotyped in an unbalanced randomized complete block design (RCBD) in two environments, irrigated lowland and irrigated upland, at the IRRI farm (approximately 14°09′50.7″ N, 121°15′50.5″ E) in the dry season of 2018. After establishment in seedbeds on January 8, 2018, seedlings were transplanted at the three-leaf stage. Transplanting occurred on January 31, 2018, at the irrigated lowland site and on February 8, 2018, at the irrigated upland site. The plants were harvested the week of May 14, 2018. Basal NPK fertilizer was applied at a rate of 30 kg/ha, and zinc was applied at a rate of 5 kg/ha. N fertilizer was also applied at 28–30 days after transplanting and at the panicle initiation stage (42 days after transplanting) at a rate of 35 kg/ha. Rat fences were installed at both locations; bird pressure was controlled by farmworkers in the lowland environment, and by a bird net at the upland environment. Both environments were hand-weeded. Both environments were irrigated, but the lowland environment was continuously flooded to a depth of ∼10 cm, whereas water depth was allowed to vary in the upland field. Insect pressure was controlled by application of Regent® pesticide (fipronil). Temperatures were sufficient throughout the growing season to ensure seed set in the TGMS female line A07.

The field layout was designed in PBTools 1.4, which depends on the R package agricolae (IRRI, 2014; de Mendiburu, 2020). There were two blocks per environment with one replicate per genotype per block, but replicates were missing for some genotypes. Genotypes were replicated in single-row plots due to limited availability of hybrid seed, with five plants per row, and plants were spaced to 25 × 20 cm within rows. Measurements were only recorded for plants at 20 cm spacing within rows; i.e., edge plants were not measured, nor were plants with missing neighbor plants within the row.

The following traits or trait derivatives were phenotyped: plant height, number of tillers, panicle dry weight, panicle length, proportion of spikelets filled, yield per plant, and yield potential per plant (**Supplementary File 1**). For all genotype replicates, phenotypes were averaged across measured plants in a single-row plot; plants were subsamples, but were not treated as subsamples in downstream modeling because subsampling variance was not of interest. Plant height was measured from the base of the plant to the panicle tip after flowering. All tillers were assumed to be productive based on observations in a subset of samples. For panicle measurements and yield estimates, three random panicles were harvested from each sampled plant, totaling nine panicles per replicate. Panicle length was measured from the pedicel to the panicle tip and averaged across all panicles in a replicate. For a given replicate, yield per plant was calculated as panicle dry weight times tiller number. For a given replicate, yield potential per plant was calculated as average panicle dry weight divided by proportion of spikelets filled times tiller number. Yield potential per plant and proportion of spikelets filled were only measured in the irrigated lowland environment due to cost.

Molecular Marker Generation and Analysis

Genome-wide molecular markers were generated for the parents with tunable genotyping-by-sequencing® (tGBS) by data2Bio and its subsidiary, Freedom Markers, in Ames, Iowa (Ali et al., 2017; Ott et al., 2017). In general, each individual DNA sample was double-digested with restriction enzymes, then the resulting fragments were ligated to a uniquely barcoded adapter at the 5′ end. At the 3′ end, the fragments were ligated to a universal sequencing adapter. However, in the first subsequent PCR amplification of the library, the complementary primer for the universal sequencing adapter was extended by 1-3 nucleotides; only fragments in which the genomic sequence complemented the extension were amplified. Then, the libraries were amplified with Ion Proton sequencing primers.

The male and female parents were sequenced in separate Ion Proton runs. For the male parents, a total of ten Ion Proton runs were done; the female parents were sequenced in two Ion proton runs as part of a larger set. The Ion Proton sequencing reads were trimmed by the manufacturer to remove adapter sequence and bases with PHRED quality scores less than 15 in the software Lucy (Chou and Holmes, 2001). For the female parent A07, additional RAD-sequencing was done in-house. In brief, DNA was extracted from mature leaf tissue of the A07 parent and digested separately with one of three enzyme combinations: *Ape*KI-*Pst*I, *Hin*P1I-*Pst*I, or *Ape*KI only. Then, each digestion was ligated separately to unique barcoded adapters and subsequently pooled. Fragments were selected for sizes ranging from 200–500 bp, and the libraries were amplified by PCR. Then, the libraries were sequenced for single-read 100 bp reads with an Illumina NovaSeq6000 SP. All reads were aligned to the Nipponbare IRGSP-1.0 v7 reference genome in GSNAP 2017-11-15 (Kawahara et al., 2013; Wu et al., 2016). Then, variants were called in BCFtools 1.7 (SAMtools, 2018).

Variant sites were filtered separately within the male and female parent sets in TASSEL 5.0 (Glaubitz et al., 2014). Within sets, only the 2 most major alleles of the variant were considered, and at most 50% of the individuals were permitted to be heterozygous. In the females, the minimum site count was

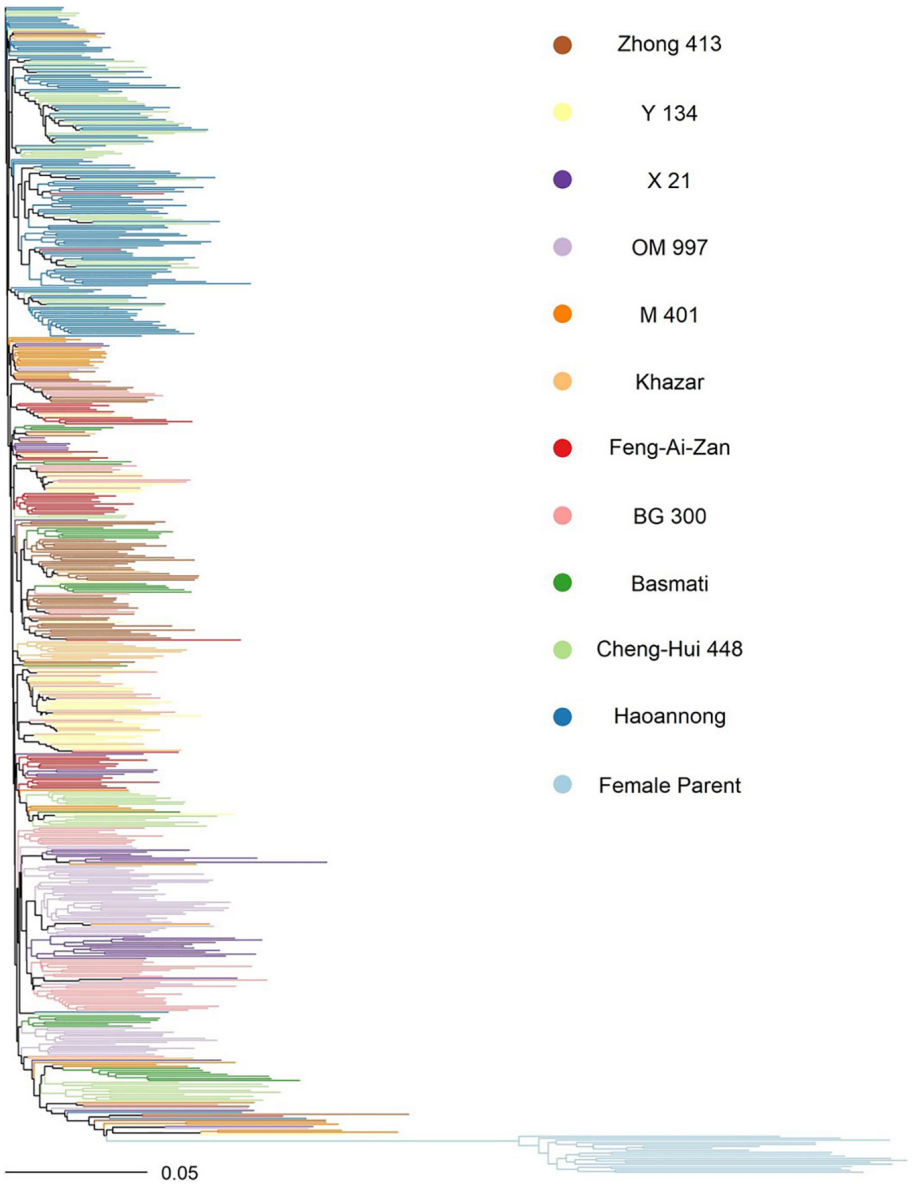

FIGURE 1 | Neighbor-joining tree of the parental lines used in the study. Color indicates the female parents (light blue) and, for the male parents, their subfamily by donor parent (all other colors). The recurrent parent for all males was WTR-1.

2 (corresponding to a minimum site presence of 10% in all females), and the minimum minor allele count was 2, yielding 77,709 polymorphic sites among the females. In the males, the minimum site count was 188 (corresponding to a minimum site presence of 33% in the males), and the minimum minor allele count was 3, yielding 148,922 total polymorphic sites among the males. The genotypes were imputed separately for males and females in Beagle 5.0 (Browning et al., 2018). Principal components analysis of the male and female parent genotypes was done using 20,000 sites common to both using the *glPca* function in the R package adegenet (Jombart and Ahmed, 2011). The F_1 hybrid genotypes were inferred from the same 20,000 sites common to males and females using the *build.HMM* function in the R package sommer version 3.8 (Covarrubias-Pazaran, 2016, 2018). A phylogenetic neighbor-joining tree of the male and female parent genotypes was constructed in TASSEL 5.0 using the same common 20,000 sites, and the tree was visualized in the R package ape version 5.5 (**Figure 1**; Paradis and Schliep, 2019).

The additive relationship matrices of each the females, males, and F_1 hybrids, denoted respectively as $\mathbf{G}_F$, $\mathbf{G}_M$, and $\mathbf{G}_H$, were calculated with the *A.mat* function in sommer by the method of Endelman and Jannink (2012). The 77,709 imputed female sites were used to estimate $\mathbf{G}_F$, and the 148,922 imputed male sites were used to estimate $\mathbf{G}_M$. The hybrid genotypes inferred from the 20,000 sites common to males and females via the *build.HMM* function were used to estimate $\mathbf{G}_H$. The specific combining ability

(SCA) relationship matrix $\mathbf{G}_{\mathrm{FM}}$ was the Kronecker product of the male and female additive relationship matrices (Bernardo, 2002). Pairwise SDAF, a hypothetical predictor of mid-parent heterosis, was calculated as:

$$SDAF_{ij} = \frac{\sum_{n=1}^{N} (p_{i_n} - p_{j_n})^2}{N} \quad (1)$$

Where $SDAF_{ij}$ was the squared difference in allele frequency between the i^{th} female parent and the j^{th} male parent, $p_{i_n} - p_{j_n}$ was the difference allele frequency between the i^{th} female parent and j^{th} male parent at the n^{th} variant site, and N was the total number of variant sites.

Modeling and Statistical Analysis

All linear models were fit in a single step with the *mmer* function in sommer (Covarrubias-Pazaran, 2016, 2018). For a given trait, i.e., plant height, tiller number, panicle length, average yield per plant, proportion of spikelets filled, or potential yield per plant, the genotype replicates (i.e., single-row plots) were considered the experimental unit. Genetic variances for plant height, tiller number, panicle length, and average yield per plant traits were estimated with models of the following form:

$$Y_{ijkl} = \mu + H_{\mathrm{i}} + E_j + HE_{\mathrm{ij}} + B_{(k)j} + \varepsilon_{ijkl} \quad (2)$$

where Y_{ijkl} was the random phenotypic response of the i^{th} single-cross hybrid genotype in the k^{th} block nested in the j^{th} environment from the l^{th} replicate, μ was the grand mean, H_{i} was the random effect of the i^{th} hybrid genotype with $N(0,\ \mathbf{I}\sigma^2_H)$, E_j was the random effect of the j^{th} environment with $N(0,\ \mathbf{I}\sigma^2_E)$, HE_{ij} was the random interaction of the i^{th} hybrid genotype and the j^{th} environment with $N(0,\ \mathbf{I}\sigma^2_{HE})$, $B_{(k)j}$ was the effect of the k^{th} block nested within the j^{th} environment with $N(0,\ \mathbf{I}\sigma^2_B)$, and ε_{ijkl} was the random error associated with each replicate with $N(0,\ \mathbf{I}\sigma^{\ 2}_e)$.

The genetic variance for proportion of spikelets filled and potential yield per plant was estimated using the following model in (3), without the environment term and its associated interactions, because the traits were only phenotyped in the irrigated lowland environment:

$$Y_{ijk} = \mu + H_{\mathrm{i}} + B_j + \varepsilon_{ijk} \quad (3)$$

Y_{ijk} was the random phenotypic response of the i^{th} hybrid genotype in the j^{th} block from the k^{th} replicate, μ was the grand mean, H_{i} was the random effect of the i^{th} hybrid genotype with $N(0,\ \mathbf{I}\sigma^2_H)$, B_j was the random effect of the j^{th} block with $N(0,\ \mathbf{I}\sigma^2_B)$, and ε_{ijk} was the random error associated with each replicate with $N(0,\ \mathbf{I}\sigma^{\ 2}_e)$.

The entry-mean heritability was estimated for each of height, tiller number, panicle length, and yield per plant by (4) following the method of Holland et al. (2003) for unbalanced RCBDs:

$$H^2 = \frac{\sigma^2_H}{\sigma^2_H + \frac{\sigma^2_{HE}}{h_j} + \frac{\sigma^2_e}{h_t}} \quad (4)$$

where σ^2_H was the variance among hybrid genotypes from models of the form in (2), σ^2_{HE} was the variance of the interaction of the hybrid genotype and environment, σ^2_e was the error variance, h_j was the harmonic mean of the number of total observations of each hybrid genotype within an environment, and h_t was the harmonic mean of the total number of observations per hybrid genotype.

For proportion of spikelets filled and potential yield, which were phenotyped in a single environment, entry-mean heritability was estimated by (5) also following Holland et al. (2003), using the following equation with the terms as described in (3):

$$H^2 = \frac{\sigma^2_H}{\sigma^2_H + \frac{\sigma^2_e}{h_t}} \quad (5)$$

Additive genetic variances were estimated from models of the form in (6) for plant height, tiller number, panicle length, and yield per plant. The terms of (6) are the same as in (2), but in (6) the random effect H was assumed to have a multivariate normal (MVN) distribution with $H \sim MVN(0,\ \mathbf{G}_{\mathrm{H}}\sigma^2_H)$, where $\mathbf{G}_{\mathrm{H}}$ was the additive genomic relationship matrix of the F_1 hybrids:

$$Y_{ijkl} = \mu + H_{\mathrm{i}} + E_j + HE_{\mathrm{ij}} + B_{(k)j} + \varepsilon_{ijkl} \quad (6)$$

Additive genetic variances for proportion of spikelets filled and potential yield were estimated from model of the form in (7), with the same terms as (3), but the random effect H was assumed to have a multivariate normal (MVN) distribution with $H \sim MVN(0,\ \mathbf{G}_{\mathrm{H}}\sigma^2_H)$, where $\mathbf{G}_{\mathrm{H}}$ was the additive genomic relationship matrix of the F_1 hybrids:

$$Y_{ijk} = \mu + H_{\mathrm{i}} + B_j + \varepsilon_{ijk} \quad (7)$$

Narrow-sense heritability, or the proportion of additive genetic variance out of total phenotypic variance, was estimated on a single-plant basis for all traits. Variance components were estimated from the models of the form in (6) for plant height, tiller number, panicle length, and yield per plant, and narrow-sense heritability was estimated with (8):

$$h^2 = \frac{\sigma^2_H}{\sigma^2_H + \sigma^2_{HE} + \sigma^2_e} \quad (8)$$

For proportion of spikelets filled and potential yield, narrow-sense heritability was estimated using (9) with variances estimated from the models of the form in (7):

$$h^2 = \frac{\sigma^2_H}{\sigma^2_H + \sigma^2_e} \quad (9)$$

For each trait, genomic best linear unbiased predictions (GBLUPs) of hybrid genetic value, male GCA, and female GCA were each estimated using two separate models: the genomic GCA model and the genomic GCA + SCA model. Model fits were compared with the Akaike information criterion (AIC) and Bayesian information criterion (BIC). The genomic GCA model was:

$$Y_{ijklm} = \mu + F_i + M_j + E_k + B_{(l)k} + FE_{ik} + ME_{jk} + \varepsilon_{ijklm} \quad (10)$$

where Y_{ijklm} was the random phenotypic response of a single-cross hybrid of the i^{th} female line and the j^{th} male line observed in the k^{th} environment, l^{th} block, and m^{th} replicate, μ was the grand mean, F_i was the random GCA effect of the i^{th} female parent with $F \sim MVN(0, \mathbf{G}_F\sigma^2_F)$ where $\mathbf{G}_F$ was the additive genomic relationship matrix among females, M_j was the random GCA effect of the j^{th} male parent with $M \sim MVN(0, \mathbf{G}_M\sigma^2_M)$ where $\mathbf{G}_M$ was the additive genomic relationship matrix among males, E_k was the effect of the k^{th} environment with $N(0, \mathbf{I}\sigma^2_E)$, $B_{(l)k}$ was the effect of the l^{th} block nested in the k^{th} environment with $N(0, \mathbf{I}\sigma^2_B)$, FE_{ik} was the random interaction of the i^{th} female and the k^{th} environment with $N(0, \mathbf{I}\sigma^2_{FE})$, ME_{jk} was the random interaction of the j^{th} male and the k^{th} environment with $N(0, \mathbf{I}\sigma^2_{ME})$, and ε_{ijklm} was the random error of each observation with $N(0, \mathbf{I}\sigma^2_e)$.

The genomic GCA + SCA model was:

$$Y_{ijklm} = \mu + F_i + M_j + E_k + B_{(l)k} + FE_{ik} + ME_{jk} + FM_{ij} + \varepsilon_{ijklm} \quad (11)$$

where terms were as described above, and FM_{ij} was the additional random SCA interaction effect of the i^{th} female and the j^{th} male, with $FM \sim MVN(0, \mathbf{G}_{FM}\sigma^2_{FM})$. $\mathbf{G}_{FM}$ was the Kronecker product of $\mathbf{G}_F$ and $\mathbf{G}_M$ (Bernardo, 2002).

Best linear unbiased predictions (BLUPs) of hybrid genetic value and male and female GCAs were also estimated without genomic information to 1) estimate the predictive ability and prediction accuracy of the genomic prediction models, and 2) estimate the accuracy of phenotypic selection. For the GCA model, all terms remained the same as in (10), but the distribution of the random effects of F and M were simply assumed to be $N(0, \mathbf{I}\sigma^2_F)$ and $N(0, \mathbf{I}\sigma^2_M)$ respectively.

$$Y_{ijklm} = \mu + F_i + M_j + E_k + B_{(l)k} + FE_{ik} + ME_{jk} + \varepsilon_{ijklm} \quad (12)$$

Similarly, for the GCA + SCA model, all terms remained the same as in (11), but the distribution of the random effects F, M and FM were assumed to be $N(0, \mathbf{I}\sigma^2_F)$, $N(0, \mathbf{I}\sigma^2_M)$, and $N(0, \mathbf{I}\sigma^2_{FM})$ respectively:

$$Y_{ijklm} = \mu + F_i + M_j + E_k + B_{(l)k} + FE_{ik} + ME_{jk} + FM_{ij} + \varepsilon_{ijklm} \quad (13)$$

There are multiple methods to estimate genomic prediction accuracy (Estaghvirou et al., 2013). Here, predictive ability was Pearson's correlation of an estimated value and a true value. Prediction accuracy was considered to be predictive ability divided by the square root of the reliability of the estimated value (Mrode, 2014). This method of estimating prediction accuracy, which is well-established in animal breeding programs, was chosen because it is relatively unbiased, precise, and stable compared to other methods (Estaghvirou et al., 2013). Predictive abilities of the genomic GCA and genomic GCA + SCA models described in (10) and (11) were each estimated for male GCA by ten-fold cross-validation (Resende et al., 2012). Sample size was inadequate to estimate genomic predictive ability and accuracy for female GCA. In each fold, the hybrid progeny phenotypes of 38 of the males were masked from the training set, and the masked training set was used to train the prediction model. Predictive abilities for male GCA, or Pearson's correlation of the GBLUP of male GCA estimated in the training set fold and the BLUP of male GCA estimated from all available observations in the full dataset, were then calculated for the 38 masked males and averaged across folds for each model. Prediction accuracy for male GCA for each model was the predictive ability divided by the square root of the reliability of the genomic prediction. For each model, the square root of the reliability of the genomic prediction was the correlation of the GBLUP of male GCA and BLUP of male GCA when all available observations were used for estimation of both, as in (10), (11), (12), and (13).

Predictive abilities and prediction accuracies were also estimated for hybrid genetic values for each phenotypic response following Technow et al. (2014). For each of 500 replications, four females and 127 males which had been crossed to at least one of the four females were randomly sampled. Then, a training set was formed by randomly sampling 150 hybrids which had phenotypic records available, with the constraint that the randomly sampled male and female lines had at least one hybrid progeny in the training set. Predictive ability, here Pearson's correlation of the hybrid genetic values estimated from the training fold and the observed hybrid genetic values, was recorded for hybrids which were not included in the training set. Predictive ability was recorded separately for hybrids which had both parents included in the training set (T2), one parent included in the training set (T1), only the female parent included in the training set (T1F), only the male parent included in the training set (T1M), and neither parent included in the training set (T0). The prediction accuracy was the predictive ability divided by the square root of the reliability of the genomic prediction, here the correlation of the genomic BLUP of hybrid genetic value and BLUP of hybrid genetic value when all available observations were used for estimation of both.

The accuracies of phenotypic selection for each trait were estimated for male GCA, female GCA, and hybrid genetic value as the square root of the reliabilities of their respective BLUPs (Falconer and Mackay, 1996; Mrode, 2014). Reliabilities of male GCA and female GCA were estimated from each the GCA model in (12) and the GCA + SCA model in (13). Reliabilities of hybrid genetic values were estimated from model (2) and (3). To estimate each reliability, the prediction error variances (*PEV*) of the appropriate BLUPs (i.e., of hybrid genetic value, female GCA, or male GCA) were obtained in sommer by inverting the coefficient matrix of the relevant model. For the BLUPs of male GCA for a given model and trait, the reliability was the average of:

$$1 - \frac{PEV_j}{\sigma^2_M} \quad (14)$$

where PEV_j was the prediction error variance of the j^{th} BLUP of male GCA and σ^2_M was the estimated male GCA variance. For the

BLUPs of female GCA for a given model and trait, the reliability was the average of:

$$1 - \frac{PEV_i}{\sigma_F^2} \quad (15)$$

where PEV_i was the prediction error variance of the i^{th} BLUP of female GCA, and σ_F^2 was the estimated female GCA variance. For the BLUPs of hybrid genetic value for a given trait from model (2), the reliability was the average of:

$$1 - \frac{PEV_i}{\sigma_H^2} \quad (16)$$

where PEV_i was the prediction error variance of the i^{th} BLUP of hybrid genetic value, and σ_H^2 was the estimated variance among hybrids.

To assess correlation of SDAF with mid-parent heterosis, mid-parent heterosis of each F_1 hybrid was estimated using BLUPs of genetic values from the following model:

$$Y_{ijklm} = \mu + F_i + G_j + E_k + GE_{jk} + B_{(l)k} + \varepsilon_{ijklm} \quad (17)$$

Y_{ijklm} was the random phenotypic response and μ was the grand mean. In the vein of Xiang et al. (2016) and Liang et al. (2018), F_i was the fixed effect of an indicator of whether the genotype was an inbred or F_1 hybrid to account for the possibility of differing inbred and hybrid group means in the presence of heterosis. G_j was the random effect of the j^{th} inbred parent, inbred commercial check, or F_1 hybrid genotype with $N(0, \mathbf{I}\sigma_G^2)$ E_k was the random effect of the k^{th} environment with $N(0, \mathbf{I}\sigma_E^2)$, GE_{jk} was the random interaction effect of the j^{th} genotype and the k^{th} environment with $N(0, \mathbf{I}\sigma_{GE}^2)$, $B_{(l)k}$was the random effect of the l^{th} block nested within the k^{th} environment with $N(0, \mathbf{I}\sigma_B^2)$, and $\varepsilon_{ijklm}$was the random error associated with each observation with $N(0, \mathbf{I}\sigma_e^2)$. The environment term and its associated interactions were dropped in estimation of yield potential and proportion of spikelets filled, because they were observed only in the irrigated lowland environment.

Mid-parent heterosis of each F_1 hybrid was obtained as:

$$MPH = \frac{\hat{H} - \widehat{MP}}{\widehat{MP}} \quad (18)$$

where MPH was mid-heterosis, $\hat{H}$ was the BLUP of the F_1 genotype value from (17), and $\widehat{MP}$ was the mid-parent value, i.e., the mean of the BLUPs from (17) of its parental genotype values (**Supplementary File 1**). The BLUP of the i^{th} genotype value was the sum of μ, $\hat{F}_i$, and $\hat{G}_j$ from (17). Because the CMS parents do not set seed, the corresponding maintainer (B) line phenotype was used to estimate heterosis. Pearson's correlation of SDAF with mid-parent heterosis was estimated among all hybrids in the study and also separately within families of hybrids with the same female parent. Student's *t* test of significance was conducted at $\alpha = 0.05$ for each correlation, with the null hypothesis that a given correlation did not significantly differ from zero and the alternate hypothesis that the given correlation significantly differed from zero.

Commercial relative performance (commercial heterosis) was estimated for each F_1 hybrid with phenotypic observations against each check as:

$$CRP = \frac{\hat{H} - \widehat{C}}{\widehat{C}} \quad (19)$$

where CRP was commercial relative performance, $\hat{H}$ was the BLUP of the F_1 hybrid genotype value from (17), and $\widehat{C}$ was the BLUP of the commercial check genotype value from (17). The BLUP of the i^{th} genotype value was the sum of μ, $\hat{F}_i$, and $\hat{G}_j$ from (17).

RESULTS

Summary Statistics and Heritabilities

Mean phenotypic values for height, tiller number, panicle length, proportion of spikelets filled, yield per plant, and potential yield were respectively 85 cm, 14 tillers, 226 mm, 0.757, 36 g per plant, and 54 g per plant (**Supplementary Figure 1**; **Table 1**). Highest entry-mean heritability observed was for height at 0.906, and lowest entry-mean heritability observed was for potential yield at 0.311 (**Table 2**). Narrow-sense heritabilities on a single-plant basis were greatest for the proportion of spikelets filled at 0.864 and least for potential yield at 0.271 (**Table 2**). Because the narrow-sense heritability of proportion of spikelets filled was high, perhaps due to differing genetic architectures between TGMS and CMS lines for the trait, we also estimated the narrow-sense heritability in the CMS lines only as 0.922 by removing observations of the TGMS line A07 from the model. Principal components analysis of the parental genotypes showed clustering of the males and females, but divergence of the male and female parents was not due to historical reciprocal recurrent selection (**Supplementary Figure 2**).

Model Fit, Predictive Ability, and Prediction Accuracy

For all traits, model fit was superior for the genomic GCA + SCA model compared to the genomic GCA model as assessed by

TABLE 1 | Trait mean phenotypic values and standard deviations overall and within environments for the F_1 hybrids.

Trait	Mean ± Standard Deviation, Overall	Mean ± Standard Deviation, Irrigated Lowland	Mean ± Standard Deviation, Irrigated Upland
Height (cm)	85 ± 15	89 ± 13	70 ± 11
Tiller Number	14 ± 5	16 ± 4	10 ± 3
Panicle Length (mm)	226 ± 24	231 ± 22	213 ± 25
Proportion of Spikelets Filled*		0.757 ± 0.149	
Yield per Plant (g)	36 ± 19	44 ± 17	18 ± 7
Potential Yield* (g)		54 ± 20	

**Proportion of spikelets filled and potential yield were only phenotyped in the irrigated lowland environment.*

TABLE 2 | Estimates and their standard errors for each trait of entry-mean heritability of the F_1 hybrids and narrow-sense heritability on a single-plant basis.

Trait	Entry-Mean Heritability	Narrow-Sense Heritability
Height	0.906 ± 0.007	0.719 ± 0.029
Tiller Number	0.505 ± 0.034	0.376 ± 0.054
Panicle Length	0.824 ± 0.014	0.542 ± 0.046
Proportion of Spikelets Filled	0.780 ± 0.014	0.864 ± 0.014*
Yield per Plant	0.433 ± 0.038	0.328 ± 0.055
Potential Yield	0.311 ± 0.046	0.271 ± 0.063

**In the CMS lines only, the narrow-sense heritability for proportion of spikelets filled was 0.922 ± 0.008.*

either AIC or BIC (**Supplementary Tables 2, 3**). However, the predictive abilities and accuracies of the genomic GCA + SCA models were not substantially different from the genomic GCA models (**Supplementary Tables 4, 5**; **Tables 3**, **4**). Mean prediction accuracies for male GCA ranged from 0.215 to 0.318 for the genomic GCA models and 0.233 to 0.332 in the genomic GCA + SCA models (**Table 4**). Prediction accuracies for untested females were not estimated. For hybrids in the T0 set, mean genomic GCA model accuracies ranged 0.039 to 0.394, while the genomic GCA + SCA model accuracies ranged from 0.043 to 0.490. For hybrids in the T1 set, mean genomic GCA model accuracies ranged from 0.476 to 0.806, while the genomic GCA + SCA model accuracies ranged from 0.509 to 0.827. For hybrids in the T1F set, mean genomic GCA model accuracies ranged from 0.310 to 0.908, while the genomic GCA + SCA model accuracies ranged from 0.364 to 0.943. For hybrids in the T1M set, mean genomic GCA model accuracies ranged from 0.537 to 0.742, while the genomic GCA + SCA model accuracies ranged from 0.423 to 0.785. For hybrids in the T2 set, mean genomic GCA model accuracies ranged from 0.769 to 0.948, while the genomic GCA + SCA model accuracies ranged from 0.772 to 0.956 (**Table 3**). However, in the hybrid T0 case, accuracy for yield per plant was increased from 0.215 to 0.490 by inclusion of the SCA effect. No other trait had more than a 10% increase in accuracy by inclusion of the SCA effect in the T0 case, and substantial increases in accuracy with inclusion of the SCA effect were not observed for yield per plant in scenarios besides T0. Except in the case of proportion of spikelets filled, accuracy in the T1F scenarios was always substantially higher than the T1M scenarios.

The accuracy of phenotypic selection for hybrid genetic value ranged from 0.566 to 0.952 among traits (**Table 5**). For the GCA model, the accuracy of phenotypic selection for male GCA ranged from 0.484 to 0.861, and the accuracy of phenotypic selection for female GCA ranged from 0.853 to 0.910 (**Tables 6**, **7**; **Supplementary Tables 6, 7**). For the GCA + SCA model, the accuracy of phenotypic selection for male GCA ranged from 0.253 to 0.798, and the accuracy of phenotypic selection for female GCA ranged from 0.850 to 0.910 (**Tables 6**, **7**; **Supplementary Tables 6, 7**).

Hybrid Genetic Value and Parental GCA

The genomic GCA + SCA model was used to rank hybrid genetic values. The maximum predicted F_1 hybrid yield was 43.352 grams per plant, which scaled to 8.670 tons per hectare (**Supplementary Table 8**). The maximum predicted F_1 hybrid potential yield was 77.401 grams per plant, scaling to 15.479 tons per hectare (**Supplementary Table 9**). Over half of the top 20 F_1 hybrids in terms of yield per plant had phenotypic observations available, though the top-ranked hybrid did not. The relative performance of the F_1 hybrids compared to the commercial inbred checks (commercial heterosis) ranged from −43.9% to 70.0% for yield per plant (**Supplementary Figure 3**; **Table 8**). The maximum

TABLE 3 | Mean prediction accuracies ± standard error thereof in cross-validation of the genomic prediction models for hybrids.

Trait	T0	T1	T1F	T1M	T2
Genomic GCA model					
Height	0.345 ± 0.013	0.793 ± 0.009	0.908 ± 0.007	0.562 ± 0.006	0.948 ± 0.003
Tiller Number	0.162 ± 0.005	0.628 ± 0.014	0.644 ± 0.015	0.537 ± 0.008	0.812 ± 0.007
Panicle Length	0.394 ± 0.012	0.806 ± 0.009	0.906 ± 0.008	0.574 ± 0.005	0.942 ± 0.005
Proportion of Spikelets Filled	0.039 ± 0.003	0.476 ± 0.003	0.310 ± 0.006	0.742 ± 0.003	0.769 ± 0.003
Yield per Plant	0.215 ± 0.012	0.688 ± 0.008	0.741 ± 0.008	0.646 ± 0.007	0.820 ± 0.005
Potential Yield	0.380 ± 0.015	0.719 ± 0.011	0.842 ± 0.009	0.566 ± 0.011	0.855 ± 0.006
Genomic GCA + SCA model					
Height	0.414 ± 0.019	0.820 ± 0.010	0.943 ± 0.008	0.498 ± 0.014	0.956 ± 0.005
Tiller Number	0.190 ± 0.007	0.640 ± 0.016	0.676 ± 0.016	0.442 ± 0.012	0.772 ± 0.013
Panicle Length	0.446 ± 0.020	0.827 ± 0.011	0.940 ± 0.011	0.423 ± 0.018	0.928 ± 0.011
Proportion of Spikelets Filled	0.043 ± 0.003	0.509 ± 0.004	0.364 ± 0.007	0.785 ± 0.005	0.809 ± 0.004
Yield per Plant	0.490 ± 0.022	0.715 ± 0.009	0.795 ± 0.009	0.532 ± 0.014	0.822 ± 0.006
Potential Yield	0.415 ± 0.016	0.723 ± 0.012	0.847 ± 0.010	0.525 ± 0.014	0.851 ± 0.007

For each of the 500 cross-validation folds, 4 female parents and 127 male parents were chosen to have hybrid progeny included in the training set. The total number of hybrid genotypes sampled for training was 150. Accuracies are reported for hybrids which were not included in the training set which had neither parent in the training set (T0), one parent in the training set (T1), the female parent only in the training set (T1F), the male parent only in the training set (T1M), and both parents in the training set (T2).

TABLE 4 | Model prediction accuracy and standard error for male GCA for each trait as estimated by ten-fold cross-validation.

Trait	Genomic GCA model	Genomic GCA + SCA model
Height	0.232 ± 0.056	0.233 ± 0.067
Tiller Number	0.215 ± 0.060	0.261 ± 0.065
Panicle Length	0.224 ± 0.043	0.269 ± 0.046
Proportion of Spikelets Filled	0.318 ± 0.083	0.332 ± 0.079
Yield per Plant	0.219 ± 0.072	0.241 ± 0.079
Potential Yield	0.292 ± 0.078	0.233 ± 0.096

TABLE 5 | Reliabilities and accuracies of phenotypic selection for hybrid performance.

Trait	Reliability	Accuracy
Height	0.906	0.952
Tiller Number	0.533	0.730
Panicle Length	0.828	0.910
Proportion of Spikelets Filled	0.791	0.889
Yield per Plant	0.466	0.682
Potential Yield	0.321	0.566

TABLE 6 | Accuracies of phenotypic selection for male GCA.

Trait	GCA model	GCA + SCA model
Height	0.809	0.627
Tiller Number	0.644	0.570
Panicle Length	0.680	0.253
Proportion of Spikelets Filled	0.861	0.798
Yield per Plant	0.640	0.562
Potential Yield	0.484	0.454

TABLE 7 | Accuracies of phenotypic selection for female GCA.

Trait	GCA model	GCA + SCA model
Height	0.906	0.906
Tiller Number	0.860	0.858
Panicle Length	0.910	0.910
Proportion of Spikelets Filled	0.898	0.876
Yield per Plant	0.853	0.850
Potential Yield	0.900	0.899

genomic predicted GCA for yield per plant in females and males respectively were 36.341 and 34.047; both of the female and male lines top-ranked for GCA had phenotypic observations available (**Supplementary Tables 10, 11**).

Mid-Parent Heterosis and Parental SDAF

Average mid-parent heterosis was positive, though not extremely so, for all traits except height and proportion of spikelets filled (**Figure 2** and **Table 9**). Yield per plant and its component trait, tiller number, showed the highest average heterosis; average heterosis for yield was 32.0%. Parental SDAF ranged from 0.200 to 0.285 in the phenotyped F_1 hybrids (**Supplementary Figure 4**).

TABLE 8 | Mean, standard deviation, and range of relative yield per plant of the F_1 hybrids compared to each inbred check.

Check	Mean Relative Performance ± SD	Range
WTR-1	0.206 ± 0.143	−0.238—0.679
Y 134 (DP 6)	0.191 ± 0.141	−0.248—0.659
Khazar (DP 8)	0.283 ± 0.147	−0.218—0.724
OM 997 (DP 10)	0.283 ± 0.152	−0.190—0.786
M 401 (DP 17)	0.179 ± 0.140	−0.256—0.642
X 21 (DP 19)	0.289 ± 0.153	−0.186—0.796

Overall, in all hybrids, parental SDAF was significantly correlated with mid-parent heterosis for all traits except proportion of spikelets filled (**Figure 3** and **Table 10**). Interestingly, the direction of the correlation was negative for all traits except tiller number. The strongest correlation of mid-parent heterosis and SDAF for hybrids overall was for panicle length. However, when hybrids were grouped into families by female parent, there were no significant correlations between parental SDAF and mid-parent heterosis.

DISCUSSION

The objectives of this study were (1) to identify high-yielding F_1 hybrids from crosses of IRRI male sterile lines with stress-tolerant male lines, (2) to identify parental lines with high GCA for use in future reciprocal recurrent genomic selection programs, and (3) to evaluate genomic prediction and phenotypic selection accuracies in our hybrid breeding population, with the end goal of developing stress-tolerant hybrid rice varieties. Compared to inbred commercial checks (which were also progenitors of the male lines), phenotyped F_1 hybrids showed genetic yield advantages of up to 80% and warrant further testing (**Supplementary Figure 3**; **Table 8**). Although the genetic yield of the top-performing F_1 hybrid observed in the study environment was 8.670 tons per hectare, this measure pertains to the study environment only—which included both standard irrigated conditions and stressful upland conditions—and only the plant population densities used, which were lower than those observed in farmers' fields (**Supplementary Table 8**). Relative to the mid-parent, the F_1 hybrids showed on average 32.0% mid-parent heterosis for yield, which is consistent with literature averages of 10 to 30% in rice (**Figure 2** and **Table 9**; Janaiah and Xie, 2010; Longin et al., 2012; Spielman et al., 2013). However, the maximum mid-parent heterosis observed for yield was 89.2%, and heterosis up to 48.3% was observed within a single standard deviation of the mean (**Figure 2** and **Table 9**). Because heterosis is present in the F_1 hybrids and SCA as well as GCA variance was detected, recurrent selection for GCA in the male and female lines tested should allow development of heterotic pools. However, we did not evaluate the relative efficiency of hybrid vs. line breeding for our population given the estimated GCA and SCA variance detected in our population or other relevant factors, and further investigation of this topic is warranted.

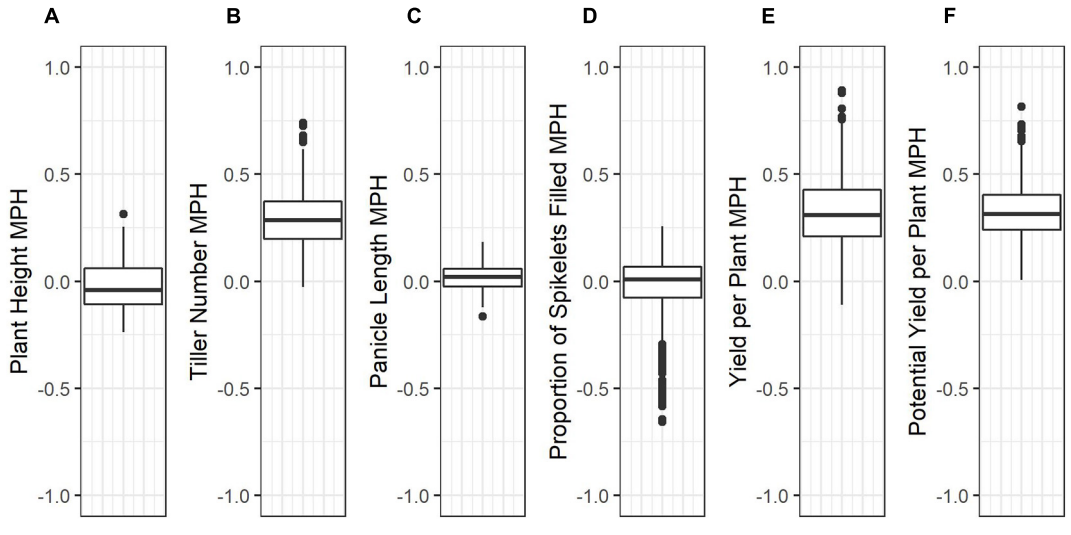

FIGURE 2 | Box plots of mid-parent heterosis for each trait in the phenotyped F_1 hybrids. **(A)** Plant height (cm). **(B)** Tiller number. **(C)** Panicle length (mm). **(D)** Proportions of spikelets filled. **(E)** Yield per plant (g). **(F)** Potential yield per plant (g).

Fertility Restoration and Genetic Architecture of Spikelet Filling

Interestingly, narrow-sense heritability for the proportion of spikelets filled was relatively high at 0.600 (**Table 2**). Although variation in grain fill is generally driven by environmental factors in rice, with starch synthesis and deposition depending on available assimilate, in rice hybrids absence of grain fill can be due to spikelet sterility from lack of fertility-restoring (*Rf*) alleles (Peng et al., 1998; Tang et al., 2017). Because fertility restoration is only relevant in hybrids of CMS parents, not TGMS parents, we estimated narrow-sense heritability for proportion of spikelets filled in the CMS-derived hybrids only as 0.922 (**Table 2**). The relatively high proportion of phenotypic variance explained by additive genetic variance for this trait in CMS lines suggests that segregation for fertility restoration played a role in the proportion of spikelets filled in the CMS-derived hybrids and as such in observed yield per plant. Concordantly, selection accuracy for proportion of spikelets filled was substantially higher in the T1M hybrids than T1F hybrids in cross-validation, though T1F hybrids had higher accuracies than T1M for all other traits. Screening the population for known major fertility restoration alleles at *Rf3* and *Rf4* may allow the use of marker-assisted selection to improve selection accuracy (Tang et al., 2017). It may also be possible to select for fertility restoration by genomic prediction rather than mapping fertility-restoring alleles of more minor or modifying effect. Selection for fertility restoration may effectively unlock observed yield potential in future hybrids and fix *Rf* alleles in the male heterotic pool.

Accuracies of Genomic Prediction for Hybrid Genetic Value and Parent GCA

Genomic prediction model accuracies were high in unobserved T2 F_1 hybrids, for which both parents were included in the training set (**Table 3**). All F_1 hybrids surveyed were closely related; closely related individuals have smaller effective population size, which reduces the effective number of loci controlling traits and is expected to increase prediction accuracy (**Supplementary Figure 5**; Daetwyler et al., 2010). Accuracy appeared to be driven primarily by estimation of the female line effects, and accuracy in T2 hybrids was not substantially different from T1F hybrids (**Table 3**). The exception was proportion of spikelets filled, for which the male (restorer) line effects were more relevant. As expected, accuracy in the T0 hybrids was low, though positive and improved substantially for yield per plant by the inclusion of SCA effects in the model (**Table 3**).

Accuracy was low for genomic estimated male GCA (< 0.300) despite that the males all shared a recurrent parent and as such a large proportion of their genomes (75% ± Mendelian sampling and selection; **Supplementary Figure 6**; **Table 4**). Although the male donor progenitors were diverse, multiple male lines per donor were sampled. Low accuracies of genomic predictions of male GCA may have been due to highly unbalanced crossing of males to females, with no single male crossed to all females. It was not possible to estimate accuracy for genomic estimated GCA in the females, which were also closely related but more extensively phenotyped (**Supplementary Figure 7**).

TABLE 9 | Mean, standard deviation, and range of mid-parent heterosis for each trait.

Trait	Mean Mid-Parent Heterosis ± SD	Range
Height	−0.019 ± 0.106	−0.237−0.313
Tiller Number	0.292 ± 0.132	−0.026−0.741
Panicle Length	0.017 ± 0.054	−0.163−0.183
Proportion of Spikelets Filled	−0.034 ± 0.160	−0.657−0.258
Yield per Plant	0.320 ± 0.163	−0.110−0.892
Potential Yield	0.327 ± 0.131	0.005−0.817

Phenotypic accuracies were similar to or lower than genomic prediction accuracies for hybrid performance in the T2 case (**Table 5**; Rutkoski et al., 2015). Notably, accuracy of genomic prediction of hybrid yield (0.820) was greater than accuracy of the phenotypes (0.682). However, for male GCA, the phenotypic accuracies greatly exceeded those of genomic prediction for both the GCA and the GCA + SCA models (**Table 6**). It was not possible to compare genomic and phenotypic selection accuracies for female GCA.

Inconsistent Correlations of Mid-Parent Heterosis and Parental SDAF

Considering all hybrids in the study, we observed parental SDAF to be negatively correlated with mid-parent heterosis and hybrid genetic value for all traits surveyed except tiller number, for which SDAF was positively correlated with mid-parent heterosis (**Figure 3** and **Table 10**). In many species, parental SDAF (or other measures of genetic distance) is positively correlated with heterosis due to release from inbreeding depression to a point, as dominant alleles mask deleterious recessive alleles in hybrids (Falconer and Mackay, 1996). However, as genetic distance increases, outbreeding depression eventually prevails as favorable epistatic combinations of genes are separated (Lynch and Walsh, 1998). A common manifestation of outbreeding depression is fertility barriers (Edmands, 2002). In the case of yield per plant, we cannot eliminate the possibility that genetic distance is correlated with absence of wide-cross compatibility alleles known to affect seed set, given the inter-subspecific diversity present in the male lines (Ji et al., 2005). However, given the intense selection on the male lines, it seems possible that wide-cross compatibility in the males may have also been positively and indirectly selected with yield. Genetic distance could also be correlated with absence of fertility restoring alleles by chance. However, yield potential corrects for fertility restoration by estimating yield as if all spikelets were filled to the average weight observed in the study, and overall mid-parent heterosis for yield potential was also negatively correlated with parental SDAF. Importantly, though unsurprisingly given the relationships of the BC_1F_5 male lines, the correlation of parental SDAF and mid-parent heterosis was not observed within female families of hybrids (**Table 10**). This suggests that the negative correlations observed in hybrids overall were due to differences in female genetic distance from the average male. More crucially for practical purposes, whether genetic distance is indicative of mid-parent heterosis depends on the population defined, even in closely related hybrids.

TABLE 10 | Pearson's correlation coefficient of mid-parent heterosis and SDAF with 95% confidence intervals of the coefficient and *t*-tests of significance conducted at $\alpha = 0.05$.

Trait	$r \pm 95\%$ CI	*t*	*df*	*P*
Height				
Overall	−0.330 ± 0.062	−9.791	783	< 0.001
10A	−0.041 ± 0.157	−0.512	154	0.610
2A	0.007 ± 0.176	0.077	122	0.939
4A	0.031 ± 0.145	0.419	181	0.676
6A	−0.044 ± 0.227	−0.381	73	0.705
7A	−0.073 ± 0.237	−0.592	66	0.556
A07	0.084 ± 0.146	1.118	177	0.265
Tiller Number				
Overall	0.299 ± 0.064	8.725	778	< 0.001
10A	−0.143 ± 0.155	−1.786	152	0.076
2A	0.058 ± 0.176	0.644	121	0.521
4A	0.027 ± 0.145	0.357	180	0.722
6A	0.032 ± 0.227	0.269	73	0.788
7A	−0.084 ± 0.239	−0.677	65	0.501
A07	0.045 ± 0.146	0.601	177	0.549
Panicle Length				
Overall	−0.430 ± 0.057	−13.309	783	< 0.001
10A	−0.042 ± 0.157	−0.516	154	0.607
2A	−0.084 ± 0.175	−0.935	122	0.352
4A	0.081 ± 0.144	1.094	181	0.275
6A	0.066 ± 0.226	0.568	73	0.572
7A	0.013 ± 0.238	0.103	66	0.918
A07	0.058 ± 0.146	0.772	177	0.441
Proportion of Spikelets Filled				
Overall	−0.015 ± 0.081	−0.363	588	0.717
10A	0.127 ± 0.156	1.578	151	0.117
2A	0.063 ± 0.177	0.687	120	0.493
4A	0.103 ± 0.146	1.364	174	0.174
6A	0.027 ± 0.228	0.23	72	0.818
7A	0.083 ± 0.242	0.663	63	0.510
Yield per Plant				
Overall	−0.131 ± 0.069	−3.698	781	< 0.001
10A	0.014 ± 0.157	0.169	154	0.866
2A	0.041 ± 0.176	0.450	122	0.653
4A	0.116 ± 0.144	1.562	180	0.120
6A	0.040 ± 0.227	0.346	73	0.730
7A	−0.022 ± 0.238	−0.176	66	0.861
A07	0.104 ± 0.146	1.389	176	0.167
Potential Yield				
Overall	−0.092 ± 0.08	−2.225	585	0.026
10A	−0.031 ± 0.159	−0.386	151	0.700
2A	−0.005 ± 0.178	−0.059	120	0.953
4A	0.087 ± 0.148	1.141	171	0.255
6A	0.042 ± 0.228	0.360	72	0.720
7A	−0.008 ± 0.244	−0.060	63	0.952

Overall correlations as well as correlations within female families are reported. Correlations are not available within the A07 female family for potential yield or proportion of spikelets filled because phenotypic observations of the A07 line were not available at the location in which the traits were phenotyped, irrigated lowland.

Future Directions for IRRI Hybrid Rice Breeding

Based on the study findings, we caution against the conventional wisdom that increased genetic distance between parents alone will always confer improved hybrid performance or positive heterosis. Increased genetic distance in the potential founders of heterotic pools of rice screened was not reliably associated with desired positive heterosis for yield, even though the pools could be genetically distinguished. In this population, and probably in rice more generally, empirical selection for GCA is preferable to selection based on genetic distance to breed high-performing hybrid rice.

For hybrid performance, genomic prediction accuracies were similar to or higher than phenotypic accuracies. Therefore,

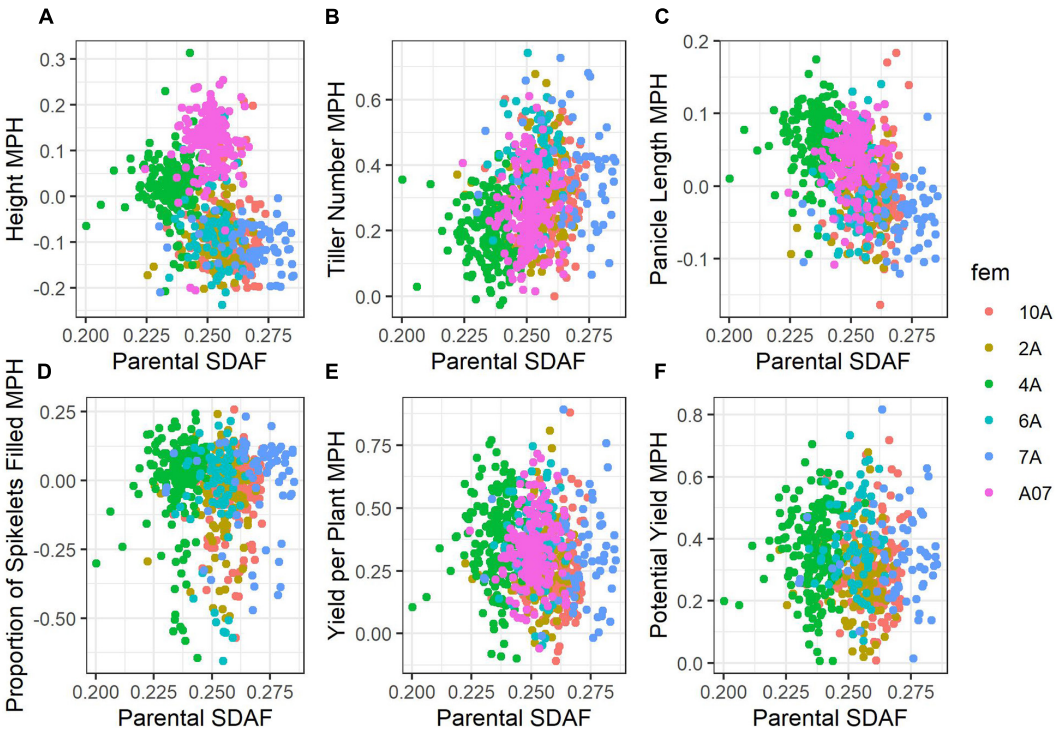

FIGURE 3 | Scatterplots of mid-parent heterosis against parental SDAF. Points are colored according to female parent. **(A)** Height. **(B)** Tiller number. **(C)** Panicle length. **(D)** Proportion grains filled. **(E)** Yield per plant. **(F)** Potential yield.

genomic prediction could be useful for product development in the population of study. Most notably, inclusion of genomic information increased prediction accuracy for hybrid yield per plant by approximately 13.8% compared to phenotype alone (**Tables 3**, **5**). Observed accuracies of prediction of unobserved hybrids with at least one parent in the training population (T1) were also positive and substantial, suggesting that on average genomic prediction could allow identification of further crosses with high value in the population of study. Genomic prediction accuracies for hybrids with neither parent observed (T0) were not as high as in the T1 case, but were nonetheless positive.

In contrast, genomic prediction accuracies for male GCA were substantially lower than phenotypic accuracies. For yield, genomic prediction accuracies were approximately three times less than phenotypic selection accuracies. However, reciprocal recurrent genomic selection for GCA can reduce cycle length by two-thirds compared to reciprocal recurrent phenotypic selection, because parents can be immediately recycled using genomic predictions of their GCA, leading to a cycle length of one (Powell et al., 2020). In conventional reciprocal recurrent selection, it is necessary to cross new parents to the opposing pool and phenotype the inter-pool crosses to estimate GCA before intra-pool recycling is possible, which increases the cycle length to three (Rembe et al., 2019). The genomic prediction accuracies observed in the study would provide comparable genetic gain to phenotypic selection if used to reduce the breeding cycle length to one, assuming that reduction in cycle length has no effect on genomic prediction accuracy and that accuracy of genomic prediction of female GCA (which could not be estimated, but for which more observations per female were available) is the same or higher than male GCA (Rembe et al., 2019; Powell et al., 2020). Increases in accuracies of genomic prediction of GCA relative to phenotypic selection are likely possible, as the training set of related hybrids would build over time in a closed population, and more complete and informative crossing designs could provide phenotypes. The potential of hybrid breeding strategies in IRRI germplasm would benefit from further assessment by simulation (Faux et al., 2016; Gaynor et al., 2020).

DATA AVAILABILITY STATEMENT

The phenotypic datasets generated and analyzed for this study can be found in **Supplementary File 1**. The raw genotype data used in this study can be found at NCBI under accession PRJNA479931, excluding the female genotype data. The female genotype data for this article are not publicly available as they are the property of the International Rice Research Institute. Requests to access the female genotype data should be directed to JA at j.ali@irri.org.

AUTHOR CONTRIBUTIONS

ML planned the study, executed the field trial and analysis, and wrote the manuscript. JA, MA, EA, MP, and MS developed

plant materials used. EA, MP, and MS directed field operations. AL supervised the statistical analysis. AS supervised the marker analysis. JR supervised the quantitative genetic analysis. JA conceived of the study, directed the study, and provided breeding insights. All authors edited and approved the final manuscript.

FUNDING

Funding for this study was provided by the United States Agency for International Development and Purdue University (Grant No. 208452). Green Super Rice (GSR) materials used in this study were developed under the funding of the GSR Project (IDOPP1130530) to JA as a research sub-grant by the Bill & Melinda Gates Foundation (BMGF). BMGF supports the open access publication fees of this manuscript.

ACKNOWLEDGMENTS

We thank the IRRI hybrid rice breeding team, past and present, for their contributions to the study. We thank Rey de la Cueva and Benvenido Bacani for technical expertise and assistance. We thank Erik J. Sacks for co-authoring Grant No. 208452 which partially funded the study. We thank Giovanny Eduardo Covarrubias-Pazaran for developing the open-source R package *sommer* and associated vignettes. We thank our reviewers for their time and helpful insights.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.692870/full#supplementary-material

REFERENCES

Abebrese, S. O., Martey, E., Dartey, P. K. A., Akromah, R., Gracen, V. E., Offei, S. K., et al. (2019). Farmer preferred traits and potential for adoption of hybrid rice in Ghana. *Sustain. Agric. Res.* 8, 38–48. doi: 10.5539/sar.v8n3p38

Ali, J., Aslam, U. M., Tariq, R., Murugaiyan, V., Schnable, P. S., Li, D., et al. (2018). Exploiting the genomic diversity of rice (*Oryza sativa* L.): SNP-typing in 11 early-backcross introgression-breeding populations. *Front. Plant Sci.* 9:849. doi: 10.3389/fpls.2018.00849

Ali, J., Xu, J. L., Gao, Y. M., Ma, X. F., Meng, L. J., Wang, Y. et al. (2017). Harnessing the hidden genetic diversity for improving multiple abiotic stress tolerance in rice (*Oryza sativa* L.). *PLoS One* 12:e0172515. doi: 10.1371/journal.pone.0172515

Amuzu-Aweh, E. N., Bijma, P., Kinghorn, B. P., Vereijken, A., Visscher, J., van Arendonk, J. A., et al. (2013). Prediction of heterosis using genome-wide SNP-marker data: application to egg production traits in white Leghorn crosses. *Heredity* 111, 530–538. doi: 10.1038/hdy.2013.77

Bernardo, R. (2002). *Breeding for Quantitative Traits in Plants*. Woodbury, MN: Stemma Press.

Boeven, P. H., Zhao, Y., Thorwarth, P., Liu, F., Maurer, H. P., Gils, M., et al. (2020). Negative dominance and dominance-by-dominance epistatic effects reduce grain-yield heterosis in wide crosses in wheat. *Sci. Adv.* 6:eaay4897. doi: 10.1126/sciadv.aay4897

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015

Chou, H. H., and Holmes, M. H. (2001). DNA sequence quality trimming and vector removal. *Bioinformatics* 17, 1093–1104. doi: 10.1093/bioinformatics/17.12.1093

Comstock, R. E., Robinson, H. F., and Harvey, P. H. (1949). A breeding procedure designed to make maximum use of both general and specific combining ability 1. *Agron. J.* 41, 360–367. doi: 10.2134/agronj1949.00021962004100080006x

Covarrubias-Pazaran, G. (2016). Genome assisted prediction of quantitative traits using the R package sommer. *PLoS One* 11:e0156744. doi: 10.1371/journal.pone.0156744

Covarrubias-Pazaran, G. (2018). Software update: moving the R package sommer to multivariate mixed models for genome-assisted prediction. *bioRxiv [Preprint]* doi: 10.1101/354639

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185, 1021–1031. doi: 10.1534/genetics.110.116855

de Mendiburu, F. (2020). *agricolae: Statistical Procedures for Agricultural Research. R Package Version 1.3-3*.

Edmands, S. (2002). Does parental divergence predict reproductive compatibility? *Trends Ecol. Evol.* 17, 520–527. doi: 10.1016/s0169-5347(02)02585-5

Endelman, J. B., and Jannink, J. L. (2012). Shrinkage estimation of the realized relationship matrix. *G3* 2, 1405–1413. doi: 10.1534/g3.112.004259

Estaghvirou, S. B. O., Ogutu, J. O., Schulz-Streeck, T., Knaak, C., Ouzunova, M., Gordillo, A., et al. (2013). Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. *BMC Genomics* 14:860. doi: 10.1186/1471-2164-14-860

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction into Quantitative Genetics*. Essex. Prentice Hall.

Faux, A. M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., et al. (2016). AlphaSim: software for breeding program simulation. *Plant Genome* 9:Plantgenome2016.02.0013.

Feng, F., Li, Y., Qin, X., Liao, Y., and Siddique, K. H. (2017). Changes in rice grain quality of Indica and Japonica type varieties released in China from 2000 to 2014. *Front. Plant Sci.* 8:1863. doi: 10.3389/fpls.2017.01863

Gaynor, R. C., Gorjanc, G., and Hickey, J. M. (2020). AlphaSimR: an R-package for breeding program simulations. *bioRxiv [Preprint]* doi: 10.1101/2020.08.10.245167

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9:e90346. doi: 10.1371/journal.pone.0090346

Hochholdinger, F., and Baldauf, J. A. (2018). Heterosis in plants. *Curr. Biol.* 28, R1089–R1092.

Holland, J. B., Nyquist, W. E., and Cervantes-Martínez, C. T. (2003). Estimating and interpreting heritability for plant breeding: an update. *Plant Breed. Rev.* 22, 9–112. doi: 10.1002/9780470650202.ch2

IRRI (2014). *Biometrics and Breeding Informatics, PBGB Division*. Los Baños: International Rice Research Institute.

Janaiah, A., and Xie, F. (2010). *Hybrid rice adoption in India: farm-level impacts and challenges. IRRI Technical Bulletin*. Los Baños, Philippines: International Rice Research Institute, 20.

Ji, Q., Lu, J., Chao, Q., Gu, M., and Xu, M. (2005). Delimiting a rice wide-compatibility gene S 5 n to a 50 kb region. *Theor. Appl. Genet.* 111, 1495–1503. doi: 10.1007/s00122-005-0078-0

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4. doi: 10.1186/1939-8433-6-4

Kwon, S. J., Ha, W. G., Hwang, H. G., Yang, S. J., Choi, H. C., Moon, H. P., et al. (2002). Relationship between heterosis and genetic divergence in 'Tongil'-type rice. *Plant Breed.* 121, 487–492. doi: 10.1046/j.1439-0523.2002.00760.x

Liang, Z., Gupta, S. K., Yeh, C. T., Zhang, Y., Ngu, D. W., Kumar, R., et al. (2018). Phenotypic data from inbred parents can improve genomic prediction in pearl millet hybrids. *G3* 8, 2513–2522. doi: 10.1534/g3.118.200242

Longin, C. F. H., Mühleisen, J., Maurer, H. P., Zhang, H., Gowda, M., and Reif, J. C. (2012). Hybrid breeding in autogamous cereals. *Theor. Appl. Genet.* 125, 1087–1096. doi: 10.1007/s00122-012-1967-7

Lu, Z. M., and Xu, B. Q. (2010). On significance of heterotic group theory in hybrid rice breeding. *Rice Sci.* 17, 94–98. doi: 10.1016/s1672-6308(08)60110-9

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*, Vol. 1. Sunderland, MA: Sinauer, 535–557.

Melchinger, A. E. (1999). "Genetic diversity and heterosis," in *Genetics and Exploitation of Heterosis in Crops*, eds J. G. Coors and S. Pandey (Madison: American Society of Agronomy, Inc.), 99–118. doi: 10.2134/1999.geneticsandexploitation.c10

Mottaleb, K. A., Mohanty, S., and Nelson, A. (2015). Factors influencing hybrid rice adoption: a Bangladesh case. *Aust. J. Agric. Resour. Econ.* 59, 258–274. doi: 10.1111/1467-8489.12060

Mrode, R. A. (2014). *Linear Models for the Prediction of Animal Breeding Values*. Wallingford: CABI.

Nalley, L., Tack, J., Barkley, A., Jagadish, K., and Brye, K. (2016). Quantifying the agronomic and economic performance of hybrid and conventional rice varieties. *Agron. J.* 108, 1514–1523. doi: 10.2134/agronj2015.0526

Nalley, L., Tack, J., Durand, A., Thoma, G., Tsiboe, F., Shew, A., et al. (2017). The production, consumption, and environmental impacts of rice hybridization in the United States. *Agron. J.* 109, 193–203. doi: 10.2134/agronj2016.05.0281

Ott, A., Liu, S., Schnable, J. C., Yeh, C. T., Wang, C., and Schnable, P. S. (2017). Tunable genotyping-by-sequencing (tGBS§) enables reliable genotyping of heterozygous loci. *bioRxiv [Perpint]* doi: 10.1101/100461

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Peng, S., Yang, J., Garcia, F. V., Laza, M. R. C., Visperas, R. M., Sanico, A. L., et al. (1998). "Physiology-based crop management for yield maximization of hybrid rice," in *Advances in Hybrid Rice Technology*, eds S. S. Virmani, E. A. Siddiq, and K. Muralidaran (Los Baños: IRRI), 157–176.

Powell, O. M., Gaynor, R. C., Gorjanc, G. M., Werner, C. R., and Hickey, J. M. (2020). A two-part strategy using genomic selection in hybrid crop breeding programs. *bioRxiv [Preprint]* doi: 10.1101/2020.05.24.113258

Rembe, M., Zhao, Y., Jiang, Y., and Reif, J. C. (2019). Reciprocal recurrent genomic selection: an attractive tool to leverage hybrid wheat breeding. *Theor. Appl. Genet.* 132, 687–698. doi: 10.1007/s00122-018-3244-x

Resende, M. F. R., Munoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., et al. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics* 190, 1503–1510. doi: 10.1534/genetics.111.137026

Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L., et al. (2015). Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. *Plant Genome* 8, 1–10.

SAMtools. (2018). *BCFtools*. https://github.com/samtools/bcftools/releases/tag/1.7.

Spielman, D. J., Kolady, D. E., and Ward, P. S. (2013). The prospects for hybrid rice in India. *Food Secur.* 5, 651–665. doi: 10.1007/s12571-013-0291-7

Sprague, G. F., and Tatum, L. A. (1942). General vs. specific combining ability in single crosses of corn 1. *Agron. J.* 34, 923–932. doi: 10.2134/agronj1942.00021962003400100008x

Tang, H., Xie, Y., Liu, Y. G., and Chen, L. (2017). Advances in understanding the molecular mechanisms of cytoplasmic male sterility and restoration in rice. *Plant Reprod.* 30, 179–184. doi: 10.1007/s00497-017-0308-z

Technow, F., Schrag, T. A., Schipprack, W., Bauer, E., Simianer, H., and Melchinger, A. E. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* 197, 1343–1355. doi: 10.1534/genetics.114.165860

Tracy, W. F., and Chandler, M. A. (2006). "The historical and biological basis of the concept of heterotic patterns in corn belt dent maize," in *Plant Breeding: The Arnel R Hallauer International Symposium*, eds K. R. Lamkey and M. Lee (Ames, IA: Blackwell Publishing), 219–233. doi: 10.1002/9780470752708.ch16

Virmani, S. S., and Wan, B. H. (1988). "Development of CMS lines in hybrid rice breeding," in *Hybrid Rice*, ed. IRRI (Manila: International Rice Research Institute), 103–114.

Waters, D. L., Subbaiyan, G. K., Mani, E., Singh, S., Vaddadi, S., Baten, A., et al. (2015). Genome wide polymorphisms and yield heterosis in rice (*Oryza sativa* subsp. indica). *Trop. Plant Biol.* 8, 117–125. doi: 10.1007/s12042-015-9156-x

Wu, T. D., Reeder, J., Lawrence, M., Becker, G., and Brauer, M. J. (2016). "GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality," In *Statistical Genomics* E. Mathé, and S. Davis (New York, NY: Humana Press), 283–334. doi: 10.1007/978-1-4939-3578-9_15

Xiang, T., Christensen, O. F., Vitezica, Z. G., and Legarra, A. (2016). Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. *Genet. Sel. Evol.* 48, 1–14.

Xiao, J., Li, J., Yuan, L., McCouch, S. R., and Tanksley, S. D. (1996). Genetic diversity and its relationship to hybrid performance and heterosis in rice as revealed by PCR-based markers. *Theor. Appl. Genet.* 92, 637–643. doi: 10.1007/s001220050173

Yu, S., Ali, J., Zhang, C., Li, Z., and Zhang, Q. (2020). Genomic breeding of green super rice varieties and their deployment in Asia and Africa. *Theor. Appl. Genet.* 133, 1427–1442. doi: 10.1007/s00122-019-03516-9

Yuan, L., Virmani, S. S., and Changxiong, M. (1989). "Hybrid rice: achievements and outlook," in *Proceedings of the International Rice Research Conference*, (Hangzhou: IRRI),

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 July 2021
doi: 10.3389/fgene.2021.689319

An Assessment of the Factors Influencing the Prediction Accuracy of Genomic Prediction Models Across Multiple Environments

Sarah Widener[1], George Graef[2], Alexander E. Lipka[1] and Diego Jarquin[2]**

[1] Department of Crop Sciences, University of Illinois, Urbana, IL, United States, [2] Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE, United States

The effects of climate change create formidable challenges for breeders striving to produce sufficient food quantities in rapidly changing environments. It is therefore critical to investigate the ability of multi-environment genomic prediction (GP) models to predict genomic estimated breeding values (GEBVs) in extreme environments. Exploration of the impact of training set composition on the accuracy of such GEBVs is also essential. Accordingly, we examined the influence of the number of training environments and the use of environmental covariates (ECs) in GS models on four subsets of n = 500 lines of the soybean nested association mapping (SoyNAM) panel grown in nine environments in the US-North Central Region. The ensuing analyses provided insights into the influence of both of these factors for predicting grain yield in the most and the least extreme of these environments. We found that only a subset of the available environments was needed to obtain the highest observed prediction accuracies. The inclusion of ECs in the GP model did not substantially increase prediction accuracies relative to competing models, and instead more often resulted in negative prediction accuracies. Combined with the overall low prediction accuracies for grain yield in the most extreme environment, our findings highlight weaknesses in current GP approaches for prediction in extreme environments, and point to specific areas on which to focus future research efforts.

Keywords: genotype-by-environment (GE) interaction, soybean nested association mapping (SoyNAM) populations, genomic selection (GS), extreme environmental conditions, environmental covariates (ECs)

OPEN ACCESS

Edited by:
Michelle Lacey,
Tulane University, United States

Reviewed by:
Just Jensen,
Aarhus University, Denmark
Julio Isidro Sanchez,
Universidad Politécnica de Madrid, Spain

****Correspondence:***
Alexander E. Lipka
alipka@illinois.edu
Diego Jarquin
diego.jarquin@gmail.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *31 March 2021*
Accepted: *07 June 2021*
Published: *23 July 2021*

Citation:
Widener S, Graef G, Lipka AE and Jarquin D (2021) An Assessment of the Factors Influencing the Prediction Accuracy of Genomic Prediction Models Across Multiple Environments. Front. Genet. 12:689319. doi: 10.3389/fgene.2021.689319

INTRODUCTION

The impacts of climate change are adversely affecting the availability of food, feed, fuel, and fiber security worldwide, with prior research suggesting a crop yield loss of 5% for each degree Celsius above historically observed weather patterns (Nelson et al., 2010; Zhao et al., 2017). Accelerated climate change has already been observed in specific regions that have low food security, which in turn could exacerbate crises in areas of the world that already struggle with a lack of available and affordable food (Whitford et al., 2013). It is therefore critical that research efforts focus on refining breeding tools so that the overall genetic gain of crops that humanity relies on continues to increase,

even in the face of extreme and fluctuating environments. Such work in breeding for optimal crop varieties are essential because agricultural efficiency in use of land and inputs are maximised whenever growers select the best crop variety for their environment, whereas varieties ill-suited to their environment will be more susceptible to disease, pest, and weather events (Zhao et al., 2017).

Genomic prediction (GP) is an emergent methodology that revolutionised plant and animal breeding, and is grounded in a statistical framework that uses genome-wide markers to predict breeding values of agronomically important traits (Bernardo, 1994; Meuwissen et al., 2001). Bernardo (1994) was the first who proposed the use of genomic information as covariates for predicting untested genotypes. Later on, Meuwissen et al. (2001) proposed a new methodology to cope with the challenge of fitting prediction models when the number of genomic covariates (p), delivered with the advancements of sequencing technologies, surpass by far the number data points (n) available to fit models ($p \gg n$).

A typical breeding program using GP begins with model training in which individual plants, grouped in a training population, are genotyped and phenotyped for the trait(s) of interest (Heffner et al., 2009). These training data are then used to fit a prediction model that quantifies the relationship between the p genotyped markers and phenotypic traits. This fitted model exclusively uses genotypic information collected from a breeding population to predict genomic estimated breeding values (GEBVs) of un-phenotyped genotypes, leveraging the genomic relationships between individuals in testing and training sets. The main application of this fitted GP model is to find which individuals have optimal GEBVs. Arguably, the most important advantages of GP are that it allows breeders (1) to determine which varieties should be discarded (screening), (2) to identify superior individuals to advance, and (3) to select best parents with desirable characteristics to be used in the next improvement cycles. In this way, GP has been shown to increase the genetic gain per field season compared to marker-assisted selection approaches that rely on phenotypic selection (Heffner et al., 2010).

One challenge that GP has already been shown to be well-suited for is the prediction of GEBVs across multiple environments (Burgueño et al., 2012; Heslot et al., 2014; Jarquín et al., 2014; Lopez-Cruz et al., 2015). To accurately make such predictions, GP models are typically augmented with additional terms to account for variability attributable to environments and their interaction with the genotype. These augmented GP models take on two main forms, specifically naïve or non-informed and informed. The first, naïve or non-informed, is to include a main random effect for the environment, as well as a two-way interaction effect between each marker genotype and each environment. This so-called G × E model has been shown to improve prediction accuracies (Jarquín et al., 2014; Lopez-Cruz et al., 2015) relative to conventional GP models that only include genotype and environment main effects. The second approach (informed) takes into account environmental covariates (ECs) measured at each environment, and then uses kernel-based methods to incorporate such information via the variance-covariance structures, which ultimately account for the interaction between environmental factors and marker genotypes. The resulting model (called the G × W model) incorporates quantifications of the interactions between each marker genotype and each EC into the prediction of GEBVs, and it could potentially outperform the naive G × E models (Jarquín et al., 2014; Basnet et al., 2019).

Given the promising prediction accuracies of the G × E and G × W models reported in these previous studies (Basnet et al., 2019), it is critical that their potential to predict GEBVs in extreme environments are explored. If these two models end up yielding a low or similar prediction accuracies under extreme environmental conditions, then future research will need to focus on either refining these GP models, exploring the genetic and environmental diversity required to yield decent prediction accuracies, or both. Therefore, the purpose of this study was to explore the impact of training set composition on the ability of the G × E and G × W models to accurately predict GEBVs in an extreme environment. The resulting analysis was conducted using a subset of the publicly available genotypic, phenotypic, and EC data from the soybean nested association mapping (SoyNAM) panel (Song et al., 2017; Diers et al., 2018) collected across multiple years and locations across the US-North Central Region. We used the phenotypic and EC data available at the nine resulting environments to determine which of the nine resulting environments were most and least similar among them. We then explored which subsets of environments yielded the highest prediction accuracies in these two targeted environments. Our working hypothesis was that the currently available genotypic, phenotypic, and EC data were insufficient for enabling the G × E and G × W GP models to accurately predict GEBVs in extreme environments. Thus, we predicted that these two GP models would provide inaccurate GEBVs at the most different of the nine environments that we considered in this study.

MATERIALS AND METHODS

The SoyNAM panel has been previously described (Song et al., 2017). Briefly, this panel consists of 40 recombinant inbred line (RIL) families derived from crossing a diverse parent to a common parent (IA3023). On average, each family consists of approximately 140 RILs, resulting in a total sample size of 5,600 individuals. To conduct our analysis, we considered a total of 5,000 markers that were genotyped from 17 lines that are elite public germplasm; 15 have diverse ancestry and 8 are plant introductions (Xavier et al., 2015). Genotypic and phenotypic data for the SoyNAM are publicly available at https://www.soybase.org/SoyNAM. These markers were then filtered to remove all markers that contain more than 50% of missing values and a minor allele frequency smaller than 0.03, resulting in a total of 4,450 markers being used for all downstream analyses.

Phenotypic Data and Field Trials

The phenotypic data were collected across 10 different locations in the US-North Central Region over 3 years. The trait that we analysed was grain yield (kg ha^{-1}), which has been

previously described (Hunter et al., 2017). The experimental design at each of the resulting environments have already been presented in Diers et al. (2018) and Xavier et al. (2018). However, not all locations were observed in all years, which resulted in a total of 18 location × year combinations (environments) (Xavier et al., 2018). Of these 18 environments, we analysed only a subset of 9 environments for which (1) we were able to obtain weather information and (2) have a common set of overlapping genotypes. Thus, a total of 2,336 genotypes were observed across all 9 environments. At each of the 9 environments, best linear unbiased predictions (BLUPs) grain yield, which already have been presented in Diers et al. (2018) and Xavier et al. (2018), were used in our analyses. To ensure the most and the least similar environments based on weather data also were the most and the least similar environments based on phenotypic data, random samples of 500 individuals were selected and mean phenotypic correlation between environments was computed until these matched. For a given environment, the mean phenotypic correlation is defined as the mean Pearson correlation of grain yield and the grain yield at the remaining environments. Thus, four random samples were considered for this study, where the difference of the mean phenotypic Pearson correlation between the most and the least correlated environment ranged from 0.185, 0.190, 0.191 to 0.198. Within each environment, heritabilities for grain yield were estimated as the ratio between the variability explained by the genetic component and the total variance $\hat{H}^2 = \frac{\widehat{\sigma}_L^2}{\widehat{\sigma}_L^2+\widehat{\sigma}_E^2}$, where $\widehat{\sigma}_L^2$ and $\widehat{\sigma}_E^2$ are, respectively, the variance component estimates of a line random effect and residual random effect fitted from a mixed linear model with grain yield as the response variable and lines included as a random effects (please see Holland et al., 2003) for an overview of calculating heritability).

Weather Data

At each of the 9 environments, we obtained ECs in the form of historical weather data extrapolated from Google Cloud.[1] These data were from weather stations distanced at most 57 km from the field location. After downloading the data from the cloud using a custom R script (Available from GitHub[2]), we selected three ECs that were both common to all 9 weather stations and recorded in 24-hr increments. Specifically, these three ECs were mean minimum daily temperature (measured in tenths of degrees Celsius), mean maximum daily temperature (tenths of degrees Celsius), and mean daily precipitation (inches). We chose to not convert mean daily precipitation to SI units because we wanted to leave the historical weather data from Google Cloud unaltered. For each location, weather data were collected starting on the planting date and continued until the 125th day after planting. Thus, the total number of ECs totaled 3 × 125 = 375, for a total of 9 × 375 = 3,375 total weather records across all 9 locations. A total of six weather records were missing; these values were imputed with the mean value between the previous and the following day within the same environment.

[1]https://cloud.google.com/public-datasets/weather

[2]https://github.com/sarahwidener/Frontiers_Paper/blob/master/weather_retrieval/Get%20all%20weather%20data%20for%20project.R

Statistical Analyses Conducted on ECs to Quantify Similarity Across Environments

At each environment, we assessed the distribution of the values of each EC on the first day of planting and the following 125 days. Additionally, we conducted a principal component analysis (Morrison et al., 1976) of all 375 ECs (3 ECs measured across 125 days) to explore their degree of similarity across the 9 environments. These analyses enabled the identification of which environments were most and least similar among them.

GS Models

We considered three genomic selection models (M1–M3) in our analysis; however, first we introduce the most elemental linear predictor (M0) because it is useful for deriving the other models.

M0: E + L. Consider that the yield performance y_{ij} of the ith (i = 1, 2,..., 500) genotype observed in the jth (j = 1, 2,..., 9) environment can be represented as a sum of a common constant effect across genotypes and environments (μ) plus a line effect L_i, an environmental effect E_j and an error term e_{ij} addressing the non-explained phenotypic variability as follows in M0:

$$y_{ij} = \mu + E_j + L_i + e_{ij} \quad (0)$$

where E_j and L_i are considered random terms such that these are assumed to be independent and identically distributed (IID) outcomes from a normal density such that $E_j \sim N\left(0, \sigma_E^2\right)$ and $L_j \sim N\left(0, \sigma_L^2\right)$, with σ_E^2 and σ_L^2 being the corresponding variance components; and $e_{ij} \sim N\left(0, \sigma^2\right)$ with σ^2 representing the residual variance. One disadvantage of M0 is that it does not allow the prediction of unobserved genotypes because it relies only on phenotypic information.

M1: E + G. To allow the prediction of untested genotypes, genomic relationships between individuals in training and testing sets should be established first. For this, we construct a covariance structure whose entries contain the genomic similarities between pairs of individuals. Assuming that the marker effects b_k in the linear combination involving p markers, $g_i = \sum_{k=1}^{p} x_{ik} b_k$, follows IID normal densities $N(0, \sigma_b^2)$ and using results from the multivariate normal density we have that that the vector of genomic effects $\mathbf{g} = \{g_i\}$ follows a multivariate normal distribution such that $\mathbf{g} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_g^2\right)$, where $\boldsymbol{G} = \frac{\mathbf{X}\mathbf{X}'}{p}$, $\boldsymbol{X}$ is the centered and scaled (by columns) matrix of molecular markers, and $\sigma_g^2 = p \times \sigma_b^2$. Thus, we have the following linear predictor for M1:

$$y_{ij} = \mu + E_j + g_i + e_{ij}, \quad (1)$$

where all terms are as previously described. One of the disadvantages of M1 is that it returns the same genomic

effect across environments; thus the direct influence of stimuli unique to particular environments are not taken into consideration.

M2: E + G + G × E. To allow estimations of particular genomic effects within environments, Jarquín et al. (2014) proposed a model that conceptually considered the interaction between each molecular marker and each environment. This model is based on the cell-by-cell product between two covariance structures, one for environmental factors using only the identification of the environments and another for genotypes based on the genomic relationship matrix. Thus, the genotype-by-environment interaction effects can be predicted thought $\mathbf{gE} = \left\{gE_{ij}\right\}$ with $\mathbf{gE} \sim N\left(\mathbf{0}, Z_E Z_E^{'} \# Z_g \mathbf{G} Z_g^{'} \sigma^2_{gE}\right)$ where Z_E and Z_g are the corresponding incidence matrices that connect phenotypic observations with environments and genotypes; σ^2_{gE} is the corresponding variance component; and # represents the Hadamard (cell-by-cell) product between two matrices. Hence, we have that the resulting linear predictor for M2 can be written as follows:

$$y_{ij} = \mu + E_j + g_i + gE_{ij} + e_{ij}, \quad (2)$$

where all terms are as previously described. This model not only allows the inclusion of the G × E interaction in a naïve way but potentially also offers the opportunity of including the genotype-by-environment interaction component in an informed way. One approach for accomplishing this is to include ECs that describe environmental similarities between pairs of environments. Such information is incorporated into the final GP model we consider, as described below.

M3: E + G + G × W. Analogous to the derivation of the kinship matrix **G**, the information on ECs can be considered in the development of an environmental kinship matrix **Ω** describing environmental similarities between pairs of environments. Jarquín et al. (2014) proposed a model that allows the incorporation of the ECs to interact with molecular markers. To accomplish this, it is necessary to first model the main effect of the ECs. Consider that the environmental effect w_j corresponding to j^{th} environment can be written as a linear combination between q ECs and their corresponding effects $w_j = \sum_{l=1}^{q} W_{jl}\gamma_l$ with $\gamma_l \sim N\left(0, \sigma^2_W\right)$ and σ^2_W defined as the corresponding variance component. Then we have that the vector of environmental effects follows a multivariate normal density such that $\mathbf{w} = \left\{w_j\right\} \sim N\left(0, \Omega\sigma^2_\Omega\right)$; where $\Omega = \frac{WW^{'}}{q}$, W is the centered and scaled (by columns) matrix of ECs (i.e., measurements of mean minimum daily temperature, mean maximum daily temperature, and mean daily precipitation across 125 days, as previously described), $\sigma^2_\Omega = q\sigma^2_W$ is the corresponding variance component. To include the main effect of the ECs in the prediction model, we have to expand **Ω** using the incidence matrix that connects phenotypes with environments such as $Z_E \Omega Z_E^{'}$ is the resulting covariance structure. In order to include the ECs in interaction with marker genotypes, we substitute the expanded covariance matrix in the covariance structure of the **gE** term such as $\mathbf{gw} \sim N\left(\mathbf{0}, Z_E \Omega Z_E^{'} \# Z_g \mathbf{G} Z_g^{'} \sigma^2_{g\Omega}\right)$ with $\sigma^2_{g\Omega}$ acting as the corresponding variance component. The resulting linear predictor for M3 can be written as follows:

$$y_{ij} = \mu + E_j + g_i + gw_{ij} + e_{ij}, \quad (3)$$

where all terms are as previously described. Conceptually, this model allows the inclusion of the interaction between each molecular marker and each ECs.

Cross-Validation Scheme

The main objective of this cross-validation scheme was to identify the training environments and GP model that yielded the highest possible prediction accuracies in (1) the environment that had the lowest mean phenotypic correlation with the other eight environments, and then to contrast this result with (2) the environment that had the highest mean phenotypic correlation with the other eight environments. For both (1) and (2), we conducted a CV00 cross validation scheme (Jarquín et al., 2017; Jarquin et al., 2020) where none of the genotypes from the test environment were used to train the GS model.

For a given random sample of 500 genotypes (i.e., RILs from the SoyNAM population) observed in all 9 environments, we randomly selected a set of 200 genotypes to be the testing set in the unobserved environment. We used the phenotypic information of the remaining 300 genotypes observed in the remaining 8 environments. Because we were interested in the impact of training set composition on prediction accuracies in a given test environment, we evaluated the ability of all possible subsets of the remaining eight remaining environments to train each GP model and accurately predict GEBVs. The resulting numbers of possible combinations of environments to include in the training set are described in **Table 1**. For a given test environment, training set, and GP model, prediction accuracy was measured as the Pearson correlation between the observed (phenotypic) and predicted (GEBV) values. This procedure was repeated three additional times so that the performance of the GP models could be evaluated across all 4 random subsets of 500 genotypes.

TABLE 1 | The number of possible combinations (right column) of subsets of eight environments (left column) considered as a training sets for the genomic selection models.

Subset of environments	Number of combinations
1	8
2	28
3	56
4	70
5	56
6	28
7	8
8	1

In summary, the value of the right column of the i^{th} row is $\binom{8}{i}$.

RESULTS

Similar Distribution of EC Values Across Nine Environments, With IA_2013 Displaying the Most Unique EC Values

A biplot of the first two principal components of 3 × 125 = 375 ECs suggests that many of the locations have similar environmental conditions (**Figure 1A**). This result is supported by similar distributions of values of the three ECs (across the 125 days since planting) within each of the 9 locations (**Figures 1B–D**). Collectively, these results suggest that there is not a substantial amount of environmental diversity among the 9 environments that were tested. Nevertheless, among these 9 environments, IL_2013, IN_2013, and IA_2013 appeared to be the most divergent.

Phenotypic Data on Grain Yield Were Most Unique Within IA_2013, While Grain Yield Within NE_2011 Was Most Similar to the Remaining Environments

Across the 4 replicate random samples of 500 genotypes, we observed similar trends in phenotypic distributions of yield performance (ka ha^{-1}) across the 9 environments. For brevity, we focus on the results for the second random sample in the main text of the manuscript, and then provide similar details for the remaining three random samples in the **Supplementary Material** section. We observed that IA_2013, IA_2012, and IL_2011 were the environments that tended to yield the least, while IN_2013 and NE_2011 were the environments that yielded the greatest (**Figure 2**).

We then quantified the phenotypic correlation between environments to determine which were least and most similar. As such, the mean phenotypic correlation of each environment with the remaining eight environments is presented in **Table 2**. The two environments that showed the lowest and the highest mean correlation with the remaining eight environments were IA_2013 (0.137) and NE_2011 (0.327), respectively (as depicted under the column labeled "Rep 2" under "Average Correlation" in **Table 2**). Across the four replicates, we also calculated the heritabilities at each of the environments. These heritabilities were relatively stable across the 9 tested environments with around 50% of the phenotypic variability explained by the additive genetic component within each environment (**Table 2**). Based on the collective information on trait correlations across the 9 environments and the ECs, we determined that IA_2013 was the most extreme environment and that NE_2011 was the least extreme environment.

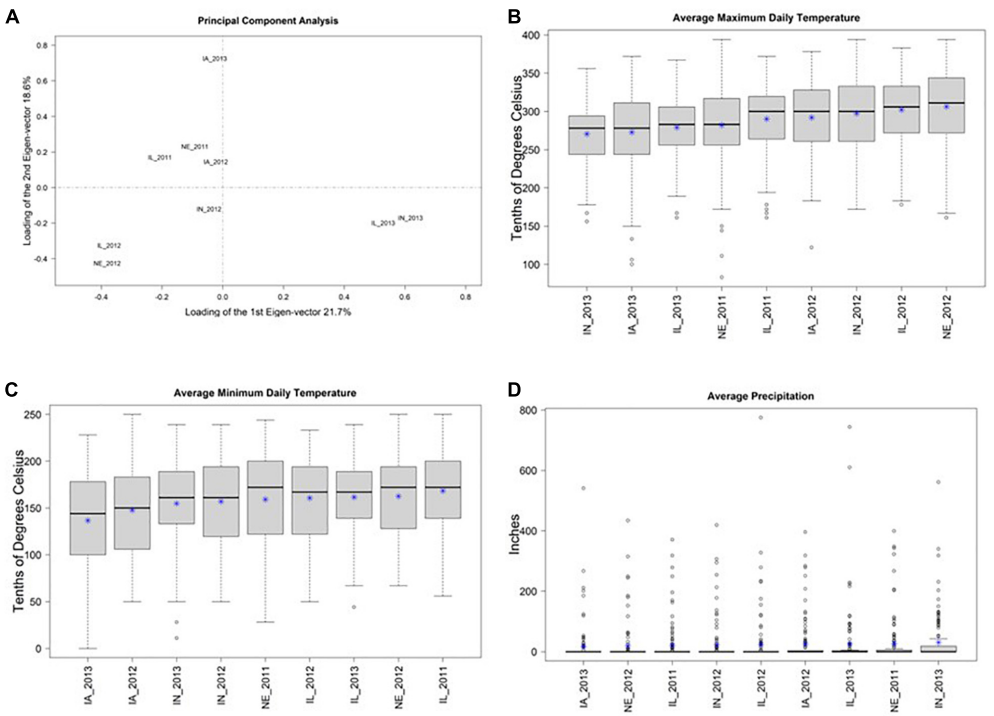

FIGURE 1 | (A) Shows the first two principal components (PCs) from a principal component analysis of three environmental covariates (ECs) measured over 125 days at nine environments in the US North Central Region. The *X*-axis shows PC1 and the *Y*-axis shows PC2. The percent of total variation explained in each PC is provided in the axis labels. This plot suggests that IN_2013, IL_2013, and IA_2013 are the environments with EC values that are most distinct from the remaining environments. **(B–D)** Display the distributions of mean daily values of three environmental covariates (*Y*-axis) at these nine environments (*X*-axis) and * represents the mean. **(B,C)** Show maximum and minimum daily temperature, respectively, within each environment in tenths of degrees Celsius. **(D)** Shows precipitation in inches. Collectively **(B–D)** suggest that there the observed EC values are similar among these 9 environments.

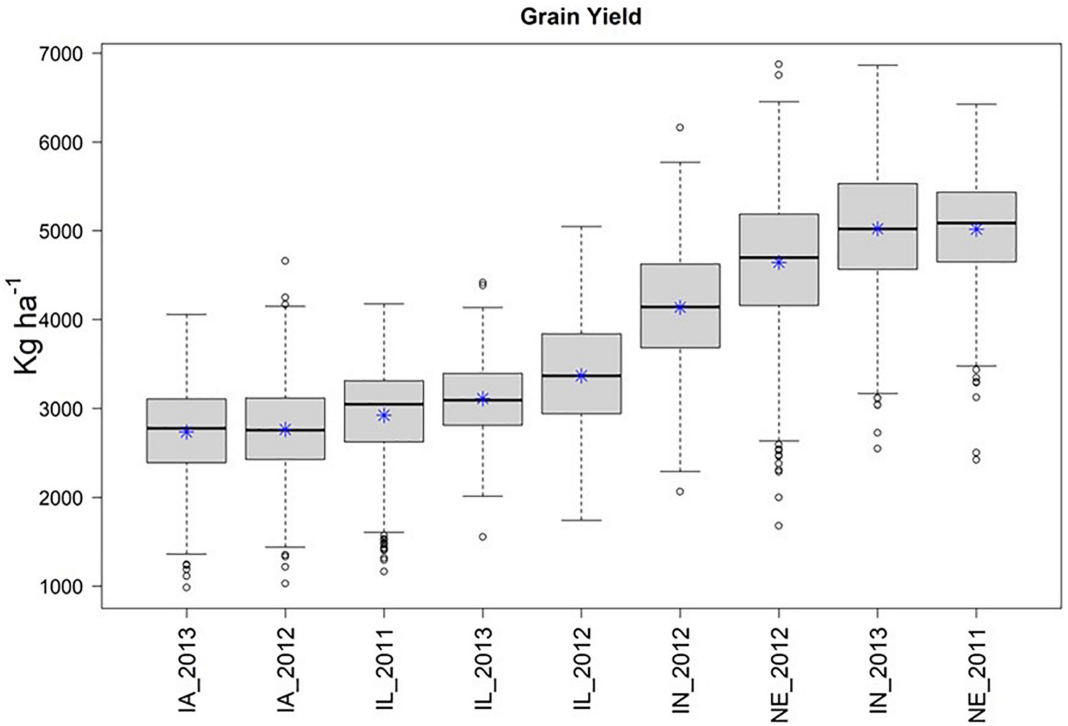

FIGURE 2 | Boxplot of yield in kg ha^{-}1 (*Y*-axis), by environment (*X*-axis) for the second random sample of 500 genotypes from the SoyNAM panel, and * represents the sampling mean. Environments IA_2013, IA_2012, and IL_2011 had the lowest yield, while IN_2013 and NE_2011 had the highest yield.

TABLE 2 | Mean Pearson correlation coefficient of grain yield (in kg ha^{-}1) between each environment and the remaining eight environments (presented under the columns labeled "Average Correlation"), as well as the observed heritability of grain yield within each environment (presented under the columns labeled "Heritability").

Environment	Average Correlation				Heritability			
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 4
IA_2013	0.164	0.137	0.150	0.158	0.503	0.485	0.492	0.497
IA_2012	0.268	0.216	0.228	0.254	0.511	0.499	0.483	0.496
IL_2011	0.221	0.203	0.179	0.177	0.500	0.521	0.486	0.514
IL_2013	0.277	0.277	0.257	0.291	0.509	0.510	0.507	0.482
IL_2012	0.263	0.261	0.236	0.247	0.513	0.481	0.497	0.504
IN_2012	0.290	0.277	0.262	0.287	0.507	0.495	0.500	0.483
NE_2012	0.293	0.245	0.235	0.266	0.506	0.484	0.478	0.481
IN_2013	0.289	0.283	0.301	0.276	0.517	0.534	0.502	0.467
NE_2011	0.349	0.327	0.340	0.356	0.502	0.500	0.510	0.499

The columns labeled "Rep 1", . . . , "Rep 4" present these summary statistics for each of the 4 random samples of 500 randomly selected individuals.

Relatively Small Number of Environments Needed to Yield Accurate Predictions for IA_2013

We evaluated the ability of M1–M3 to predict GEBVs in the most extreme environment, IA_2013, using the all-possible subsets of the 8 remaining environments, as described in the Section "Materials and Methods" and **Table 1**. **Figure 3** presents the correlation between the predicted and observed values for IA_2013 considering the 255 different ways for combining the remaining 8 environments for model calibration across the 4 replicates of 500 randomly selected genotypes. In general, low and sometimes negative prediction accuracies were observed, with the highest observed prediction accuracy being less than 0.36. The optimal number of training environments (i.e., that yielded the highest prediction accuracies from M1, M2, and M3) changed considerably across the four replicates, but we frequently observed that a relatively small number of environments was needed to achieve the highest possible prediction accuracy. Across the 4 replicates of 500 random samples, we never observed an instance where the model accounting for ECs (i.e., M3) yielded definitively higher prediction accuracies than M1 or M2. Moreover, there were many combinations of training

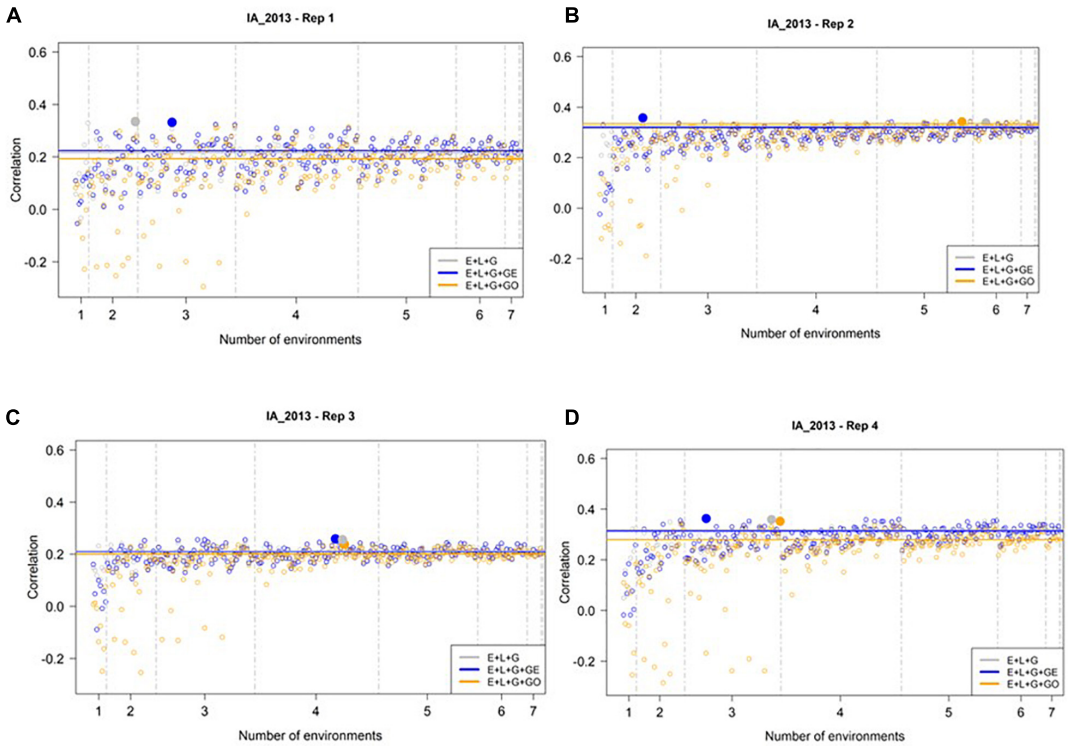

FIGURE 3 | Observed prediction accuracy of grain yield in kg ha^{-1} at IA_2013 across multiple genomic prediction (GP) models and training environments. Four random samples of 500 genotypes from the SoyNAM panel are presented in panels **(A–D)**. For each panel, the *X*-axis is the specific number of environments considered for training the GP model, sorted from smallest to largest number of training environments; and the *Y*-axis shows the prediction accuracy, quantified as the Pearson correlation coefficient between the observed phenotypic values and the genomic estimated breeding values. The results in grey depict the GP model without any genotype-by-environment (G × E) interaction effects, while the results in blue depict the GP model with G × E interaction effects, and finally the results in yellow depict the GP model with G × E interaction effects that incorporates environmental covariates (ECs). The highest observed prediction accuracies across any training set from each GP model are highlighted by a solid circle of the same color, while the prediction accuracies of the three models obtained using all eight of the possible environments in the training set are shown as horizontal lines of the same color. These panels show that not all eight environments are needed to obtain the maximum possible prediction accuracies.

environments where M3 clearly yielded lower, and often negative, prediction accuracies.

Slightly Larger Number of Training Environments Needed to Maximize Prediction Accuracy in NE_2011

We then conducted a similar analysis to assess the predictive ability of M1–M3 to predict GEBVs in the least extreme environment (NE_2011, see **Figure 4**). In general, we observed higher prediction accuracies at NE_2011 relative to those observed in the most extreme environment (IA_2013). Similar to IA_2013, the number of optimal environments needed for M1, M2, and M3 differed across reps. However, the general trend we observed was that a larger number of training environments were needed for maximizing the prediction accuracy in NE_2011 relative to IA_2013. Finally, we did not observe any evidence that including ECs in the model improved prediction accuracy. That is, the highest prediction accuracy observed for M3 (∼0.53) was not noticeably different than those of M1 and M2, and the lowest prediction accuracies observed across the four replicates were from M3.

DISCUSSION

We compared the ability of various subsets of environments to accurately predict GEBVs in (1) a target environment that was the most different from the remaining environments with respect to phenotypic correlation and observed ECs, and (2) a target environment that was the most similar using these same metrics. Although we observed lower prediction accuracies in (1), the ensuing analysis highlighted similar trends in model performance for both (1) and (2). Using three different GS models that accounted for environmental information to varying degrees, we discovered that maximum prediction accuracies could be achieved by using only a subset of the 8 environments to train the GP models. Additionally, we found that the inclusion of ECs into GP models did not substantially boost the prediction accuracies of the target environments. Finally, when using a reduced number of environments to train the GP models, we occasionally observed extremely low and negative prediction accuracies when including ECs into the GP model. Thus, we identified potential areas of weakness in existing GP models when they are applied to predicting GEBVs in specific environments and underscored the

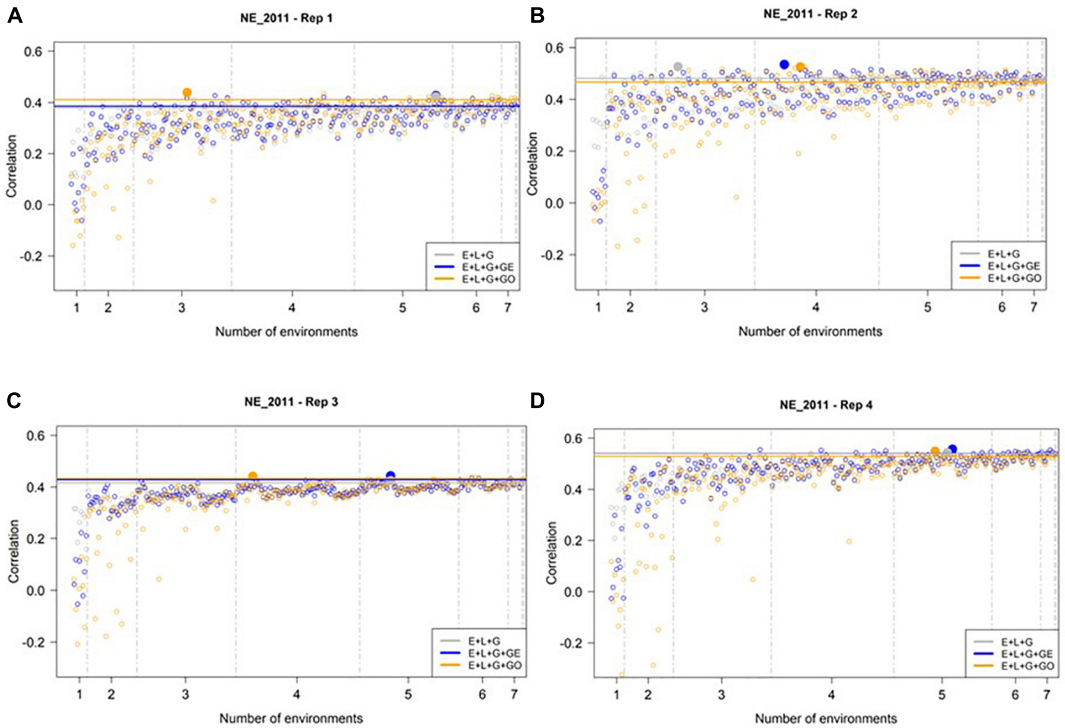

FIGURE 4 | Observed prediction accuracy of grain yield in kg ha^{-1} at NE_2011 across multiple genomic prediction (GP) models and training environments. Four random samples of 500 genotypes from the SoyNAM panel are presented in panels **(A–D)**. For each panel, the *X*-axis is the specific number of environments considered for training the GP model, sorted from smallest to largest number of training environments; and the *Y*-axis shows the prediction accuracy, quantified as the Pearson correlation coefficient between the observed phenotypic values and the genomic estimated breeding values. The results in grey depict the GP model without any genotype-by-environment (G × E) interaction effects, while the results in blue depict the GP model with G × E interaction effects, and finally the results in yellow depict the GP model with G × E interaction effects that incorporates environmental covariates (ECs). The highest observed prediction accuracies across any training set from each GP model are highlighted by a solid circle of the same color, while the prediction accuracies of the three models obtained using all eight of the possible environments in the training set are shown as horizontal lines of the same color. These panels show that not all eight environments are needed to obtain the maximum possible prediction accuracies.

critical need to explore which factors influence the development of training environments that can lead to the most accurate of such predictions.

The Inclusion of ECs Into the GP Model Did Not Result in Substantially Higher Prediction Accuracies

For the environment with the least similar phenotypic correlations and ECs relative to the remaining environments (IA_2013), we observed low prediction accuracies, as expected. These low accuracies indicate that there is room for improvement for developing approaches to predict GEBVs in extreme environments. Nevertheless, the trends that we observed in our analysis point to areas for further exploration and refinement. To illustrate this point, consider the results from the second random sample of 500 individuals we considered in this study. For this replicate of our analysis, the inclusion of the G × E interaction in the model without weather data (M2) returned the highest predictive ability (0.357). In this case, only two environments were needed (IL_2012, IN_2012) for model calibration, and the relative improvements were, respectively, 7, 12, and 10% relative to using all of the eight environments in the training set under M1, M2, and M3.

These results also identified important shortcomings of using ECs directly in the GP model. For instance, the fact that M3 occasionally yielded prediction accuracies that were lower than those of M1 and M2 suggests that the inclusion of ECs into the GP model is not guaranteed to increase the accuracy of GEBVs. This suggests that further research into the development of GP models that effectively incorporate these ECs is needed. Combined with the observation that M3 yielded negative prediction accuracies more often than M1 and M2, we also infer that further investigation similar to Gillberg et al. (2019) is needed into dissecting which EC values are most likely to contribute to the highest possible prediction accuracies. These two avenues for future research could ultimately facilitate the development robust statistical models for GP in this paradigm, as well as identification of the ideal environments and ECs to use to train these GP models.

We observed similar trends between the performance of the three GP models in most similar environment (NE_2011). In particular, we noted that the incorporation of such weather data to predict GEBVs in NE_2011 (i.e., through M3) often

resulted in accuracies that were either negative or worse than M1 and M2. Because we observed a higher and more stable prediction accuracies as the number of environments used in the training set increased (a trend that was also observed for IA_2013), we infer that the collective information from multiple similar environments is critical for accurate prediction GEBVs in targeted environments with similar weather characteristics.

Minimal Genotypic and Environmental Diversity Are Limiting Factors of This Study

There are several important shortcomings of this study. First, we limited our analysis to only one species. Given the relatively narrow genetic diversity of soybean (Hyten et al., 2006), our study potentially did not fully explore the full extent to which M1–M3 could robustly predict breeding values in species with more diverse germplasm. Although we would expect to observe low prediction accuracies for such scenarios (as suggested by the findings of, e.g., Lorenz and Smith, 2015), it would nevertheless be worthwhile to quantify these accuracies. Similarly, all 9 of the environments that we evaluated were from a relatively narrow geographical range in the midwestern United States. Even though we were able to observe differences in the prediction accuracy of the GP models between the two test environments (IA_2013 and NE_2011), it is critical that follow-up studies conduct the analysis presented in this work in data from a wider range of locations.

In general, the incorporation of ECs into GP in a manner analogous to the incorporation of genome-wide marker data is rapidly maturing into the field of enviromics (Resende et al., 2021), and the findings from this study and others (Alves et al., 2021) could be useful for the establishment of best practices for collecting and utilizing environmental data. For example, one notable constraint of our study was that the observed ECs were common for all genotypes within the same environment. Given the potential for significant differences in EC values within a field, we were unable to capture these potentially important sources of variability. Combined with our use of only three ECs that were common across the 9 environments, we postulate that the inclusion of more ECs, potentially with differing values within locations, will reveal how sensitive or insensitive the GP models are at predicting breeding values when used in cases of extreme environments.

CONCLUSION

Even with the relatively narrow scope of genomic and environmental diversity observed in our data, we identified notable weaknesses in both the current GP models and training data used to predict GEBVs in different environments. We observed that (1) most accurate GEBVs were from GP models trained on only a subset of the available environments, and (2) at best the inclusion of ECs into the GP model did not substantially improve the prediction accuracies of the GEBVs. Nevertheless, the fact that we observed such diversity in prediction accuracies across the possible combinations of training sets suggest that a substantial amount of research is needed to explore which properties of training sets are responsible for the highest prediction accuracies. Coupled with the generally low prediction accuracies for the most extreme environment, we ultimately conclude that dedicated future research endeavors are needed to make genomic prediction better suited for extreme environments.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/ **Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SW wrote the initial draft, collected weather data, and participated in the conceptualization of the study. GG edited the manuscript and provided guidance to better understand the implementation of GS in soybeans. AL edited the manuscript, provided oversight for the study, and participated in the conceptualization of the study. DJ edited the manuscript, conducted the prediction studies, and participated in the conceptualization of the study. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by USDA-NIFA Grant Number 2018-68005-27937. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.689319/full#supplementary-material

Supplementary Figure 1 | Boxplot of yield in kg ha^{-1} (*X* axis), by environment (*Y* axis) for the (first, third or fourth) random sample of 500 genotypes from the SoyNAM panel. Environments IA_2013, IA_2012, and IL_2011 had the lowest yield, while IN_2013 and NE_2011 had the highest yield.

Supplementary Figure 2 | Boxplot of yield in kg ha^{-1} (*X* axis), by environment (*Y* axis) for the (first, third or fourth) random sample of 500 genotypes from the SoyNAM panel. Environments IA_2013, IA_2012, and IL_2011 had the lowest yield, while IN_2013 and NE_2011 had the highest yield.

Supplementary Figure 3 | Boxplot of yield in kg ha^{-1} (*X* axis), by environment (*Y* axis) for the (first, third or fourth) random sample of 500 genotypes from the SoyNAM panel. Environments IA_2013, IA_2012, and IL_2011 had the lowest yield, while IN_2013 and NE_2011 had the highest yield.

REFERENCES

Alves, F. C., Galli, G., Matias, F. I., Vidotti, M. S., Morosini, J. S., and Fritsche-Neto, R. (2021). Impact of the complexity of genotype by environment and dominance modeling on the predictive accuracy of maize hybrids in multi-environment prediction models. *Euphytica* 217:37.

Basnet, B. R., Crossa, J., Dreisigacker, S., Pérez–Rodríguez, P., Manes, Y., Singh, R. P., et al. (2019). Hybrid wheat prediction using genomic, pedigree, and environmental covariables interaction models. *Plant Genome* 12:180051. doi: 10.3835/plantgenome2018.07.0051

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci.* 34, 20–25. doi: 10.2135/cropsci1994.0011183x003400010003x

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. *Crop Sci.*52, 707–719. doi: 10.2135/cropsci2011.06.0299

Diers, B. W., Specht, J., Rainey, K. M., Cregan, P., Song, Q., Ramasubramanian, V., et al. (2018). Genetic architecture of soybean yield and agronomic traits. *G3* 8, 3367–3375.

Gillberg, J., Marttinen, P., Mamitsuka, H., and Kaski, S. (2019). Modelling G× E with historical weather information improves genomic prediction in new environments. *Bioinformatics* 35, 4045–4052. doi: 10.1093/bioinformatics/btz197

Heffner, E. L., Lorenz, A. J., Jannink, J. L., and Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci.* 50, 1681–1690. doi: 10.2135/cropsci2009.11.0662

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.1007/978-3-319-63170-7_1

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127, 463–480. doi: 10.1007/s00122-013-2231-5

Holland, J. B., Nyquist, W. E., and Cervantes-Martínez, C. T. (2003). Estimating and interpreting heritability for plant breeding: an update. *Plant Breed. Rev.* 22, 9–112.

Hunter, M. C., Smith, R. G., Schipanski, M. E., Atwood, L. W., and Mortensen, D. A. (2017). Agriculture in 2050: recalibrating targets for sustainable intensification. *Bioscience* 67, 386–391. doi: 10.1093/biosci/bix010

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., et al. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. U.S.A.* 103, 16666–16671. doi: 10.1073/pnas.0604379103

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607. doi: 10.1007/s00122-013-2243-1

Jarquin, D., De Leon, N., Romay, M. C., Bohn, M. O., Buckler, E. S., Ciampitti, I., et al. (2020). Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. *Front. Genet.* 11:592769. doi: 10.3389/fgene.2020.592769

Jarquín, D., Lemes da Silva, C., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., et al. (2017). Increasing genomic-enabled prediction accuracy by modeling genotype× environment interactions in Kansas wheat. *Plant Genome* 10, 1–15. doi: 10.3835/plantgenome2016.12.0130

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J. L., et al. (2015). Increased prediction accuracy in wheat breeding trials using a marker× environment interaction genomic selection model. *G3* 5, 569–582. doi: 10.1534/g3.114.016097

Lorenz, A. J., and Smith, K. P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Sci.* 55, 2657–2667. doi: 10.2135/cropsci2014.12.0827

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Morrison, D. F., Marshall, L. C., and Sahlin, H. L. (1976). *Multivariate Statistical Methods* (New York, NY: McGraw-Hill)

Nelson, G. C., Rosegrant, M. W., Palazzo, A., Gray, I., Ingersoll, C., Robertson, R., et al. (2010). *Food Security, Farming, And Climate Change to 2050: Challenges to 2050 and Beyond*. Washington, DC: International Food Policy Research Institute.

Resende, R. T., Piepho, H.-P., Rosa, G. J., Silva-Junior, O. B., e Silva, F. F., de Resende, M. D. V., et al. (2021). Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theor. Appl. Genet.* 134, 95–112. doi: 10.1007/s00122-020-03684-z

Song, Q., Yan, L., Quigley, C., Jordan, B. D., Fickus, E., Schroeder, S., et al. (2017). Genetic characterization of the soybean nested association mapping population. *Plant Genome* 10:lantgenome2016.2010.0109.

Whitford, R., Fleury, D., Reif, J. C., Garcia, M., Okada, T., Korzun, V., et al. (2013). Hybrid breeding in wheat: technologies to improve hybrid wheat seed production. *J. Exp. Bot.* 64, 5411–5428. doi: 10.1093/jxb/ert333

Xavier, A., Beavis, W., Specht, J., Diers, B., Muir, W., Mian, R., et al. (2015). *SoyNAM: Soybean Nested Association Mapping Dataset. R package version 1.*

Xavier, A., Jarquin, D., Howard, R., Ramasubramanian, V., Specht, J. E., Graef, G. L., et al. (2018). Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3* 8, 519–529. doi: 10.1534/g3.117.300300

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., et al. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl. Acad. Sci. U.S.A.* 114, 9326–9331. doi: 10.1073/pnas.1701762114

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 September 2021
doi: 10.3389/fgene.2021.637133

Impact of Genomic Prediction Model, Selection Intensity, and Breeding Strategy on the Long-Term Genetic Gain and Genetic Erosion in Soybean Breeding

Éder David Borges da Silva[1], Alencar Xavier[2,3] and Marcos Ventura Faria[1]*

[1] Department of Agronomy, Universidade Estadual do Centro-Oeste, Guarapuava, Brazil, [2] Department of Biostatistics, Corteva Agriscience™, Johnston, IA, United States, [3] Department of Agronomy, Purdue University, West Lafayette, IN, United States

Genomic-assisted breeding has become an important tool in soybean breeding. However, the impact of different genomic selection (GS) approaches on short- and long-term gains is not well understood. Such gains are conditional on the breeding design and may vary with a combination of the prediction model, family size, selection strategies, and selection intensity. To address these open questions, we evaluated various scenarios through a simulated closed soybean breeding program over 200 breeding cycles. Genomic prediction was performed using genomic best linear unbiased prediction (GBLUP), Bayesian methods, and random forest, benchmarked against selection on phenotypic values, true breeding values (TBV), and random selection. Breeding strategies included selections within family (WF), across family (AF), and within pre-selected families (WPSF), with selection intensities of 2.5, 5.0, 7.5, and 10.0%. Selections were performed at the F4 generation, where individuals were phenotyped and genotyped with a 6K single nucleotide polymorphism (SNP) array. Initial genetic parameters for the simulation were estimated from the SoyNAM population. WF selections provided the most significant long-term genetic gains. GBLUP and Bayesian methods outperformed random forest and provided most of the genetic gains within the first 100 generations, being outperformed by phenotypic selection after generation 100. All methods provided similar performances under WPSF selections. A faster decay in genetic variance was observed when individuals were selected AF and WPSF, as 80% of the genetic variance was depleted within 28–58 cycles, whereas WF selections preserved the variance up to cycle 184. Surprisingly, the selection intensity had less impact on long-term gains than did the breeding strategies. The study supports that genetic gains can be optimized in the long term with specific combinations of prediction models, family size, selection strategies, and selection intensity. A combination of strategies may be necessary for balancing the short-, medium-, and long-term genetic gains in breeding programs while preserving the genetic variance.

Keywords: long-term gains, soybean breeding, genomic selections, selection intensity, genomic prediction

OPEN ACCESS

Edited by:
Waseem Hussain, International Rice Research Institute (IRRI), Philippines

Reviewed by:
Yongkang Kim, University of Colorado Boulder, United States
Nicholas B. Larson, Mayo Clinic, United States

****Correspondence:***
Éder David Borges da Silva
ederdbs@gmail.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *02 December 2020*
Accepted: *05 August 2021*
Published: *01 September 2021*

Citation:
Silva ÉDB da, Xavier A and Faria MV (2021) Impact of Genomic Prediction Model, Selection Intensity, and Breeding Strategy on the Long-Term Genetic Gain and Genetic Erosion in Soybean Breeding. Front. Genet. 12:637133. doi: 10.3389/fgene.2021.637133

INTRODUCTION

Soybean [*Glycine max* (L.)] is the most important source of protein for animal feed and an important source of oil for human consumption, biofuel, and other industrial applications. Soybeans are cultivated globally, and the largest producers include Brazil, United States, Argentina, Paraguay, and China (FAO, 2021). Soybeans are bred for several traits, but grain yield is considered as the most important.

Genome-wide prediction is a key tool in soybean breeding. It is utilized for faster and more accurate selection of superior individuals (Meuwissen et al., 2001). Methodologically, genomic models recreate the framework utilized for pedigree analysis, but using genomic relationships instead (VanRaden, 2008; Habier et al., 2011; VanRaden et al., 2011). Other factors that may have contributed to the increasing adoption of genomic selection (GS) in plants include the decreasing cost of genotyping and the availability of software tools and computing power to analyze large datasets.

Studies involving GS in plants have been mostly focused on prediction for advancement purposes, hence restricted to the evaluation of genetic gain within a single generation (Schmutz et al., 2010; Sonah et al., 2013; Jarquin et al., 2016; Xavier et al., 2016, 2018a,b; Diers et al., 2018; Smallwood et al., 2019). Studies of long-term gains based on GS are expensive and time-consuming; consequently, the literature is scarce (Wray and Goddard, 1994; Goddard, 2009; Yabe et al., 2016; Gorjanc et al., 2018; Allier et al., 2019a). In addition, evaluation with real data from breeding programs faces additional challenges, such as the ongoing changes in breeding pipelines driven by business decisions, changes in the genotyping technology, and annual changes in resources. Conversely, the deployment of simulations has become an instrumental decision tool in plant breeding. It enables the assessment of genetic gain under different scenarios. In part, the increasing popularity of simulations is due to the quantity and flexibility of software made available (Faux et al., 2016; Pook et al., 2019; Toledo et al., 2019). For instance, breeders are now capable of simulating entire breeding programs with the intent of tuning the breeding parameters to maximize genetic gains in the short and long term (Hickey et al., 2014; Gorjanc et al., 2018), along with the best allocation of resources for a given budget.

By assessing predictive models and contrasting selection strategies, this study envisioned analyzing the influence of a set of variables on long-term genetic gains based on a simulated soybean breeding program and providing insight into the best practices for optimizing genetic gains.

MATERIALS AND METHODS

Simulated Populational Parameters

The founder breeding population contained 200 individuals. Those were simulated based on the genomic parameters using the Markovian Coalescent Simulator (MaCS; Chen et al., 2009), which recreates the evolutionary process with multiple cycles of drift, mutation, and selection. The genomic parameters for the simulations reproduce the soybean genome with detailed information (Schmutz et al., 2010). We considered a genetic map architecture of 20 chromosomes with 115 cM average length, which collectively spanned 950 Mb. For each chromosome, 1,000 segregating sites were assigned.

Our study focused on the simulation of grain yield (in tons per hectare) as the primary trait of interest. The genetic architecture of the simulated trait was assumed to be infinitesimal with 70% of all segregating sites, which were not necessarily utilized as markers, having a non-zero effect sampled from a normal distribution. The genotype-by-environment variance provided a non-heritable variation attributed to the season. Residual variance remained constant throughout the simulation, causing a reduction in heritability overtime as the genetic variance decreased. Simulations began assuming an average yield of 3.00 t ha^{-1}. The function *addTraitAEG* from the AlphaSimR package was utilized for the simulation of the phenotypic values. All simulation code is available on GitHub.[1]

Additive genetic effects, genotype-by-environment interaction, and residuals were simulated from Gaussian distribution using variance components estimated from the SoyNAM dataset (Diers et al., 2018; Xavier et al., 2018a) as $\sigma^2_a = 25$, $\sigma^2_{G\times E} = 49$, $\sigma^2_e = 121$, and $h^2 = 0.12$. The parameter estimation from the SoyNAM dataset was based on a multivariate genomic best linear unbiased prediction (GBLUP) model with unstructured genetic covariance and diagonal residual covariance, fitting grain yield from all 18 environments as response variables and using as explanatory variables the overall mean (fixed) and a polygenic term (random). The final estimates of the variance components for σ^2_a, σ^2_e, and h^2 were obtained as averages across the 18 environments, whereas $\sigma^2_{G\times E}$ was computed as the average off-diagonal of the variance–covariance matrix.

The main simulation settings followed a soybean breeding program with 300 families per cycle and with 50 individuals per family, producing a total of 15,000 individuals per cycle. After crossing, the populations were inbred *via* single seed descendent (SSD) until $F_{2:4}$, as shown in **Figure 1**, where lines were evaluated in field trials and genotyped with a single nucleotide polymorphism (SNP) array similar to the Soybean 6K SNP chip (Akond et al., 2013). Individuals were then selected to become parents of the upcoming breeding cycle using the phenotypic and genotypic information. The calibration of genomic prediction leveraged data from the previous three breeding cycles, thus leveraging information from up to 45,000 individuals per model. The processes of selecting and crossing were repeated for 200 cycles to capture the theoretical plateau of genetic gains across all simulated parameters. Each breeding scenario was reproduced 60 times with different computational random seeds.

A second simulation with 100 breeding cycles was performed with varying numbers of families and offspring, where five combinations that use the same number of resources were chosen—300 × 50, 250 × 60, 200 × 75, 150 × 100, and 100 × 150—where the combinations correspond to the number

[1]https://github.com/Ederdbs/GenomicSelection

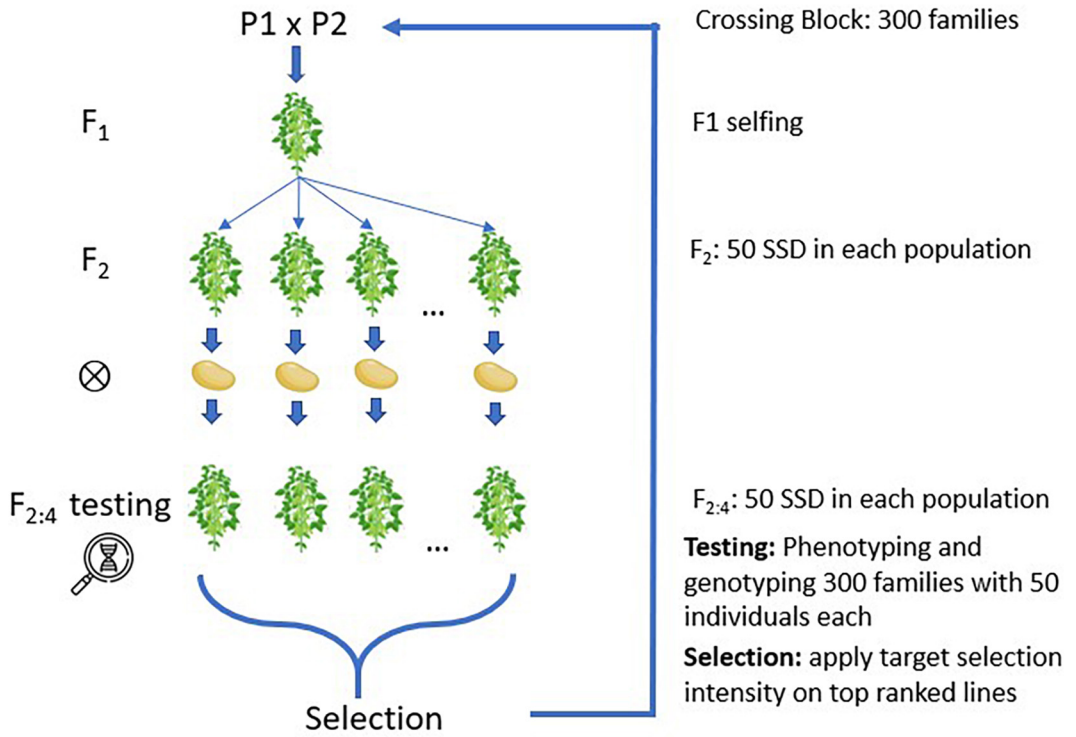

FIGURE 1 | Simulated families created and inbreeding using single seed descent (SSD).

of families and individuals per family, respectively. Each breeding scenario was reproduced 45 times with different random seeds.

Genotypic and phenotypic data were simulated with the R package AlphaSimR (Gaynor et al., 2020), reproducing the previous methodological framework (Faux et al., 2016). The software was utilized to simulate the founder population, perform selection, fingerprint individuals with the specified SNP chip, make crosses, generate offspring, inbred individuals, and simulate phenotypic values. All simulations and subsequent statistical analyses of the results were performed using R software (R Core Team, 2020). The code was run in parallel by distributing the multiple breeding scenarios over 960 cores, requiring approximately 10 h of computation per run. The R package doParallel (Ooi et al., 2019) was utilized to parallelize the runs.

Evaluation of Simulated Scenarios

Evaluation of the breeding strategies, selection intensities, and selection models was based on previous studies (Daetwyler et al., 2013). The evaluation criteria included the population mean across breeding cycles, genetic variance, and accuracy. Analyses were performed within a generation, combining the data from the repeated simulation runs. The statistical model for the analysis of simulated data was the following:

$$y = 1\mu + \mathbf{X}_m m + \mathbf{X}_s s + \mathbf{X}_i i + \mathbf{X}_p p + \varepsilon$$

where y is the vector of the random variable of the simulated population; μ is the model intercept; $\mathbf{X}$ represents the incidence matrix, which is further divided to accommodate the three factors under evaluation ($\mathbf{X}_m$, $\mathbf{X}_s$, $\mathbf{X}_i$, and $\mathbf{X}_p$); m for the selection model; s for the breeding strategy; i for the selection intensity; p for the population design, as combinations of the number of families and individuals per family; and ε is the vector of residuals, assumed to be distributed as $\varepsilon \sim N\left(0, I\sigma_\varepsilon^2\right)$. The statistical test of multiple comparison was based on Tukey's range test with 5% probability of error fit using the built-in R function TukeyHSD. This model was used to generate **Figure 2**.

Selection Models

The following selection models are evaluated: (1) True breeding values (*TBV*)—true breeding value, which serves as the upper limit of the achievable prediction power; (2) *Random*—random selection of individual, as the worst-case scenario; (3) *Pheno*—phenotypic-based selection without the use of genomic information; (4) *GBLUP*—the genomic best linear unbiased predictor fitted with REML (restricted maximum likelihood) variance components (Nejati-Javaremi et al., 1997; Habier et al., 2007); (5) *BayesA*—Bayesian shrinkage regression that assigns a t prior to marker effects (Meuwissen et al., 2001); (6) *BayesB*—an extension of BayesA with variable selection (Meuwissen et al., 2001); (7) *FLM*—fast Laplace model (Xavier, 2019), an empirical Bayes model with a double exponential prior for marker effects; and (8) *RF*—random forest regression (Breiman, 2001), a common machine learning procedure based on bootstrapping aggregation of multiple decision trees. The models GBLUP, BayesA, BayesB, and FLM were fitted using

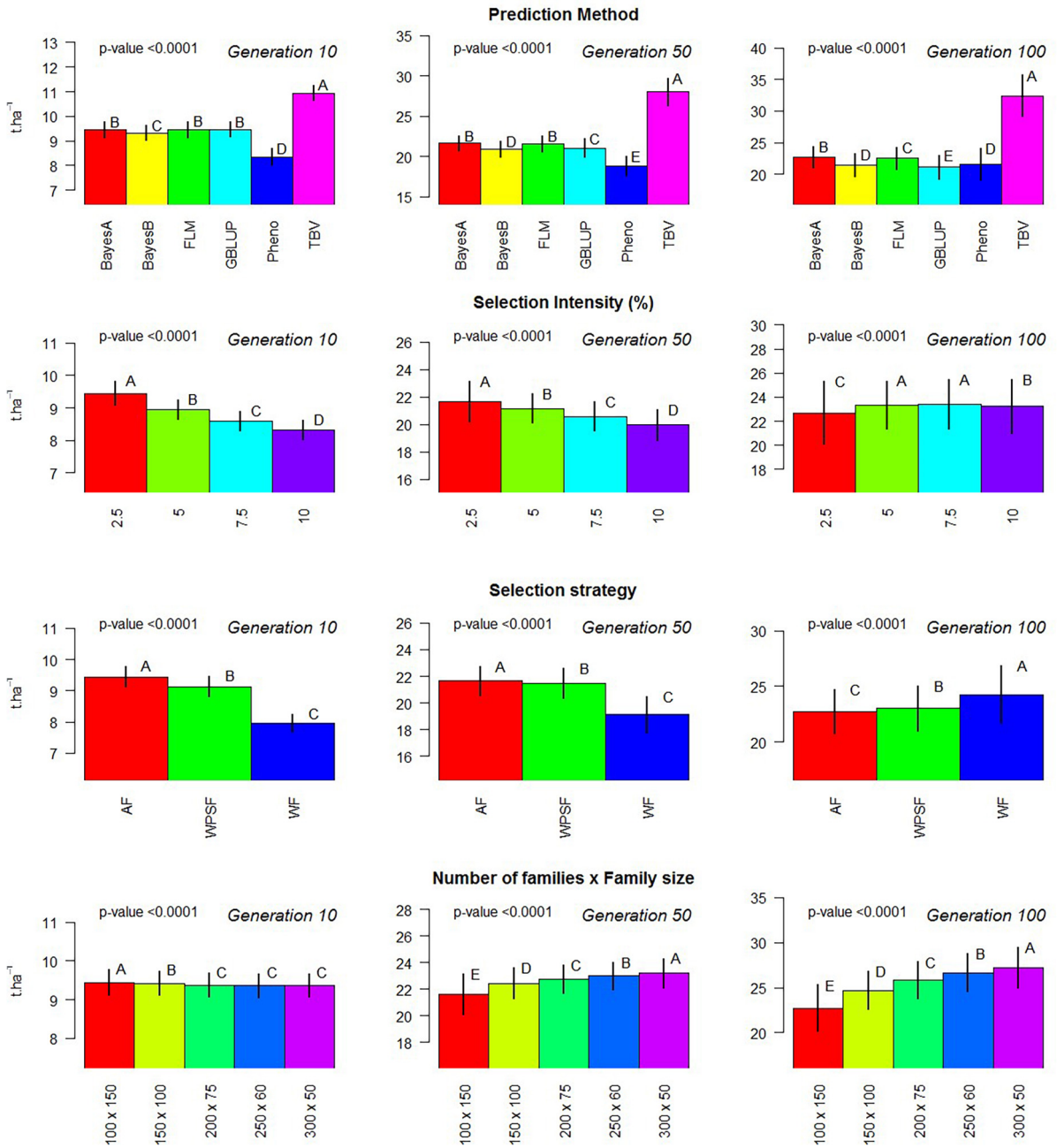

FIGURE 2 | Evaluation of individual factors on population means. Multiple comparison test: Capital letters indicate difference in means across factor with Tukey's range test with 5% alpha level contrasting the levels of each factor (prediction method, selection intensity, balance between population size and family size, and breeding strategy) on generations 10, 50, and 100.

the R package bWGR (Xavier et al., 2019) and solved *via* expectation–maximization (EM). The model RF was fitted using the R package ranger (Wright et al., 2020) with default settings.

As a brief description of the GS model, these models in function on genomic information can be written in terms of the linear model:

$$y = \mathbf{X}b + f(\mathbf{M}) + \varepsilon$$

where y is the vector of phenotypic values; $\mathbf{X}$ is the incidence matrix of the environment term treated as a fixed effect; b is a vector of environmental means; $f(\mathbf{M})$ is the function of markers that describe the genetic merit of individuals; and ε is a random vector of residuals, assumed to be distributed as $\varepsilon \sim N\left(0, \mathbf{I}\sigma_{\varepsilon}^{2}\right)$. The genetic function of markers, $f(\mathbf{M})$, varied from model to model. For GBLUP, BayesA, and FLM, the function was linear and the marker effects were strictly additive; thus, the function of markers was $f(\mathbf{M}) = \mathbf{M}\beta$. The distinction of the models was the

prior assigned to the distribution of marker effects, being normal for GBLUP, distributed as Student's t for BayesA, and distributed as a double exponential for FLM. The function describing BayesB was $f(\mathbf{M}) = \mathbf{M}\beta\gamma$, which is also linear, but with a variable selection term (γ) that caused further shrinkage to the Student's t prior assigned to the marker effects. The only non-linear model under evaluation was random forest, in which case the genetic function is a linear ensemble of multiple independent regression trees (T): $f(\mathbf{M}) = n^{-1} \sum T(m \in \mathbf{M})$.

Breeding Strategy

The breeding strategies were based on soybean breeding designs previously described in the literature (Backes et al., 2003; Sebastian et al., 2010; de Cássia Pereira et al., 2017; da Silva et al., 2018; Smallwood et al., 2019). The following approaches were considered in this study:

AF: across-family selection. Genotypes are selected across families based on their estimated genetic merit, without regard for their family structure or any constraint for selecting multiple individuals from the same pedigree.

WF: within-family selection. In this strategy, all families were equally represented in the advancements. The best genotypes from each family are selected to become parents in the upcoming generations.

WPSF: within the pre-selected family. This strategy comprises two steps. Firstly, the family level selection is performed to identify the best-performing families (top 30%). Secondly, the selection of individuals occurs within the family. With fewer families to select from, more individuals per family will be parenting the upcoming generation compared to WF.

Selection Intensity

Four levels of selection intensity were considered: 2.5, 5.0, 7.5, and 10.0%. These values represent the percentages of individuals selected to be used as parents of the next generation. The selection of parental combinations was performed at random; thus, it is possible that not all selected individuals served as parents.

RESULTS

Genetic Gains

The simulation results presented in **Figure 3** summarize the population means over the course of 200 cycles. **Supplementary Table 1** provides the population means for all combinations of treatments under evaluation in breeding cycles 10, 100, and 200. Across all scenarios, the population mean of random selection is anchored at the starting point. Selection of TBV represents the upper boundary of each scenario; hence, these are particularly useful to contrast the potential of the different scenarios. The highest long-term population means from selection on TBV occurred WF with loose selection intensities (7.5–10%). Genetic gains were generally closer to those from TBV when selections were performed WPSF.

Phenotypic selection outperformed GS over the course of 200 breeding cycles. Selection using random forest provided poor predictive performance in all scenarios, possibly due to the non-additive nature of the regression trees fitting a strictly additive genetic architecture. All linear genomic models (BayesA, BayesB, FLM, and GBLUP) provided similar outcomes. When conditioning for all other varying parameters, BayesA and FLM were the best-performing models within the first 100 breeding cycles (**Figure 2**).

After 10 cycles of selection, the highest gains were attained at the highest selection intensity (2.5%), which characterizes the short-term gain benefit from a higher selection pressure while the genetic variance is still abundant. After 100 breeding cycles, the genetic gains are affected by the combination of selection intensity and breeding strategy. For example, selection performed AF using BayesA provided the highest gains with a selection intensity of 10%, whereas, under WF, the highest gains occurred with a selection intensity of 2.5%. Such discrepancy is attributed to the amount of genetic variance left for long-term selection.

The highest long-term gains were reached when selections were performed WF. The maximum attainable, as benchmarked by selection upon TBV, resulted in a grain yield of 54 t ha^{-1} (WF), being 35% higher than AF selections and 46% higher than WPSF (**Supplementary Table 2**). The overall trend for long-term gains using GS followed the order WF > WPSF > AF. When the selections were based on phenotypic values, the genetic gains outpaced the GS run for all strategies (AF, WPSF, and WF), whereas that was not observed within the first 100 cycles (**Figure 1**). In fact, phenotypic selection WF was the third highest performing model, behind AF and WF selections performed on TBVs. The impact of each factor on the prediction accuracy over 200 breeding cycles is provided in **Supplementary Figure 1**.

Figure 2 summarizes the results of the simulation performed within 100 cycles, where different family sizes were an additional variable under evaluation. Within 10 breeding cycles, the scenario of 100 families with 150 individuals displayed the highest average, although the differences were negligible. Over the course of 50 and 100 breeding cycles, the number of families and the family sizes displayed significant differences in the genetic gains, with larger differences as generations progressed. The overall trend was that a greater number of families increase the gain in the long term.

Diversity Loss

The decay in genetic variance overtime is presented in **Figure 4**. The number of cycles to exhaust 80% of the genetic variance is provided in **Supplementary Table 3**. The study simulates closed populations without the inflow of external variation, the existing genetic variance consumed overtime as selection takes place. Overall, a fast decay in genetic variance is observed under a higher selection pressure, whereas a lower selection pressure preserved more genetic variance in the long term. When selection was performed at random, over 80% of the initial genetic variance remained after 200 breeding cycles. The interaction between the selection intensity and selection strategy was significant ($p < 0.01$) across all selection models.

Within-family selection preserved the genetic variance for more cycles (**Figure 4**). Selection WF based on TBVs exhausted 80% of the genetic variance within 48–69 breeding cycles, whereas AF and WPSF selections on TBVs exhausted 80% of

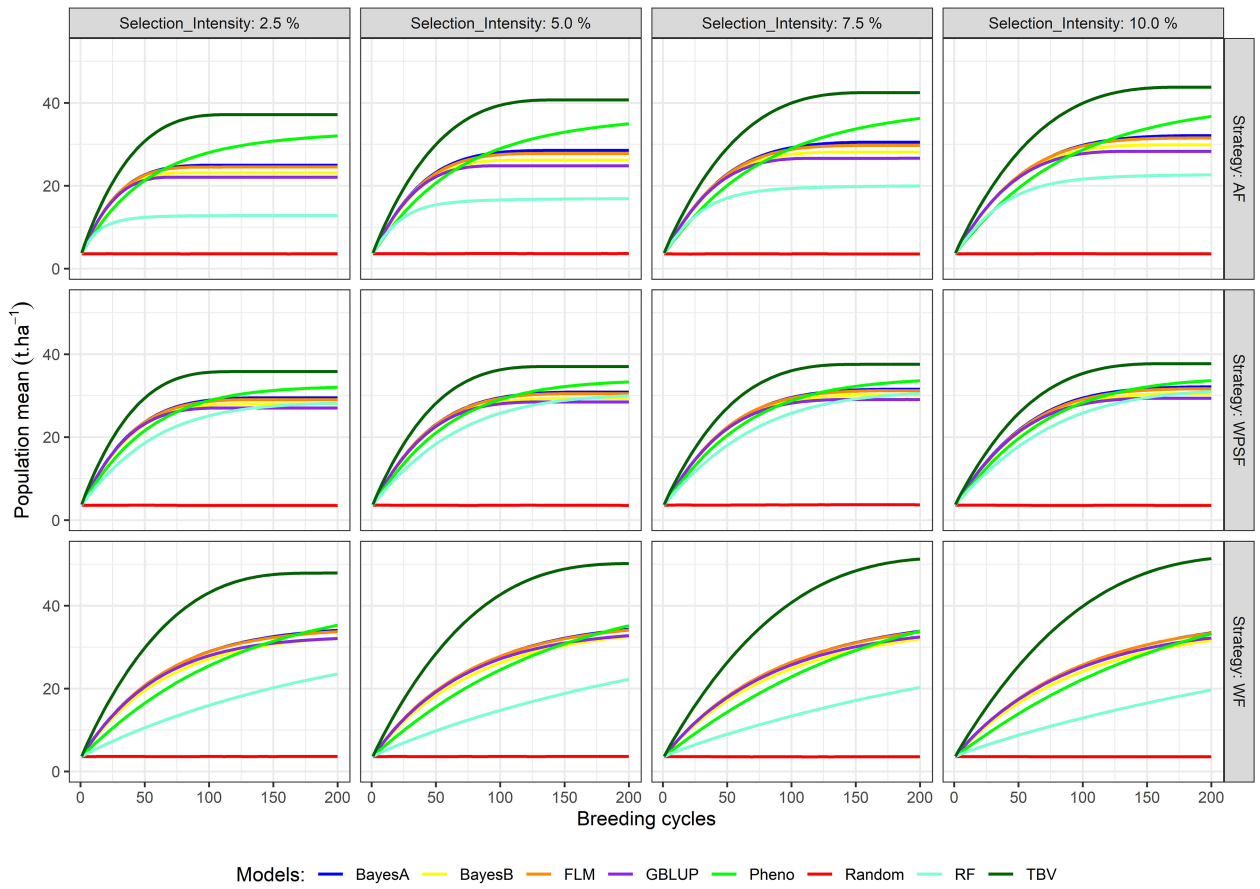

FIGURE 3 | Population means across 200 breeding cycles. Colors correspond to the different selection methods, grid columns represent the selection intensity (2.5, 5, 7.5, and 10%), and grid rows represent the breeding strategies, where individuals were selected across family (AF), within pre-selected families (WPSF), or within-family (WF).

the diversity between 25 and 42 cycles (**Supplementary Table 3**). Depletion of genetic variance was more pronounced with GS. Under the selection intensity of 10%, BayesA selection WF exhausted 80% of the variance after 184 cycles, whereas selections AF and WPSF display the same diversity loss after 54 and 58 cycles, respectively.

Diversity loss attributed to genetic drift is presented in **Figure 5**. These results assess the impact of bottlenecking the population through the various combinations of breeding strategy and selection intensity, utilizing random selections to avoid the confounding effect of directional selection. Higher rates of drift occurred under a higher selection pressure (2.5%). Strategy-wise, losses were highest for selection WPSF, with little difference across the selection intensities, ranging from −0.325 to −0.353%. The lowest rate of drift was observed under WF selection, with the rate of losses ranging from −0.199 to −0.136%.

DISCUSSION

Genomic prediction has become an important tool for selection and breeding in agriculture as it can enhance the rate of genetic gain in comparison to pedigree and phenotype-based selection by leveraging information on relationship and the linkage disequilibrium between the marker and the quantitative trait locus (QTL; Meuwissen et al., 2001; Habier and Fernando, 2009; Bernardo, 2010; Crossa et al., 2013, 2017; Daetwyler et al., 2013; de Los Campos et al., 2013). In soybean, the value of genomic prediction has been assessed and described in recent years (Jarquín et al., 2014; Xavier et al., 2016, 2018a,b; Diers et al., 2018; Matei et al., 2018; Xavier and Rainey, 2020). These studies agreed that adequate composition of the training data is imperative to successful and accurate prediction. The definition of an optimized training set entails (1) maximizing the genetic relationship between the training and target populations and (2) collecting phenotypic information from year–location combinations that represent the target population of environments. Whereas factors that affect genomic predictions for short-term gains have been well characterized, it is unclear which factors affect long-term genetic gains. The answer for that would come from long-term simulations, such as the present study. Primarily, simulations enable the optimization of the modern breeding program in animal and plant species (Yu et al., 2005; Hickey et al., 2014; Cowling et al., 2015, 2020;

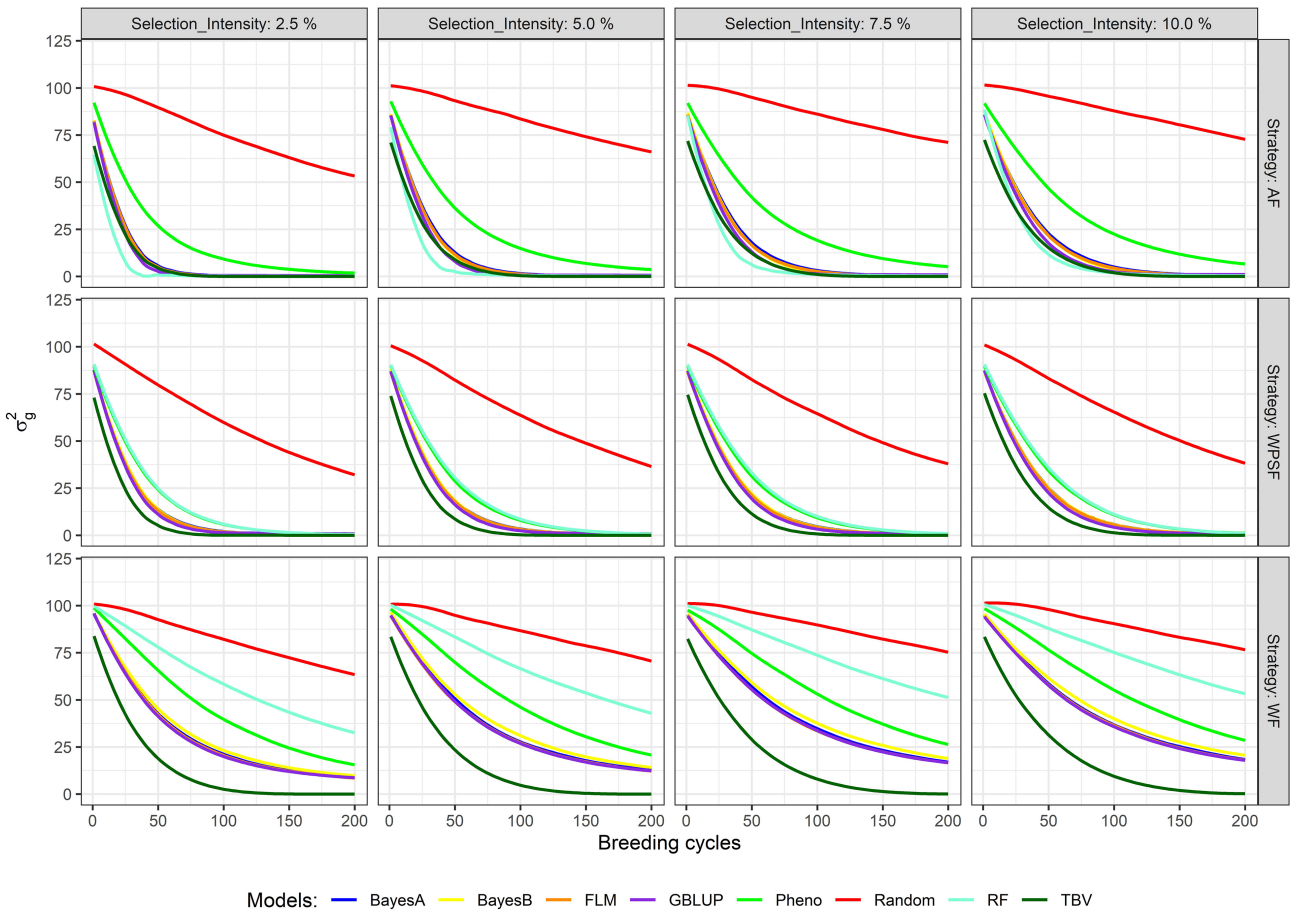

FIGURE 4 | Genetic variance across 200 breeding cycles. Colors correspond to the different selection methods, grid columns represent the selection intensities (2.5, 5, 7.5, and 10%), and grid rows represent the breeding strategies, where individuals are selected across family (AF), within pre-selected families (WPSF), or within family (WF).

Gorjanc and Hickey, 2018; Muleta et al., 2018) by enabling the assessment of the breeding conditions that increase the rate of genetic gains, the conservation of useful genetic diversity, and the best allocation of breeding resource, such as the number of field plots, genotyping density, number of crosses, and population size (Heffner et al., 2010; Gonen et al., 2017; Gorjanc et al., 2017a,b).

Simulations indicate that linear models outperformed random forest for complex traits controlled by additive genetics and additive genotype-by-environment interactions. Under different scenarios, other studies found machine learning methods to display similar performances (Li et al., 2018; Ali et al., 2020). The discrepancy in the results is likely due to the nature of trait and population under evaluation, as machine learning predictions could be suitable for more structured populations and with some degree of epistatic control (Xavier, 2019; Abdollahi-Arpanahi et al., 2020). We also acknowledge that random forest was run with default settings in this study, and parameter tuning would benefit its predictive performance.

Selection factors provided a similar outcome to the findings in other studies (Gorjanc and Hickey, 2018; Santantonio and Robbins, 2020), where the authors assessed balancing short- and long-term sustainable gains in plant breeding. Their results indicate that higher population sizes provide higher long-term gains. An alternative framework for the maximization of long-term response to selection is proposed by Goddard (2009) based on the use of selection indexes that account for allele frequency aiming to account for the value of rare loci and in short- and long-term gains. Under limited resources, our simulations indicate that a lower selection pressure generally contributes to long-term gains at the cost of compromising short-term gains. Across breeding strategies, WPSF appears to provide reasonable gains in both the short and the long term while having the range of gains being less influenced by selection pressure. WPSF is an intermediate between AF and WF, and the results are, in fact, intermediary between the short-term gains provided by AF selections and the long-term gains provided by WF selection.

The real-life trend of genetic gains in soybeans is positive, but variable across geographies. In North America, the rates of genetic gain have been estimated to be 23.4 kg ha^{-1} $year^{-1}$ (Fox et al., 2013), 26.5 kg ha^{-1} $year^{-1}$ (Koester et al., 2014), and 16.8 kg ha^{-1} $year^{-1}$ (Rogers et al., 2015). In the southern regions of Brazil, the rates of genetic gains were estimated to be

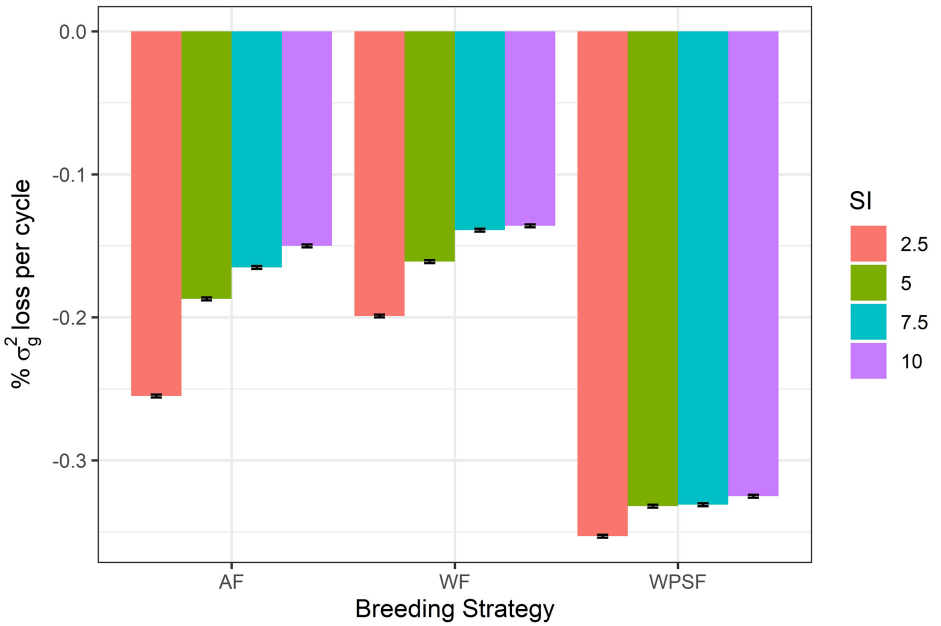

FIGURE 5 | Genetic drift per cycle under random selection across family (AF), within pre-selected families (WPSF), or within family (WF) in different selection intensities (SI).

71.5 kg ha^{-1} year^{-1} (Lange and Federizzi, 2009) and 40.0 kg ha^{-1} year^{-1} (Todeschini et al., 2019); in Argentina, the rate has been reported to be 44.3 kg ha^{-1} year^{-1} (de Felipe et al., 2016). These reports provide insight from the perspective of traditional breeding progress before the deployment of GS and, in most cases, with lengthy breeding cycles with the choice of parents taking place in advanced generations and commercial products. Our simulations provided higher annual gains than what has been reported; however, with the advent of earlier evaluations and increasing trust in genomic prediction, it is likely that annual genetic gains will be progressively and iteratively optimized for multiple factors, including those evaluated in the present study (model, selection intensity, family size, and breeding strategy).

The selection of unproven parents from earlier generations is often interpreted as gambling with high risk and high rewards, even though much of the risk is mitigated with the use of genomic information with robust statistical models calibrated with phenotypic data from multiple years. In addition to advancements, more opportunities arise with the use of genomics to predict and select the best combinations for crossing that further increase the probability of generating elite offspring. Previous studies have evaluated population-level selection strategies in further detail (Bernardo, 2010; Jannink, 2010; Kemper et al., 2012; Daetwyler et al., 2015; Ma et al., 2016; Goiffon et al., 2017; Matei et al., 2018) with the goal of preserving the segregation of low-frequency haplotypes for long-term gains (Beukelaer et al., 2017). Balancing the number of families and the family size can be a fundamental part of the strategy to continue the steady gains overtime (**Figure 2**), and, whereas the difference is not perceived in the short term, the magnitude of grain increases significantly overtime. Yet, multiple factors should be taken into account when allocating resources in terms of the number of families and family size (Lindgren et al., 1997; Fu, 2015).

Scenarios simulated as provided herein were based on the parental selection at the F4 stage, which is commonly perceived as an early generation for recycling as the quality and the quantity of phenotypic data are still scarce, of doubtful quality, and in many cases, without replication. Nevertheless, early recycling is a promising framework for speeding up the rate of genetic gain by shortening the length of the breeding cycles. In fact, shortening the breeding cycles while inducing multiple cycles a year reproduces a framework referred to as "speed breeding" (Hickey et al., 2019; Nagatoshi and Fujita, 2019; Jähne et al., 2020). Recent studies often support recombination in the early stages of inbred development (Gaynor et al., 2017), more so as the accuracy of selection in the early stages benefits greatly from the GS. Another important aspect of parental selection regards the management of genetic diversity in modern plant breeding, which is largely ignored and not always adequately measured (Fu, 2015). Our results indicate that the multiple factors in the breeding design can affect the rate of diversity loss, mainly selection pressure and selection strategy (**Supplementary Table 2**), and that one must consider to balance these factors to attain the desired gain in the short term without compromising long-term gains. That is particularly the case for soybeans, whose germplasm-wise genetic diversity is considered low when compared to that of other species (Martin, 1982). Some Canadian soybean breeding programs have maintained diversity through decades of breeding while fixing maturity genes (Bruce et al., 2019). In the United States, soybean population structures and diversity varied by maturity group (Vaughn and Li, 2016), which

suggests that new sources of variation could be obtained through the introgression of material from different regions.

The diversity available in breeding programs affects the accuracy of breeding values by dictating the amount of existing genetic signals to select upon an effective population size (Meuwissen, 2009). With restricted diversity, the genotyping density and marker distribution can be optimized to capture the existing variation in the target population with the goal of increasing genomic prediction accuracy (Ma et al., 2016). Of course, the long-term impacts of selection on genetic variance also vary depending on the genetic variance of interest, as the prominence of additive and non-additive variances is not the same over multiple cycles of selection (Paixão and Barton, 2016).

In soybeans, the management of diversity is necessary to ensure useful variability for future breeding objectives, such as yield performance under drought or waterlogging (Valliyodan et al., 2017), the seed oil and protein content profiles (Stewart-Brown et al., 2019), and disease resistance (de Azevedo Peixoto et al., 2017). Monitoring genetic diversity in the genomic era can be performed through tracking overtime changes in allele frequencies (Allier et al., 2019b; de Castro Lara et al., 2020; Meuwissen et al., 2020). We showed that selection could quickly exhaust genetic diversity under closed breeding systems, and breeding systems can benefit from balancing short gains to preserve diversity and assure long-term gains. Such balance had been the focal point of recent studies (Cowling et al., 2017; Gorjanc et al., 2018; Ru and Bernardo, 2019, 2020; Santantonio and Robbins, 2020) seeking for avenues to extend genetic resources with genomic tools, including the selection of material from germplasm collection to expend the genetic basis of elite programs. In addition to germplasm introgression, increases in genetic diversity in soybeans have been done in the past through mutagenic agents (Curtin et al., 2011; Khan, 2013; Haun et al., 2014; Demorest et al., 2016) and more recently, through genome editing techniques based on CRISPR-Cas9 (Cai et al., 2015, 2018a,b; Jacobs et al., 2015; Sun et al., 2015; Zheng et al., 2020) and target recombination for directional backcrossing (Ru and Bernardo, 2019, 2020).

The simulations performed in our study indicate that GS enables higher rates of genetic gain in the short and medium term compared with phenotype selection, but also led to faster extinction of the genetic variance. Thus, genomic prediction and selection must be applied mindfully with the purpose of maximizing gains while maintaining genetic variance. We found that a breeding strategy that balances selection at the family level, and within and across family at the individual level, can mitigate losses in genetic variance while providing satisfying genetic gains in the short term. Simulation is a powerful and inexpensive tool to test hypotheses, and for future studies, we envision addressing the importance of other important breeding parameters. Namely, future studies should focus on investigating (1) the optimal generation to select the parents and its trade-off with the accuracy of selection; (2) the influence of non-additive and non-infinitesimal genetic architecture and how machine learning would perform in such conditions; (3) the long-term effect of different models designed to select parental combinations; (4) the impact of different island models where new sources of variation are constantly infused into the main breeding panel; and (5) what would be the potential benefit of breeding hybrid soybeans assuming there are variable levels of dominance.

CONCLUSION

Long-term gains were influenced by the interaction among GS models, breeding strategy, and selection intensity. Adequate handling of these factors will aid breeding programs to ensure genetic gains in short, medium, and long term. Therefore, the breeding strategy is the most influential factor and, therefore, is a key criterion to conserve genetic variance and obtain the highest population mean overtime. The absolute impact of the selection intensity is lower than that of the breeding strategy and GS model. The benefits of balancing family size and the number of families were not perceived on short-term gains. Additive GS models (BayesA, BayesB, FLM, and GBLUP) have similar behaviors in selecting the best individuals, whereas RF has poor predictive performance when implemented with default settings. In summary, a combination of strategies may be necessary for balancing the short-, medium-, and long-term genetic gains in breeding programs while preserving genetic variance.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/Ederdbs/GenomicSelection.

AUTHOR CONTRIBUTIONS

ÉS and AX implemented the research, contributed with ideas to the algorithms, and wrote the manuscript. MF implemented the research, contributed ideas, and wrote the manuscript. All authors approved the final version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGMENTS

The authors are grateful to the Universidade Estadual do Centro-Oeste (UNICENTRO), because this manuscript is part of the thesis of the ÉS and to Corteva Agriscience for support in computational resources. The authors also thank the reviewers for their helpful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.637133/full#supplementary-material

REFERENCES

Abdollahi-Arpanahi, R., Gianola, D., and Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel. Evolut.* 52:12. doi: 10.1186/s12711-020-00531-z

Akond, M., Liu, S., Schoener, L., Anderson, J. A., Kantartzi, S. K., Meksem, K., et al. (2013). A SNP-Based genetic linkage map of soybean using the SoySNP6K Illumina infinium BeadChip genotyping array. *Plant Genet. Genomics Biotechnol.* 1, 80–89. doi: 10.5147/pggb.v1i3.154

Ali, M., Zhang, L., DeLacy, I., Arief, V., Dieters, M., Pfeiffer, W. H., et al. (2020). Modeling and simulation of recurrent phenotypic and genomic selections in plant breeding under the presence of epistasis. *Crop J.* 8, 866–877. doi: 10.1016/j.cj.2020.04.002

Allier, A., Lehermeier, C., Charcosset, A., Moreau, L., and Teyssèdre, S. (2019a). Improving short- and long-term genetic gain by accounting for within-family variance in optimal cross-selection. *Front. Genet.* 10:1006. doi: 10.3389/fgene.2019.01006

Allier, A., Teyssèdre, S., Lehermeier, C., Claustres, B., Maltese, S., Melkior, S., et al. (2019b). Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program. *Theor. Appl. Genet.* 132, 1321–1334. doi: 10.1007/s00122-019-03280-w

Backes, R. L., Reis, M. S., Cruz, C. D., Sediyama, T., and Sediyama, C. S. (2003). Correlation estimates and assessment of selection strategies in five soybean populations. *CBAB* 3, 107–116. doi: 10.12702/1984-7033.v03n02a03

Bernardo, R. (2010). Genomewide selection with minimal crossing in self-pollinated crops. *Crop Sci.* 50, 624–627. doi: 10.2135/cropsci2009.05.0250

Beukelaer, H. D., Badke, Y., Fack, V., and Meyer, G. D. (2017). Moving beyond managing realized genomic relationship in long-term genomic selection. *Genetics* 206, 1127–1138. doi: 10.1534/genetics.116.194449

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bruce, R. W., Torkamaneh, D., Grainger, C., Belzile, F., Eskandari, M., and Rajcan, I. (2019). Genome-wide genetic diversity is maintained through decades of soybean breeding in Canada. *Theor. Appl. Genet.* 132, 3089–3100. doi: 10.1007/s00122-019-03408-y

Cai, Y., Chen, L., Liu, X., Guo, C., Sun, S., Wu, C., et al. (2018a). CRISPR/Cas9-mediated targeted mutagenesis of GmFT2a delays flowering time in soya bean. *Plant Biotechnol. J.* 16, 176–185. doi: 10.1111/pbi.12758

Cai, Y., Chen, L., Sun, S., Wu, C., Yao, W., Jiang, B., et al. (2018b). CRISPR/Cas9-mediated Deletion of large genomic fragments in soybean. *Int. J. Mol. Sci.* 19:3835. doi: 10.3390/ijms19123835

Cai, Y., Chen, L., Liu, X., Sun, S., Wu, C., Jiang, B., et al. (2015). CRISPR/Cas9-mediated genome editing in soybean hairy roots. *PLoS One* 10:e0136064. doi: 10.1371/journal.pone.0136064

Chen, G. K., Marjoram, P., and Wall, J. D. (2009). Fast and flexible simulation of DNA sequence data. *Genome Res.* 19, 136–142. doi: 10.1101/gr.083634.108

Cowling, W. A., Gaynor, R. C., Antolín, R., Gorjanc, G., Edwards, S. M., Powell, O., et al. (2020). In silico simulation of future hybrid performance to evaluate heterotic pool formation in a self-pollinating crop. *Sci. Rep.* 10:4037. doi: 10.1038/s41598-020-61031-0

Cowling, W. A., Li, L., Siddique, K. H. M., Henryon, M., Berg, P., Banks, R. G., et al. (2017). Evolving gene banks: improving diverse populations of crop and exotic germplasm with optimal contribution selection. *J. Exp. Bot.* 68, 1927–1939. doi: 10.1093/jxb/erw406

Cowling, W. A., Stefanova, K. T., Beeck, C. P., Nelson, M. N., Hargreaves, B. L. W., Sass, O., et al. (2015). Using the animal model to accelerate response to selection in a self-pollinating crop. *G3* 5, 1419–1428. doi: 10.1534/g3.115.018838

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., et al. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3* 3, 1903–1926. doi: 10.1534/g3.113.008227

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Curtin, S. J., Zhang, F., Sander, J. D., Haun, W. J., Starker, C., Baltes, N. J., et al. (2011). Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. *Plant Physiol.* 156, 466–473. doi: 10.1104/pp.111.172981

da Silva, F. M., De MatosPereira, E., Val, B. H. P., Perecin, D., Mauro, A. O. D., Unêda-Trevisoli, S. H., et al. (2018). Strategies to select soybean segregating populations with the goal of improving agronomic traits. *Acta Scientiarum. Agronomy* 40:39324. doi: 10.4025/actasciagron.v40i1.39324

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi: 10.1534/genetics.112.147983

Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., and Hayes, B. J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics* 200, 1341–1348. doi: 10.1534/genetics.115.178038

de Azevedo Peixoto, L., Moellers, T. C., Zhang, J., Lorenz, A. J., Bhering, L. L., Beavis, W. D., et al. (2017). Leveraging genomic prediction to scan germplasm collection for crop improvement. *PLoS One* 12:e0179191. doi: 10.1371/journal.pone.0179191

de Cássia Pereira, F., Bruzi, A. T., de Matos, J. W., Rezende, B. A., Prado, L. C., and Nunes, J. A. R. (2017). Implications of the population effect in the selection of soybean progeny. *Plant Breed.* 136, 679–687. doi: 10.1111/pbr.12512

de Castro Lara, L. A., Pocrnic, I., Gaynor, R. C., and Gorjanc, G. (2020). Temporal and genomic analysis of additive genetic variance in breeding programmes. *bioRxiv* [Preprint]. doi: 10.1101/2020.08.29.273250

de Felipe, M., Gerde, J. A., and Rotundo, J. L. (2016). Soybean Genetic gain in maturity Groups III to V in argentina from 1980 to 2015. *Crop Sci.* 56, 3066–3077. doi: 10.2135/cropsci2016.04.0214

de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

Demorest, Z. L., Coffman, A., Baltes, N. J., Stoddard, T. J., Clasen, B. M., Luo, S., et al. (2016). Direct stacking of sequence-specific nuclease-induced mutations to produce high oleic and low linolenic soybean oil. *BMC Plant Biol.* 16:225. doi: 10.1186/s12870-016-0906-1

Diers, B. W., Specht, J., Rainey, K. M., Cregan, P., Song, Q., Ramasubramanian, V., et al. (2018). Genetic architecture of soybean yield and agronomic traits. *G3* 8, 3367–3375. doi: 10.1534/g3.118.200332

FAO (2021). *FAO Global Statistical Yearbook, FAO Regional Statistical Yearbooks*. Available online at: http://www.fao.org/faostat/en/#data/QC (accessed January 6, 2021).

Faux, A.-M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., et al. (2016). AlphaSim: software for breeding program simulation. *Plant Genome* 9, 1–14. doi: 10.3835/plantgenome2016.02.0013

Fox, C. M., Cary, T. R., Colgrove, A. L., Nafziger, E. D., Haudenshield, J. S., Hartman, G. L., et al. (2013). Estimating soybean genetic gain for yield in the Northern United States—Influence of cropping history. *Crop. Sci.* 53, 2473–2482. doi: 10.2135/cropsci2012.12.0687

Fu, Y.-B. (2015). Understanding crop genetic diversity under modern plant breeding. *Theor. Appl. Genet.* 128, 2131–2142. doi: 10.1007/s00122-015-2585-y

Gaynor, C., Gorjanc, G., Wilson, D., Hickey, J., and Money, D. (2020). *AlphaSimR: Breeding Program Simulations*. Available online at: https://CRAN.R-project.org/package=AlphaSimR (accessed July 6, 2020).

Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., et al. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Sci.* 57, 2372–2386. doi: 10.2135/cropsci2016.09.0742

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Goiffon, M., Kusmec, A., Wang, L., Hu, G., and Schnable, P. S. (2017). Improving response in genomic selection with a population-based selection strategy: optimal population value selection. *Genetics* 206, 1675–1682. doi: 10.1534/genetics.116.197103

Gonen, S., Ros-Freixedes, R., Battagin, M., Gorjanc, G., and Hickey, J. M. (2017). A method for the allocation of sequencing resources in genotyped livestock populations. *Genet. Sel. Evol.* 49:47. doi: 10.1186/s12711-017-0322-5

Gorjanc, G., Battagin, M., Dumasy, J.-F., Antolin, R., Gaynor, R. C., and Hickey, J. M. (2017a). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Sci.* 57, 216–228. doi: 10.2135/cropsci2016.06.0526

Gorjanc, G., Dumasy, J.-F., Gonen, S., Gaynor, R. C., Antolin, R., and Hickey, J. M. (2017b). Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. *Crop Sci.* 57, 1404–1420. doi: 10.2135/cropsci2016.08.0675

Gorjanc, G., Gaynor, R. C., and Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theor. Appl. Genet.* 131, 1953–1966. doi: 10.1007/s00122-018-3125-3

Gorjanc, G., and Hickey, J. M. (2018). AlphaMate: a program for optimizing selection, maintenance of diversity and mate allocation in breeding programs. *Bioinformatics* 34, 3408–3411. doi: 10.1093/bioinformatics/bty375

Habier, D., and Fernando, R. L. (2009). Genomic selection using low-density marker panels. *Genetics* 182, 343–353. doi: 10.1534/genetics.108.100289

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186. doi: 10.1186/1471-2105-12-186

Haun, W., Coffman, A., Clasen, B. M., Demorest, Z. L., Lowy, A., Ray, E., et al. (2014). Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. *Plant Biotechnol. J.* 12, 934–940. doi: 10.1111/pbi.12201

Heffner, E. L., Lorenz, A. J., Jannink, J.-L., and Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci.* 50, 1681–1690. doi: 10.2135/cropsci2009.11.0662

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Sci.* 54, 1476–1488. doi: 10.2135/cropsci2013.03.0195

Hickey, L. T., Hafeez, N. A., Robinson, H., Jackson, S. A., Leal-Bertioli, S. C. M., Tester, M., et al. (2019). Breeding crops to feed 10 billion. *Nat. Biotechnol.* 37, 744–754. doi: 10.1038/s41587-019-0152-9

Jacobs, T. B., LaFayette, P. R., Schmitz, R. J., and Parrott, W. A. (2015). Targeted genome modifications in soybean with CRISPR/Cas9. *BMC Biotechnol.* 15:16. doi: 10.1186/s12896-015-0131-2

Jähne, F., Hahn, V., Würschum, T., and Leiser, W. L. (2020). Speed breeding short-day crops by LED-controlled light schemes. *Theor. Appl. Genet.* 133, 2335–2342. doi: 10.1007/s00122-020-03601-4

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42:35. doi: 10.1186/1297-9686-42-35

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., et al. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15:740. doi: 10.1186/1471-2164-15-740

Jarquin, D., Specht, J., and Lorenz, A. (2016). Prospects of genomic prediction in the USDA soybean germplasm collection: historical data creates robust models for enhancing selection of accessions. *G3* 6, 2329–2341. doi: 10.1534/g3.116.031443

Kemper, K. E., Bowman, P. J., Pryce, J. E., Hayes, B. J., and Goddard, M. E. (2012). Long-term selection strategies for complex traits using high-density genetic markers. *J. Dairy Sci.* 95, 4646–4656. doi: 10.3168/jds.2011-5289

Khan, H. (2013). A review on induced mutagenesis in soybean. *J. Cereals Oilseeds* 4, 19–25. doi: 10.5897/JCO10.004

Koester, R. P., Skoneczka, J. A., Cary, T. R., Diers, B. W., and Ainsworth, E. A. (2014). Historical gains in soybean (*Glycine max* Merr.) seed yield are driven by linear increases in light interception, energy conversion, and partitioning efficiencies. *J. Exp. Bot.* 65, 3311–3321. doi: 10.1093/jxb/eru187

Lange, C. E., and Federizzi, L. C. (2009). Estimation of soybean genetic progress in the South of Brazil using multi-environmental yield trials. *Sci. Agric.* 66, 309–316. doi: 10.1590/S0103-90162009000300005

Li, B., Zhang, N., Wang, Y.-G., George, A. W., Reverter, A., and Li, Y. (2018). Genomic prediction of breeding values using a subset of snps identified by three machine learning methods. *Front. Genet.* 9:237. doi: 10.3389/fgene.2018.00237

Lindgren, D., Wei, R.-P., and Lee, S. J. (1997). How to calculate optimum family number when starting a breeding program. *For. Sci.* 43, 206–212. doi: 10.1093/forestscience/43.2.206

Ma, Y., Reif, J. C., Jiang, Y., Wen, Z., Wang, D., Liu, Z., et al. (2016). Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). *Mol. Breed.* 36:113. doi: 10.1007/s11032-016-0504-9

Martin, S. K. S. (1982). Effective population size for the soybean improvement program in maturity groups 00 to IV1. *Crop Sci.* 22, 151–152. doi: 10.2135/cropsci1982.0011183X002200010035x

Matei, G., Woyann, L. G., Milioli, A. S., de Bem Oliveira, I., Zdziarski, A. D., Zanella, R., et al. (2018). Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. *Mol Breed.* 38:117. doi: 10.1007/s11032-018-0872-4

Meuwissen, T. H. E. (2009). Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* 41:35. doi: 10.1186/1297-9686-41-35

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

Meuwissen, T. H. E., Sonesson, A. K., Gebregiwergis, G., and Woolliams, J. A. (2020). Management of genetic diversity in the era of genomics. *Front. Genet.* 11:880. doi: 10.3389/fgene.2020.00880

Muleta, K. T., Pressoir, G., and Morris, G. P. (2018). Optimizing genomic selection for a sorghum breeding program in haiti: a simulation study. *G3*, 9, 391–401. doi: 10.1534/g3.118.200932

Nagatoshi, Y., and Fujita, Y. (2019). Accelerating Soybean Breeding in a CO2-Supplemented Growth Chamber. *Plant Cell Physiol.* 60, 77–84. doi: 10.1093/pcp/pcy189

Nejati-Javaremi, A., Smith, C., and Gibson, J. P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.* 75, 1738–1745. doi: 10.2527/1997.7571738x

Ooi, H., Corporation, M., Weston, S., and Tenenbaum, D. (2019). *doParallel: Foreach Parallel Adaptor for the "parallel" Package (Version 1.0.16)*. Available online at: https://CRAN.R-project.org/package=doParallel (accessed July 2, 2020).

Paixão, T., and Barton, N. H. (2016). The effect of gene interactions on the long-term response to selection. *PNAS* 113, 4422–4427. doi: 10.1073/pnas.1518830113

Pook, T., Schlather, M., and Simianer, H. (2019). MoBPS - modular breeding program simulator. *bioRxiv* [Preprint]. doi: 10.1101/829333

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rogers, J., Chen, P., Shi, A., Zhang, B., Scaboo, A., Smith, S. F., et al. (2015). Agronomic performance and genetic progress of selected historical soybean varieties in the southern USA. *Plant Breed.* 134, 85–93. doi: 10.1111/pbr.12222

Ru, S., and Bernardo, R. (2019). Targeted recombination to increase genetic gain in self-pollinated species. *Theor. Appl. Genet.* 132, 289–300. doi: 10.1007/s00122-018-3216-1

Ru, S., and Bernardo, R. (2020). Predicted genetic gains from introgressing chromosome segments from exotic germplasm into an elite soybean cultivar. *Theor. Appl. Genet.* 133, 605–614. doi: 10.1007/s00122-019-03490-2

Santantonio, N., and Robbins, K. (2020). A hybrid optimal contribution approach to drive short-term gains while maintaining long-term sustainability in a modern plant breeding program. *bioRxiv* [Preprint]. doi: 10.1101/2020.01.08.899039

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Sebastian, S. A., Streit, L. G., Stephens, P. A., Thompson, J. A., Hedges, B. R., Fabrizius, M. A., et al. (2010). Context-specific marker-assisted selection for improved grain yield in elite soybean populations. *Crop Sci.* 50, 1196–1206. doi: 10.2135/cropsci2009.02.0078

Smallwood, C. J., Saxton, A. M., Gillman, J. D., Bhandari, H. S., Wadl, P. A., Fallen, B. D., et al. (2019). Context-specific genomic selection strategies outperform phenotypic selection for soybean quantitative traits in the progeny row stage. *Crop Sci.* 59, 54–67. doi: 10.2135/cropsci2018.03.0197

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., et al. (2013). An Improved Genotyping by Sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* 8:e54603. doi: 10.1371/journal.pone.0054603

Stewart-Brown, B. B., Song, Q., Vaughn, J. N., and Li, Z. (2019). Genomic selection for yield and seed composition traits within an applied soybean breeding program. *G3* 9, 2253–2265. doi: 10.1534/g3.118.200917

Sun, X., Hu, Z., Chen, R., Jiang, Q., Song, G., Zhang, H., et al. (2015). Targeted mutagenesis in soybean using the CRISPR-Cas9 system. *Sci. Rep.* 5:10342. doi: 10.1038/srep10342

Todeschini, M. H., Milioli, A. S., Rosa, A. C., Dallacorte, L. V., Panho, M. C., Marchese, J. A., et al. (2019). Soybean genetic progress in South Brazil: physiological, phenological and agronomic traits. *Euphytica* 215:124.

Toledo, F. H., Pérez-Rodríguez, P., Crossa, J., and Burgueño, J. (2019). isqg: a binary framework for in silico quantitative genetics. *G3* 9, 2425–2428. doi: 10.1534/g3.119.400373

Valliyodan, B., Ye, H., Song, L., Murphy, M., Shannon, J. G., and Nguyen, H. T. (2017). Genetic diversity and genomic strategies for improving drought and waterlogging tolerance in soybeans. *J. Exp. Bot.* 68, 1835–1849. doi: 10.1093/jxb/erw433

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R., and Weigel, K. A. (2011). Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* 43:10. doi: 10.1186/1297-9686-43-10

Vaughn, J. N., and Li, Z. (2016). Genomic signatures of North American soybean improvement inform diversity enrichment strategies and clarify the impact of hybridization. *G3* 6, 2693–2705. doi: 10.1534/g3.116.029215

Wray, N., and Goddard, M. (1994). Increasing long-term response to selection. *Genet. Sel. Evol.* 26, 431–451. doi: 10.1186/1297-9686-26-5-431

Wright, M. N., Wager, S., and Probst, P. (2020). *ranger: A Fast Implementation of Random Forests*. Available online at: https://CRAN.R-project.org/package=ranger (accessed July 2, 2020).

Xavier, A. (2019). Efficient estimation of marker effects in plant breeding. *G3* 9, 3855–3866. doi: 10.1534/g3.119.400728

Xavier, A., Jarquin, D., Howard, R., Ramasubramanian, V., Specht, J. E., Graef, G. L., et al. (2018a). Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3* 8, 519–529. doi: 10.1534/g3.117.300300

Xavier, A., Muir, W. M., and Rainey, K. M. (2016). Assessing predictive properties of genome-wide selection in soybeans. *G3* 6, 2611–2616. doi: 10.1534/g3.116.032268

Xavier, A., Muir, W. M., and Rainey, K. M. (2019). bWGR: bayesian whole-genome regression. *Bioinformatics*. 36, 1957–1959. doi: 10.1093/bioinformatics/btz794

Xavier, A., and Rainey, K. M. (2020). Quantitative genomic dissection of soybean yield components. *G3* 10, 665–675. doi: 10.1534/g3.119.400896

Xavier, A., Thapa, R., Muir, W. M., and Rainey, K. M. (2018b). Population and quantitative genomic properties of the USDA soybean germplasm collection. *Plant Genet. Resour.* 16, 513–523. doi: 10.1017/S1479262118000102

Yabe, S., Yamasaki, M., Ebana, K., Hayashi, T., and Iwata, H. (2016). Island-model genomic selection for long-term genetic improvement of autogamous crops. *PLoS One* 11:e0153945. doi: 10.1371/journal.pone.0153945

Yu, J., Arbelbide, M., and Bernardo, R. (2005). Power of in silico QTL mapping from phenotypic, pedigree, and marker data in a hybrid breeding program. *Theor. Appl. Genet.* 110, 1061–1067. doi: 10.1007/s00122-005-1926-7

Zheng, N., Li, T., Dittman, J. D., Su, J., Li, R., Gassmann, W., et al. (2020). CRISPR/Cas9-based gene editing using egg cell-specific promoters in *Arabidopsis* and Soybean. *Front. Plant Sci.* 11:800. doi: 10.3389/fpls.2020.00800

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 September 2021
doi: 10.3389/fgene.2021.675500

Strategies to Assure Optimal Trade-Offs Among Competing Objectives for the Genetic Improvement of Soybean

Vishnu Ramasubramanian[1,2]* *and William D. Beavis*[1]

[1] George F. Sprague Population Genetics Group, Department of Agronomy, Ames, IA, United States, [2] Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, IA, United States

Plant breeding is a decision-making discipline based on understanding project objectives. Genetic improvement projects can have two competing objectives: maximize the rate of genetic improvement and minimize the loss of useful genetic variance. For commercial plant breeders, competition in the marketplace forces greater emphasis on maximizing immediate genetic improvements. In contrast, public plant breeders have an opportunity, perhaps an obligation, to place greater emphasis on minimizing the loss of useful genetic variance while realizing genetic improvements. Considerable research indicates that short-term genetic gains from genomic selection are much greater than phenotypic selection, while phenotypic selection provides better long-term genetic gains because it retains useful genetic diversity during the early cycles of selection. With limited resources, must a soybean breeder choose between the two extreme responses provided by genomic selection or phenotypic selection? Or is it possible to develop novel breeding strategies that will provide a desirable compromise between the competing objectives? To address these questions, we decomposed breeding strategies into decisions about selection methods, mating designs, and whether the breeding population should be organized as family islands. For breeding populations organized into islands, decisions about possible migration rules among family islands were included. From among 60 possible strategies, genetic improvement is maximized for the first five to 10 cycles using genomic selection and a hub network mating design, where the hub parents with the largest selection metric make large parental contributions. It also requires that the breeding populations be organized as fully connected family islands, where every island is connected to every other island, and migration rules allow the exchange of two lines among islands every other cycle of selection. If the objectives are to maximize both short-term and long-term gains, then the best compromise strategy is similar except that the mating design could be hub network, chain rule, or a multi-objective optimization method-based mating design. Weighted genomic selection applied to centralized populations also resulted in the realization of the greatest proportion of the genetic potential of the founders but required more cycles than the best compromise strategy.

Keywords: island model selection, recurrent selection, tradeoffs, optimization, genetic algorithms, genetic response, genomic selection, recurrence models

OPEN ACCESS

Edited by:
Diego Jarquin, University of Nebraska-Lincoln, United States

Reviewed by:
Reka Howard, University of Nebraska System, United States
Miguel Angel Lopez Murcia, Research Center of Sugar Cane, Colombia

***Correspondence:**
Vishnu Ramasubramanian
ivanvishnu@gmaill.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *03 March 2021*
Accepted: *17 August 2021*
Published: *24 September 2021*

Citation:
Ramasubramanian V and Beavis WD (2021) Strategies to Assure Optimal Trade-Offs Among Competing Objectives for the Genetic Improvement of Soybean. Front. Genet. 12:675500. doi: 10.3389/fgene.2021.675500

BACKGROUND

Responses to the selection of commodity crops have been enabled by decreasing the number of years per cycle of recurrent selection, by increasing the number of replicable genotypes (selection intensity), and by increasing the number of replicated field trials (heritability on an entry mean basis). In other words, genotypic improvements from responses to selection in commodity crops over the last 50 years (Specht et al., 2014) required monetary investments that became part of increased seed costs during the same time (Byrum et al., 2017; USDA-ERS, 2020). Since the emergence and adoption of Genomic Selection (GS), it has been possible to increase the numbers of genotypes that are evaluated, i.e., selection intensity, without significant increases in numbers of field plots (Bernardo and Yu, 2007; Bernardo, 2008; Asoro et al., 2011; Heslot et al., 2012; Nakaya and Isobe, 2012; Combs and Bernado, 2013; Crossa et al., 2014; Beyene et al., 2015; Bassi et al., 2016; Marulanda et al., 2016; Jonas and de Koning, 2016; Hickey et al., 2017; Goiffon et al., 2017).

While the initial interest in GS has been to increase genetic gains, plant breeders are aware that increased selection intensities are associated with faster losses of genetic potential in the founder populations (Robertson, 1960; Hill and Robertson, 2008; Bulmer, 1971).

Between the two limiting cases of no response to selection and the infeasible ideal response of realizing maximum genotypic potential among founders in a single cycle of selection, there are many possible recurrent selection response curves, two of which are illustrated in **Figure 1**. One of the curves depicts high rates of gain in the early cycles, which is favored for immediate short-term gains. However, the maximum average genotypic value approaches a limit that is less than 40% of the genotypic potential of the founders. The other curve depicts a response with slower rates than the previous one in early cycles, but with greater genotypic values before approaching a limit due to loss of genetic potential from selection. This response pattern is desirable for maximizing gains while preserving genetic variability. For a fixed set of evaluation resources, the differences between the two response curves could be due to differences in selection intensities or selection methods or both. For example, simulation studies of recurrent GS methods indicate that GS provides faster genetic gains than phenotypic selection (PS) for five to 10 cycles of recurrent selection; PS provides continued genetic gains after response to GS becomes limited (Goddard, 2009; Jannink, 2010; Liu et al., 2015). A question for the breeder is which possible curve most accurately represents the relative importance of short-term gains versus retention of valuable alleles for future generations of plant breeders. For commercial plant breeders, competition in the marketplace forces greater emphasis on maximizing immediate genetic gains. In contrast public plant breeders have an opportunity, perhaps an obligation, to place greater emphasis on minimizing the loss of useful genetic alleles while realizing genetic gains that are close to the maximum.

In spite of these general statements about the relative importance for commercial and public genetic improvement projects, each genetic improvement project has unique objectives and constraints. Previously, we (Ramasubramanian and Beavis, 2020) reported responses for combinations of selection intensity, GS methods, and training sets applied recurrently to populations composed of 2000 F_5-derived lines from contemporary soybean germplasm belonging to maturity groups II and III. The combinatorial set of factors consisted of Phenotypic Selection (PS) and four commonly used GS methods, training sets, selection intensity, number of QTL (nQTL), and broad sense heritability (H) on an entry mean basis. While interactions among all factors affected all response metrics, only the impacts of GS methods, selection intensity, and training sets are factors that plant breeders can control.

All GS methods provided greater responses than PS for at least five cycles, but PS provided better responses to selection as response from GS methods reached a limit. These results are consistent with reports by Goddard (2009); Jannink (2010), and Liu et al. (2015) that demonstrated that the full genotypic potential of the founders is eliminated more quickly with GS than PS. In terms of factors that a soybean breeder can control, a selection intensity of 1.75 and Ridge Regression Genomic Prediction (RRGP) models provided rapid response in the early cycles of selection. It also allowed the retention of genetic diversity for continued response to selection in later cycles, when the models are updated with training data from previous cycles. We also suggested that further improvements might be made if the populations were organized into families or islands and mating designs that optimize parental contributions to retain greater genetic potential in the populations are used (Ramasubramanian and Beavis, 2020). Herein, we investigate strategies that soybean breeders can employ to find optimal trade-offs between maximizing genetic gain from selection and retaining useful genetic diversity.

Given that there are constraints on the size of the breeding program, including the number of lines to evaluate and the number of field plots, it is important to reveal as many response curves as possible for possible breeding strategies. While breeders can observe these curves and identify one that most closely reflects the relative importance of the two objectives, we conjectured that it should be possible to design additional breeding strategies that are better, in the sense of minimizing the trade-offs, than those we previously investigated. One of the approaches is to use a trade-offs table to identify the best strategy for a given set of relative weights for the short-term and long-term objectives of the program.

The challenge of realizing genetic gains from selection and retaining useful genetic diversity in closed populations has been of interest since it was demonstrated that there are theoretical limits for response to selection in closed populations (Hill and Robertson, 2008; Bulmer, 1971). Trade-offs among objectives don't prohibit finding optima as long as optimality is defined as a compromise among competing objective functions (Deb, 2003;

Abbreviations: SM, Selection Method; MD, Mating Design; MP, Migration policy; PS, Phenotypic selection; HN, Hub Network; BI, Best Island; GS, Genomic selection; CR, Chain Rule; RB, Random Best; WGS, Weighted Genomic Selection; RM, Random Mating; FC, Fully Connected; IM, Island Model; GM, Genomic Mating; GA, Genetic Algorithm.

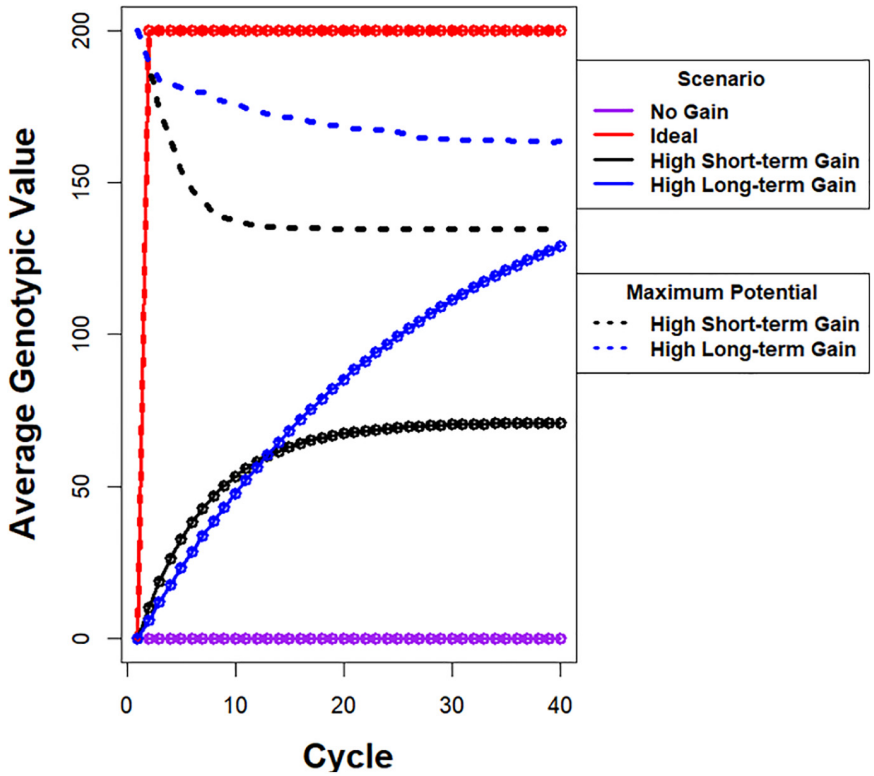

FIGURE 1 | Illustration of possible responses to recurrent selection. Average genotypic value is plotted on *y*-axis and cycle number is plotted on *x*-axis. (i) The purple line depicts a response pattern with no genetic gain, which is possible when there is no selection in a population with a large effective population. (ii) The red line depicts a hypothetical ideal response curve when the maximum genotypic potential among founders is realized in one cycle of selection. The ideal is not feasible using selection if the alleles responsible for the maximum are distributed among more than two founders. (iii) The black line depicts a response pattern with high rates of gain in the early cycles, but the maximum average genotypic value approaches a limit that is less than 40% of the genotypic potential among the founders. The dashed black line represents the corresponding rapid drop in genotypic potential among the founders. (iv) The blue line depicts a response pattern with slightly slower rates of gain in early cycles relative to the black line, but achieves greater genotypic values before approaching a limit. This is due to conservation of genetic variability as represented by the dashed blue line with slower rate of decrease in genotypic potential of the population. Note that there are potentially an infinite number of unique response curves that fall between no response and an ideal response.

Konak et al., 2006; Shoval et al., 2012; Sheftel et al., 2013; Saeki et al., 2014).

Before the development of GS, quantitative geneticists working on domestic animal systems utilized mathematical programming modeling and operations research approaches to find near-optimal solutions to the challenge of assuring genetic gain and minimizing inbreeding per cycle of selection (Wray and Goddard, 1994). The first publication using operations research approaches to address multiple objectives in plant breeding was applied to the selection of multiple traits (Johnson et al., 1988). Generally, operations research approaches involve three activities: (1) define objectives using measurable metrics; (2) translate the objectives into a model consisting of objective functions, decision variables, and constraints; and (3) find an algorithm that will provide values for the decision variables resulting in optimal solutions to the model (Rardin, 2017).

If a genetic improvement project wants to assure genetic gain and retain useful genetic diversity, then there are two competing objectives for which a trade-off needs to be optimized. This represents an example of a multi-objective optimization problem (Deb, 2003, 2011; Rardin, 2017). After translating each of the objectives into an objective function, there are several strategies for finding the optimal solution (Deb, 2003). The two most commonly used strategies are known as ε-constraint and the weighted sum. The ε-constraint method consists of identifying one of the objectives, e.g., maximize genetic gain, and translate other objectives, such as minimize inbreeding, into decision variables that can be constrained in a linear, integer, or quadratic mathematical programming model (Haimes et al., 1971); in other words, translate the multi-objective optimization mathematical model into a single objective optimization model for which there exist computational algorithms capable of finding the optimum solution (Frank and Wolfe, 1956; McCarl et al., 1977; Lazimy, 1982). Framing the ε-constraint method requires definition of metrics for genetic diversity or inbreeding. In animal breeding, this method became known as Optimum Contribution Selection (Wray and Goddard, 1994; Brisbane and Gibson, 1995; Meuwissen, 1997; Grundy et al., 1998; Meuwissen et al., 2001). Subsequent to the development of GS, Optimum Contribution Selection was modified to

maximize Genomic Estimated Breeding Values (GEBVs), and the realized relationship matrix was used to constrain inbreeding in what became known as Genomic Optimum Contribution Selection (Sonesson et al., 2010; Schierenbeck et al., 2011; Woolliams et al., 2015).

The second well-established approach to a challenge is known as the weighted sum method. The weighted sum method assigns weights, $\omega_i \in [0, 1]$ and $\sum \omega_i = 1$, to each of the "i" objective functions, and an algorithm is employed to find the values for the decision variables that minimize all objective functions simultaneously (Zadeh, 1963). The weighted sum method is equivalent to the concept of selection index familiar to breeders. In this case, the selection index is composed of weighted parameters for genetic gain and inbreeding or, equivalently, genetic diversity. If genomic information is available, GEBVs can be used to maximize genetic gain and the realized relationship matrix can be used to minimize inbreeding resulting in a genomic selection index that can be calculated for all genotypes (Carvalheiro et al., 2010; Clark et al., 2013).

Both ε-constraint and weighted sum methods are referred to as preference methods (Deb, 2003) where the constraints or relative weights have been predetermined. For defined preferences, there exist exact optimization algorithms if Karush-Kuhn-Tucker (KKT) conditions are met (Karush, 1939; Kuhn and Tucker, 1951). An exact optimization solution guarantees that no other feasible solution will be a better solution for the specified set of constraints or weights. Unfortunately, it is difficult to predetermine these values because they require forecasting the relative economic values of genetic gains and retention of useful genetic diversity in terms of immediate returns and future benefits. For commercial plant breeding projects, competition in the marketplace will force much greater emphasis on maximizing genetic gains than retaining genetic diversity to maximize immediate return on investment. In contrast, public soybean breeders have an opportunity, perhaps an obligation, to retain useful genetic diversity while realizing genetic gains for quantitative traits of agronomic importance. Since each plant breeding project has unique relative trade-offs, evolutionary algorithms have been adopted to provide multiple solutions on an efficient (Pareto) frontier of solutions to competing objectives (Deb, 2003, 2011; Konak et al., 2006). Decision makers then decide which of the solutions have the appropriate relative emphasis on the competing objectives.

Genetic improvement can be viewed as single or multiple connected search strategies in genotypic space (Podlich and Cooper, 1999; Cooper et al., 2002, 2014). The single search strategy, a.k.a. global, corresponds to the selection of lines in centralized populations, where genotypes from all the sub-populations are treated as one population (Technow et al., 2021). The multiple connected search strategy, a.k.a. local, occasional and corresponds to selection of lines in multiple domains with infrequent exchange of lines. Search strategies in genotypic space inspired the development of a class of evolutionary algorithms known as genetic algorithms (GAs). GAs are based on recurrent selection of breeding populations and are often used to find computational solutions to large combinatorial problems (Goldberg, 1989; Luque, 2011).

In a canonical GA, selected solutions are pooled together into a set of solutions. Subsequently the individual solutions are randomly sampled for pairwise "matings" to create a new set of solutions for evaluation and selection. Computational analogs of mutation or recombination referred to as mutation and recombination operators, are utilized to move the population of solutions into new domains in the solution space towards global optima. The algorithm is iterated until there are no improvements in the sets of solutions. Inspired by Wright's shifting balance theory of evolution, researchers developed a subclass of GAs, known as parallel Gas, that maintain structure among subsets of individual solutions and enable the subsets to independently find different solutions for different domains (Wright, 1967; Wright, 1988; Cantú-Paz, 2000; Luque, 2011; Yabe et al., 2016). The parallel GA is analogous to the concept of island model selection in genetic subpopulations. The term island refers to distinct sub-populations, where genotypes from any of the sub-populations cannot randomly mate with lines from other sub-populations due to restrictive rules for mating. However, Island Model/Parallel GAs allow for an exchange of solutions among subpopulations that are evolving in parallel. Island model GAs are also distinct from canonical GAs in terms of properties because evolution happens locally, within island, as well as globally, among islands. Island model parameters consist of number of islands, island size, selection pressure within each of the islands, numbers of migrants, migration frequency, connectedness or topology of islands, and emigration and immigration policies among islands (Whitley et al., 1999; Skolicki, 2007; Skolicki and Jong, 2007).

Rather than investigate the trade-off between objective functions, Jannink (2010) demonstrated that it is possible to retain useful genetic diversity in GS by weighting low-frequency alleles with favorable estimated genetic effects. Simulations with Weighted Genomic Selection (WGS) resulted in greater responses across 24 selection cycles of recurrent selection than unweighted GS, using RRBLUP estimated breeding values, for both low and high heritability traits. However, the initial rates of response using WGS were less than responses from the application of PS and less than GS. The response using WGS was better than the response from PS after 20 cycles of selection, but the responses relative to GS depended on the number of simulated QTL and heritability. Decay of linkage disequilibrium (LD) between marker and QTL is one of the factors that can slow responses using GS relative to PS (Hickey et al., 2014; Xavier et al., 2016), although decay of LD did not contribute to responses in the initial cycles using WGS. The rate of inbreeding per cycle is also greater with GS than with PS, whereas it is similar to PS when WGS is applied. The rate of fixation of favorable alleles is lower for WGS than GS resulting in larger numbers of cycles of genetic improvement before response to selection reaches a limit (Jannink, 2010). Efforts to balance the response in early cycles and later cycles have included addition of parameters to WGS (Sun and VanRaden, 2014) and dynamic weighting of rare alleles depending on the time horizon for the breeding program (Liu et al., 2015). Low-frequency favorable alleles are given greater weights, drawn from a beta distribution, in initial cycles, and the weights tend toward unity as the number of cycles of selection

approaches a predefined time horizon. This shifts the balance towards retaining greater genetic variance in earlier cycles.

Investigations of GS, WGS, genomic optimum contribution selection, and genomic selection index assume that selected individuals will be randomly mated. Typically, plant breeders do not randomly mate selected genotypes, rather, most use selected genotypes that exhibit the most desirable selection metrics, e.g., GEBVs, to serve as "hub" parents in networked crossing designs (Guo et al., 2013, 2014). Because the metaphor of hubs with spokes represents the preference for crossing most selected lines to a few "hub" lines, we refer to this mating design as a Hub Network (HN), and this is the mating design used in our previous investigation (Ramasubramanian and Beavis, 2020). A Hub Network (HN) mating design applies greater weights to genetic contributions from hub genotypes resulting in amplified loss of genetic diversity relative to Random Mating (RM) by reducing the effective population size.

As soybean breeders have become aware of the potential impacts due to loss of genetic diversity from use of GS, they have used various *ad hoc* methods to avoid crosses between related genotypes (Diers, Graef, Lorenz, Cianzio, Singh, Byrum, Xu personal communications). After quantitative geneticists working on animal breeding systems demonstrated that it is possible to use the genomic selection index strategy with an evolutionary algorithm to identify optimal pairs of mates (Kinghorn, 2011; Pryce et al., 2012; Woolliams et al., 2015), plant quantitative geneticists developed and investigated various versions of genomic selection index and genomic optimum contribution selection for plant breeding (Akdemir and Sánchez, 2016; De Beukelaer et al., 2017; Cowling et al., 2017; Lin et al., 2017; Gorjanc et al., 2018; Allier et al., 2019a,b). Notice that the computational demand to find the optimum on the non-decreasing efficiency frontier created by all possible constraint values or relative weights in all N choose 2 (NC2) mating pairs is particularly well suited for application of GAs. Also, it should be noted that Akdemir and Sánchez (2016) referred to their implementation of genomic optimum contribution selection as efficient GS. In addition to evaluating traditional PS, GS, and genomic optimum contribution selection, Akdemir and Sánchez (2016) proposed and evaluated a novel mathematical programming model, referred to as genomic mating (GM). They formulated the problem as minimizing a linear function of inbreeding plus a negative risk function for the realized relationship matrix of N_p possible parents. Inbreeding is a function of the expected genetic diversity among N_c progeny from the N_p parents and is weighted by a parameter that controls allelic diversity among all N_p parents. Risk is determined for each cross as the sum of the expected breeding values of the progeny plus the expected standard deviations of marker loci weighted by a parameter that controls the allelic heterozygosity of the relative contributions of the marker loci to the GEBVs. Thus, risk is similar to the usefulness criterion, defined by Schnell (1983) (as cited in Melchinger et al., 1988), of a selected proportion of the population and the weighting parameter reflects the breeders' emphasis of its importance. They demonstrated that their GM formulation is equivalent to an optimization problem of minimizing inbreeding subject to defined level of risk, denoted ρ. The solution needs to calculate risk and inbreeding for the range of acceptable ρ values for N_c progeny from N_p parents, i.e. (Akdemir and Sánchez, 2016) developed a Tabu-search GA to determine the efficiency frontier between inbreeding and risk. In an updated version, Akdemir et al. (2018) used a GA to find the complete set of non-dominated solutions (Deb, 2003, 2011) that comprise the efficiency frontier for the three criteria of Gain (G), Inbreeding (I), and Usefulness (U) values in the objective function. This allows the selection of a subset of solutions for evaluation, obviating the need for conducting a grid search across all possible values. More details on the GM method are provided in **Supplementary File 1**.

Akdemir and Sánchez (2016) demonstrated the utility of their genomic mating approach using simulations of recurrent selection beginning with two founders for a trait composed of simple additive genetic architecture. The QTL were evenly distributed across a simulated genome consisting of three diploid linkage groups. Their results indicated that the efficiency frontier produced responses across 20 cycles that were better than PS and as good as GS and genomic optimum contribution selection for the first five to seven cycles and better than PS, GS, and genomic optimum contribution selection thereafter (Akdemir and Sánchez, 2016). They did not include WGS for comparison in their study.

Recognizing that Island Model/Parallel GAs are very efficient at finding global optima, Yabe et al. (2016) suggested that computational island models could be used to create efficient and effective breeding plans for plant breeders. Even though computational parallel GAs allow the software developer to change mutation and recombination rates, which are not under the control of plant breeders, structures of breeding populations based on island models could offset the loss of useful genetic variability through regulation of exchange of genotypes among sub-populations. It is not unusual for plant breeders of crops that are easily self-pollinated to routinely evaluate, select, and recurrently cross lines derived from one or two specific bi-parental crosses. In the vernacular of soybean and maize breeders, this is known as "working a population." Yabe et al. (2016) demonstrated that GS on populations organized as islands provided greater response to selection than GS on a single population comprising all the islands, after the 12th of 20 cycles of recurrent GS. Their founder population consisted of lines derived from *in silico* crosses of six homozygous rice lines with an elite rice variety, i.e., a hub network. They isolated the six families of Recombinant Inbred Lines (RILs) for recurrent selection using GS with no or occasional exchange of selected lines among the family islands. While their results appeared to be similar to WGS, they did not compare their results with WGS. They also suggested that the trade-off between genetic gain and retention of useful genetic variance could be improved by adjusting the number and frequencies of migrants among sub-populations. We hypothesize that a breeding strategy consisting of breeding populations organized as family islands and in which crossing decisions are based on genomic mating will provide small soybean genetic improvement projects with the ability to minimize the trade-offs between maximizing genetic gain and minimizing the loss of useful genetic variability.

Within the populations organized as islands, we evaluated four migration policies, three selection methods, and four mating designs. Given that each of the factors show characteristic average response patterns with widely different rates and limits of responses, we also hypothesize that the combinations of all these factors will further increase the number of possible response curves due to the interaction among these factors. To evaluate the potential of these combinations of methods to realize genetic gains while retaining useful genetic diversity, we compare outcomes from simulated recurrent selection applied to contemporary soybean germplasm adapted to Maturity Group (MG) II and III using a set of metrics (Ramasubramanian and Beavis, 2020), which includes the standardized genotypic value (Rs), the most positive genotypic value (Mgv) among F_5-derived lines selected in cycle c, the standardized genotypic variance (Sgv), the average expected heterozygosity (Hs), and the lost genetic potential of populations based on the number of favorable alleles that are lost.

METHODS

Simulations

Initial sets of soybean lines were generated by simulating crosses of 20 contemporary homozygous lines representing the diversity of soybean germplasm adapted to MGs II and III with IA3023, a former widely grown variety adapted to MG III, to generate *in silico* F_1 progeny (Ramasubramanian and Beavis, 2020). Individual F_1s from each of the 20 crosses were self-pollinated *in silico* for four generations to generate 100 lines per family forming 2000 lines organized into 20 families with genotypic information at 4289 genetic loci (Song et al., 2017). Thus, the genetic structure of the initial simulated populations is similar to that used in the experimental SoyNAM investigation (Guo et al., 2010; Takuno et al., 2012; Song et al., 2015, 2017; Xavier et al., 2017; Diers et al., 2018).

As reported previously (Ramasubramanian and Beavis, 2020), there were 3818 polymorphic loci in the combined population consisting of 20 families with an average of 773 polymorphic loci within each of the families for the initial founding sets of lines. The variance of the number of polymorphic loci among families was ∼34, which indicates that the number of polymorphic loci is roughly similar among all families. Across the 20 families of Cycle 0 (C0) lines, average expected heterozygosity was 0.09 with an estimated variance of $4.4^{*}10^{-7}$ among families. The average estimated G_{st} value across the genome for the initial founding set of F_5-derived lines was 0.32, as determined by the "diff_stats" function in the mmod R package (Jombart, 2008; Ryman and Leimar, 2009; Jombart and Ahmed, 2011; Ramasubramanian and Beavis, 2020). Average pairwise "Fst" estimated using "pairwise.fst" in "hierfstat" R package (Goudet, 2005) among the 20 families in simulated data is 0.20. Pairwise "Fst" is a measure of population differentiation among pairs of populations based on Nei's genetic distance, which is estimated as the ratio of difference between the weighted average of the expected heterozygosity of pairs of families and total expected heterozygosity of the pooled populations to total expected heterozygosity of the pooled populations. For two populations "A" and "B" of size n_A and n_B, expected heterozygosity (averaged over loci) is denoted as $H_{s(A)}$ and $H_{s(B)}$, respectively. Let H_t denote the expected heterozygosity of a pooled population of "A" and "B." Then, the pairwise F_{st} between "A" and "B" is computed as: $\mathrm{Fst(A,B)} = \frac{\mathrm{Ht} - \left(\frac{\mathrm{nAHs(A)} + \mathrm{nBHs(B)}}{(nA+nB)}\right)}{\mathrm{Ht}}$ (Goudet, 2005). For comparison, the average F_{st} using genotypic data from the SoyNAM project among 40 families is 0.09 with a maximum pairwise Fst of 0.15 and a minimum Fst of 0.007 (Ramasubramanian and Beavis, 2020), whereas the average Fst among the clusters in the USDA soybean germplasm collection is 0.23 (Song et al., 2015; Xavier et al., 2018).

Combinations of Factors

We evaluated 60 combinations of factors (**Table 1**) that could influence responses to recurrently selected populations derived from a set of founder genomes representing the diversity of contemporary soybean germplasm adapted to MG II and III in North America (Mikel et al., 2010; Diers et al., 2018). The factors included structure of breeding populations, selection method, and mating design. The structure of the breeding populations, which refers to the presence of distinct sub-populations, included retaining the structure of the original 20 founder families through restrictive breeding rules, referred to as family islands, and dissolving family structures after the initial founder population was created, referred to as centralized populations. For comparison with previous studies, the centralized population structure corresponds to the bulked population in Yabe et al. (2016) and the centralized policy in Technow et al. (2021). The family island structure with migration of lines among islands is also called as distributed policy in Technow et al. (2021), whereas islands that are not connected to each other are called as isolated in Technow et al. (2021) and is the policy termed as discrete selection in Yabe et al. (2016).

Previously, we demonstrated that the development of homozygous lines for phenotypic evaluation would limit the numbers of segregating linkage blocks with effective QTL effects. Our evaluations of responses with 40, 400, and 4289 QTL showed that responses for 400 QTL followed a pattern that facilitated the study of the impact of factors on short-term and long-term responses, as the responses realized limits around 15–20 cycles, whereas for 40 and 4289 QTL, the responses reached limits within 10 cycles and around 30 cycles, respectively (Ramasubramanian and Beavis, 2020). Consequently, we chose to designate only 400 polymorphic marker loci as simulated QTL. The QTL were distributed uniformly among the SNP loci and each contributed equal additive effects of ± 0.5 units to the total genotypic value of a line. Thus, cycle C0 lines derived from the founders had an average genotypic value of zero and the potential to create genotypic values ranging from −200 to +200. Phenotypic values were simulated by adding non-genetic values sampled from N $(0, \sigma_e^2)$ distribution to the simulated genotypic values, where σ_e^2 that corresponds to non-genotypic variance was determined by the heritability ($\sigma_e^2 = ((1-H)/H)^{*} \sigma_g^2$), where σ_g^2 corresponds to genotypic variance and H corresponds to broad sense heritability. Herein we report only simulated broad sense heritability values

TABLE 1 | Treatment design representing the factors that impact responses and limits of responses that were investigated.

Factors	Levels	Values for levels
Population type	2	Centralized and Island populations
	Island model selection	
Migration frequency	1	Migration frequency of 2 corresponds to migration of lines every other cycle of selection
Migration size	1	Migration of 2 lines per migration event (20%)
Migration policy	4	(i) Isolated selection (IS) (For IS, migration frequency, size and direction are set to "0") (ii) Best island (iii) Random best (iv) Fully connected
Migration direction	1	(i) Bi-directional
	Factors common to Non-island and Island populations	
Selection method	3	(i) Phenotypic selection, (ii) Genomic selection, (iii) Weighted genomic selection
Mating design	4	(i) Hub network (ii) Chain rule (iii) Random mating (iv) Genomic mating
Genetic model parameters	1	400 QTL and 0.7 H
Total number of combinations of treatment factors	60	
Total number of simulations	**5 (replicates/combination of factors)**	**300**

on an entry mean basis of 0.7. The non-genetic variance was held constant across subsequent cycles of selection. Thus, heritability is expected to decline with every cycle of selection due to loss of additive genetic variance relative to a constant non-genetic variance.

Phenotypic selection (PS), genomic selection (GS), and weighted genomic selection (WGS) were applied recurrently to both population structures. Recurrent selection applied to the centralized populations consisted of ranking all lines in a given cycle (**Figure 2**) according to the selection metric and retaining 10% for crossing to create the next cycle of lines. In terms of standardized selection differential, this corresponds to selection intensity, $\iota = 1.75$. For selection of lines organized into family islands, 10% of the lines are selected within islands (**Figure 3**). Subsequently, 20% of lines might be migrants from other family islands depending on migration rules (**Table 1**). Metrics used for selection include simulated phenotypic values for PS, genome estimated breeding values (GEBVs) for GS, and weighted genome estimated breeding values for WGS. We used the weighting function used by Jannink (2010) for estimating weighted genome estimated breeding values (**Supplementary Table 1**). A previous study indicated that among GS methods, Ridge Regression (RR) provided the best compromise between short-term and long-term responses (Ramasubramanian and Beavis, 2020); thus, we only used RR to train GP models for GS. RR was implemented with a method that employs Expectation Maximization to obtain Restricted Maximum Likelihood Estimates of marker effects (Xavier, 2019).

For both GS and WGS, the training models were updated every cycle of selection with data sets from all prior cycles. Since average within-family prediction accuracies are less than prediction accuracies from combined training sets comprising F_5-derived lines from across all the families (Ramasubramanian and Beavis, 2020), we used a training set comprising F_5-derived lines from all the families. Training sets for each cycle were obtained by randomly sampling 1600 lines from the set of 2000 lines in each cycle. The most accurate predictions and maximum genetic responses were obtained with training data that are cumulatively added every cycle. For purposes of this manuscript, model updating refers to retraining the model with data from the current cycle as well as all prior cycles that were cumulatively added.

Subsequent to selection, four mating designs were applied to create the next cycle of lines (**Table 1**). To simulate theoretical truncation selection, selected lines were randomly mated (RM). The chain rule (CR), a.k.a., a single round-robin mating design (Yabe et al., 2016), is an alternative to RM that assures all selected lines contribute to the subsequent cycle of evaluation and selection. In contrast to the attempt to assure equal representation of selected lines through RM and CR, most

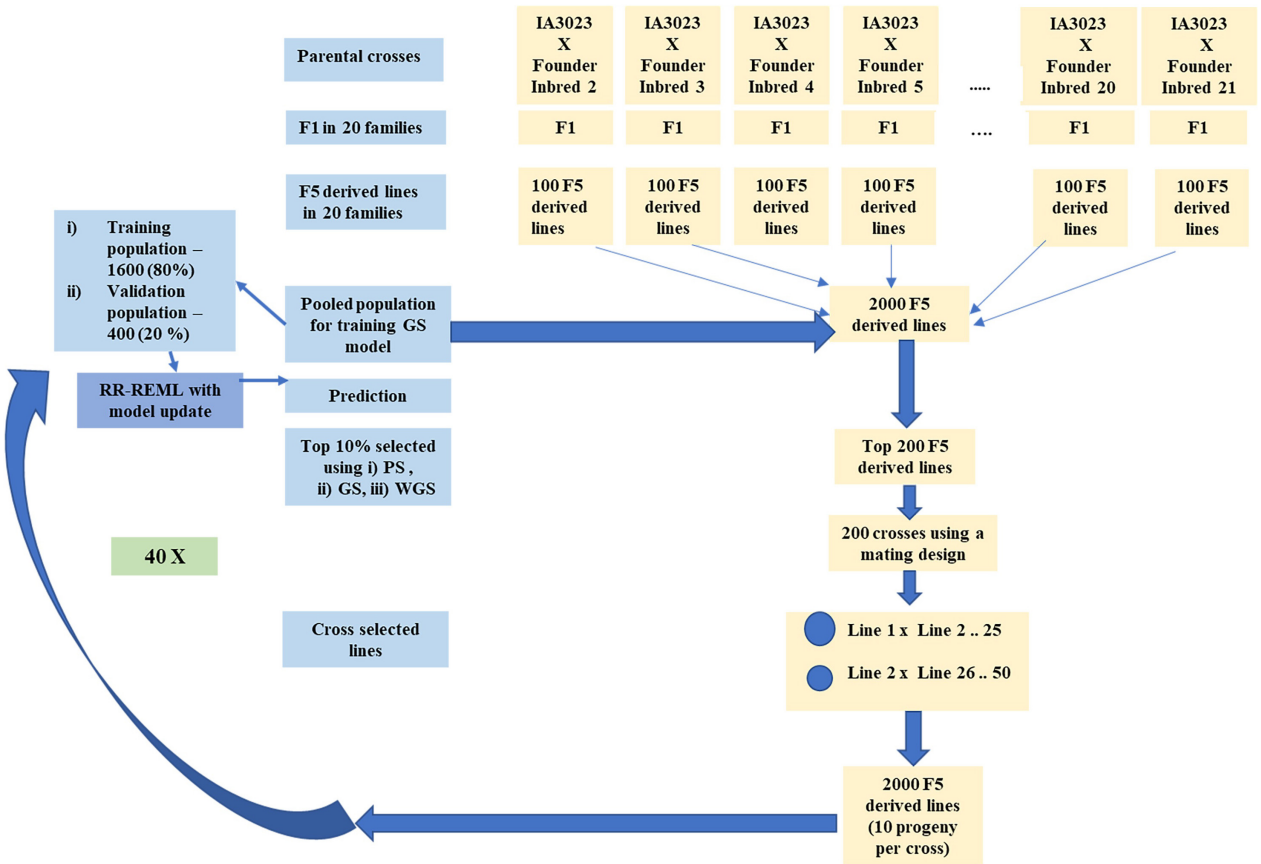

FIGURE 2 | Schematic representing simulated recurrent selection in centralized populations comprising 20 families. The schematic depicts the *in silico* steps used to generate the base population of 2000 F_5 derived lines derived from 20 founder lines crossed to IA3023. The depiction includes the model training step and the recurrent steps of prediction, sorting, truncation selection, crossing, and generation of 2000 F_5-derived lines for each cycle as well as the decision steps to check if the training set should be updated and if the recurrent process should be continued for another cycle.

soybean breeders use a mating design that assures most progeny will be derived from crosses of a few lines that exhibit the most desirable performance (Guo et al., 2013, 2014). In the hub network (HN) mating design, the hub parents with the largest selection metrics make the largest parental contributions (Ramasubramanian and Beavis, 2020). The fourth mating design, genomic mating (GM), uses mathematical objective functions to assure that defined breeding objectives are used to identify pairs of crosses from among the selected lines. GM method was implemented using the "Genomic Mating" R package (Akdemir et al., 2018). As originally described, GM combines selection and mating in a single step, but we decomposed the steps to provide comparable outcomes from all other combinations of selection methods, mating designs, and organized populations.

Genomic Mating in Centralized Families

In a selected set of 200 lines, there are 200C2 (19900) combinations of parental pairs. To solve the objective function w.r.t., an initial population of parental pairs, 250 initial populations of 200 combinations of parental pairs, is sampled from 19900 combinations (19900C200) for the genetic algorithm to solve.

Genomic Mating in Populations Organized as Family Islands

In island selection, ten lines are selected from each of the 20 family islands. Within each island, 45 (10C2) combinations of parental pairs are possible (**Supplementary Figure 1**). To solve the objective function w.r.t., an initial population of parental pairs, 250 initial populations of 10 combinations of parental pairs, is sampled with replacement to keep the population size equal to the centralized populations for the GA. For each of the 20 families, the GA is applied to the initial subset of 250 out of all possible combinations (45C10). The other parameters for the GA are the same for both centralized and family island populations. The genetic algorithm selects non-dominated elite solutions (Deb, 2003, 2011) and crosses non-dominated elite solutions for 50 iterations with a mutation probability of 0.8 (**Supplementary Figure 1**). Examples of pseudocode are provided in Akdemir and Sánchez (2016) and the "Genomic Mating". It is important to note that the parameter values in the genetic algorithm can be optimized, and the set of solutions in the pareto-front can be explored for better fsolutions using other methods such as NSGA-II, NSGA-III, SPEA-1,

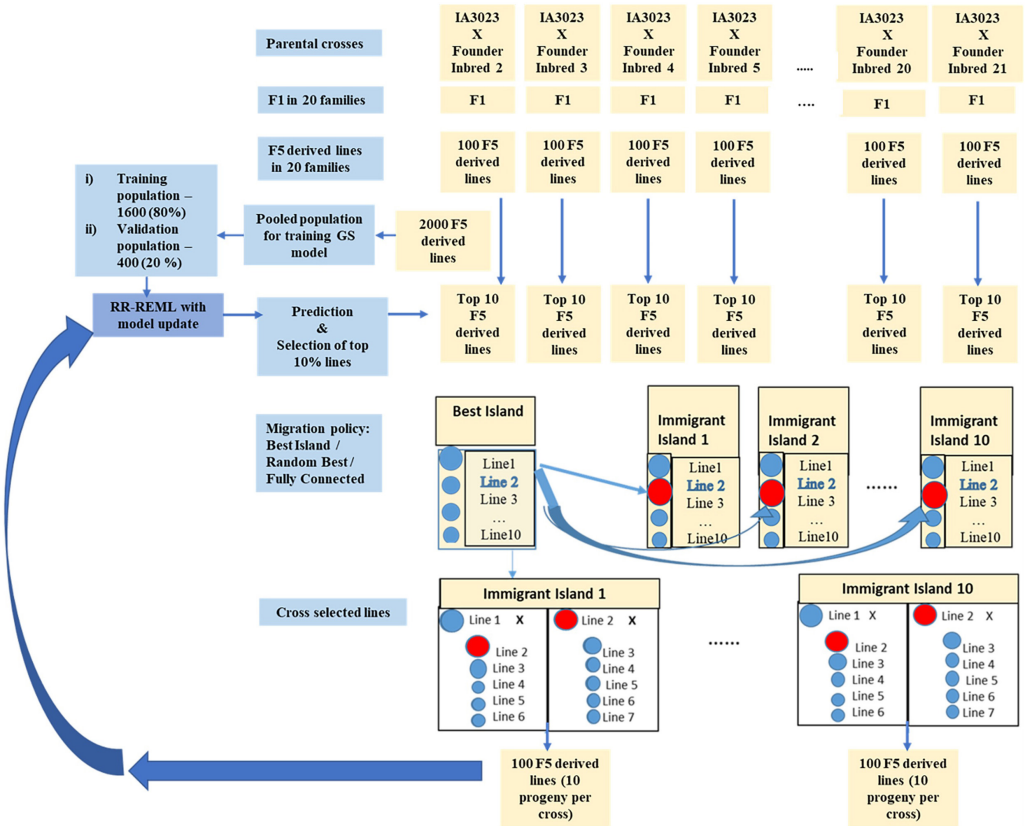

FIGURE 3 | Schematic representing simulated recurrent selection of family island populations where each of the 20 families from the founders is considered an island population. The schematic depicts the *in silico* steps used to generate the base population of 2000 F_5 derived lines derived from 20 founder lines crossed to IA3023; 100 F_5-derived lines generated from each of the crosses form a distinct island. The depiction includes the model training step and the recurrent steps of prediction, sorting, truncation selection within islands, migration, crossing, and generation of 100 F_5-derived lines per island for each cycle as well as the decision steps to check if the training set should be updated and if the recurrent process should be continued for another cycle. The blue shaded circles represent lines that are descendants of the founder populations in the islands and red shaded circles represent lines that are replaced by immigrants from the island with the largest genotypic value for the "Best Island" policy.

SPEA-2 and other recent improved versions of GAs for better convergence rate and quality of solutions, determined by the proximity to global optimum (Deb, 2011; Seada and Deb, 2018; **Supplementary Figure 1**).

Migration Rules Among Family Islands

In addition to applying selection methods and mating designs to both population structures, there are many possible rules that affect migration among islands. Migration rules that were implemented in a preliminary investigation included: (1) frequency of migration—never, once every two cycles, and every cycle of recurrent selection; (2) the proportion (10% and 20%) of immigrants that will be included in crosses responsible for creating the next cycle of lines; (3) migration can be either in one direction or it can be reciprocal among family islands. Based on the preliminary investigation (results available on request), we decided to set the migration rule as bi-directional migration between both immigrant and emigrant islands of two lines once every other cycle of selection.

Migration Policies Among Family Islands

Migration policy (MP) refers to the nature of island topology specifying connections between emigrant and immigrant islands. The four levels for migration policy included "Isolated" (IS), "Best Island" (BI), "Random Best" (RB), and "Fully Connected" (FC). For the BI policy, emigrant lines are selected from the island with most desirable average genotypic value in the islands, and selected lines can migrate to no more than 10 islands. Given a bi-directional migration rule, the emigrant island also receives two immigrants from the islands that received the emigrants. For an RB policy, an emigrant island is selected randomly from a set of 10 islands with high average genotypic values, while the migration pattern itself is similar to BI policy. For the FC policy, every island is connected to every other island, and lines migrate from emigrant islands with high values to randomly selected immigrant islands (**Supplementary Figure 2**).

Note that migration factors are irrelevant for populations that did not maintain the structure of family islands, and they are

irrelevant for isolated family islands where there is no migration. Thus, the treatment design is not a complete factorial, rather, the complete set is comprised of responses for 60 combinations of factors with five independent replicates per combination of factors. The parameter values for levels of island selection specific factors were selected based on limits of responses from a larger set of simulations (2664 combinations of factors with 10 replicates per combination of factor) performed for a preliminary study. The migration rules investigated included migration of one or two lines every cycle, or every other cycle or once in three cycles in one or both directions. Mating designs included (HN, CR, and RM) for 40, 400 and 4289 QTL with 0.7 and 0.3 H (response patterns for this set are available on request).

Modeled Response to Recurrent Selection

The averaged genotypic value for each cycle, c, of recurrent selection was modeled with a linear first-order recurrence equation:

$$f_0(c)y_{(c+1)} + f_1(c)y_{(c)} = g(c) \quad (1)$$

where c is a sequence of integers from 0 to 39 representing each cycle of recurrent selection from cycle 1 to 40 and f_0, f_1, and g are constant functions of c. By rearranging the equation, we note that the response in cycle c+1 can be represented as

$$y_{(c+1)} = -\frac{f_1(c)}{f_0(c)}y_{(c)} + \frac{g(c)}{f_0(c)} \quad (2)$$

Since the ratios $f_1(c)/f_0(c)$ and $g(c)/f_0(c)$ are constants, we can represent the response in cycle c+1 as

$$y_{(c+1)} = \alpha y_{(c)} + \beta \quad (3)$$

If y_0 specifies the average genotypic value of the first cycle of F_5 lines derived from crosses involving IA3023 and the other founders, then (3) has a unique solution (Goldberg, 1958; Ramasubramanian and Beavis, 2020):

$$\begin{aligned} y_c &= \alpha^c y_0 + \beta\tfrac{1-\alpha^c}{1-\alpha} \text{ if } \alpha \neq 1 \\ y_c &= \alpha^c y_0 + \beta c \quad \text{if } \alpha = 1 \end{aligned} \quad (4)$$

An alternative representation of (eqn 4) for the situation of $\alpha \neq 1$ is

$$\begin{aligned} &y_c = \alpha^c(y_0 - y') + y' \\ &\text{with } y' = \tfrac{\beta}{1-\alpha} \end{aligned} \quad (5)$$

where α is less than 1 for genotypic response to recurrent selection and y' represents the asymptotic limit to selection (Goldberg, 1958; Ramasubramanian and Beavis, 2020). An illustration of the values of the sequence of c = 0–39 for a range of α and β values can be found in our previous study (Ramasubramanian and Beavis, 2020). The model-derived curves can be interpreted as response to selection as a function of the frequencies of alleles with additive selective advantage, selection intensity, time, and effective population size (Robertson, 1960). The parameters, α, and β were estimated with a non-linear mixed effects method implemented in the "nlme" and "nlshelper" packages (Pinheiro and Bates, 2000; Baty et al., 2015; Pinheiro et al., 2021).

Since the limits of responses are approached asymptotically, the number of cycles required to reach half of the limits before there is no longer response to selection is referred to as the half-life of the recurrent selection process (Robertson, 1960; Dempfle, 1974; Cockerham and Burrows, 1980; Kang and Namkoong, 1980; Kang, 1983; Kang and Nienstaedt, 1987). From the first-order recurrence equation (5), the half-life is estimated as

$$t_{1/2} = \ln(0.5) \,/\, \ln(\alpha) \quad (6)$$

when y_0 is "0" and the asymptotic limit is estimated as y' (Ramasubramanian and Beavis, 2020).

Analyses of Variance (ANOVA) of Modeled Response to Recurrent Selection

Analyses of variance is used to evaluate the impact of factors and their interactions on the modeled responses to global and island recurrent selection. The analyses of variance used single-level nlme models with modeled (Eqn 5) responses grouped by combinations of treatment factors. We analyzed the variance among modeled responses using AIC, BIC, and Likelihood metrics that were grouped based on combinations of treatment variables consisting of population type, selection method, mating design, and migration policy for migration frequency, migration size, and migration direction for one genetic model consisting of 400 simulated QTL responsible for 0.7 H with equal additive effects (**Table 1**). For a discussion of the analyses of variance using non-linear mixed effects models refer to (Pinheiro and Bates, 2000; Zuur et al., 2009; Baty et al., 2015; Pinheiro et al., 2021; Oddi et al., 2019; Ramasubramanian and Beavis, 2020).

In the first phase of model fitting, we fit a random intercept model for estimating both α and β in the recurrence equation using the "nlme" R package. Estimates of modeled parameters from nlsList models were retained as starting values for fixed effects. Multiple ANOVA of "nlme" objects representing the models were used to identify combinations of factors with significant impacts on the non-linear response. The model with the lowest AIC score was selected as the best model. The best random intercept model in the first phase of model fitting process M15 and models with combinations of three factors (M11-M14) showed evidence of auto-correlation among residuals. Since auto-correlation violates the independence assumption, the correlation among residuals was modeled using AR-1 correlation structure. Since the genotypic values across cycles in recurrent selection are correlated, fitting AR-1 correlation does not remove the correlation unless cycles are used as co-variates. However, using cycles as a co-variate makes the model fitting process very time-consuming and often has larger AIC scores than models without cycles as covariates. The Model M15 with AR-1 correlation structure was further

refined by modeling variance components using "varIdent" structure in "nlme." The process for fitting, selecting, and refining mixed-effects models is similar to our previous study (Ramasubramanian and Beavis, 2020) and is described in **Supplementary File 2**.

Evaluations of Responses to Recurrent Selection

Evaluations of responses to recurrent selection were conducted on both modeled and genotypic values using a set of metrics described in Ramasubramanian and Beavis (2020) and defined below. The estimated population half-life and asymptotic limits used the estimated parameters α and β of the first-order recurrence model. The average genotypic values were used to estimate the standardized genotypic value (Rs) and maximal genotypic value (Mgv). Maximum possible genotypic potential of the founders provided a reference for number of favorable alleles retained in the population. The loss of genotypic potential is characterized by reduction in the standardized variance of genotypic values (Sgv) and estimated heterozygosity (Hs). In addition, efficiency of conversion of loss in genotypic variance into genetic gain (Rs_var) provides a way to assess gain in genotypic value and loss of genetic variance simultaneously. For selection using island models, the different impacts of selection strategies on the genotypic variance at individual island or global levels are assessed using intra-island Sgv, inter-island, and global variance of genotypic values. A schematic diagram of the processes, factors, and evaluation metrics used to characterize the responses to recurrent selection is provided in **Figure 4**.

Evaluation Metrics

The standardized genotypic value, R_s was estimated in every cycle of selection as the proportion of maximum genotypic potential (200 units) relative to the average genotypic value of 2000 lines in C0 (Eqn 7). Values range from 0 to 1 with the value of 1 corresponding to the maximum possible genotypic value with the genetic model and 0 corresponding to the average genotypic value of C_0 (Meuwissen et al., 2001; Liu et al., 2015; Ramasubramanian and Beavis, 2020).

$$Rs = \frac{Rc}{(R_m - R_0)} \quad (7)$$

R_s - Standardized genotypic value

R_0 - Average genotypic value of F5 derived lines produced by founders

R_c - Average genotypic value in cycle 'c' – R_0

R_m - Maximum possible genotypic value (200)

Since we previously evaluated the genetic improvement of soybean using PS and the HN mating design in centralized populations, we used PS with a selection intensity of 1.75 for the centralized population and HN mating design (designated as CE-PS-HN) as a reference for comparing novel combinations of selection and mating designs proposed in the study. A standardized relative genotypic response, ΔRs_c, is calculated in equation (8) as the percentage of the difference in standardized genotypic values, Rs_c, in each cycle c.

$$\text{Percent Gain in } Rs_{c\ (Design-x)} = \frac{Rs_{c\ (Design-x)} - Rs_{c(CE-PS-HN)}}{Rs_{c(CE-PS-HN)}} * 100 \quad (8)$$

$Rs_{c(Design-X)}$ - standardized response for Design-x in cycle 'c'

$Rs_{c(CE-PS-HN)}$ - standardized ersponse for CE-PS-HN design in cycle 'c'

The standardized genotypic variance (*Sgv*), defined as the change in estimated genotypic variance from the estimated genotypic variance of the initial population of lines from C0, was used to evaluate the changes in estimated genotypic variance across cycles of recurrent selection. Note that values for *Sgv* range from zero to one as it is standardized to the maximum genotypic variance among founders.

Efficiency of genetic improvement is a metric used to evaluate the proportion of genetic improvement that was obtained through loss of genetic diversity from recurrent selection (Gorjanc et al., 2018). Efficiency is estimated as the slope in linear regression in linear regions of response curves. However, responses to recurrent selection in the absence of mutation are inherently non-linear (Robertson, 1960; Bulmer, 1971; Hill and Robertson, 2008; Ramasubramanian and Beavis, 2020). For purposes of evaluating the relative contribution of lost genetic variance to genetic response in both linear and non-linear segments of the response curve, we introduce the standardized genotypic variance of the response, Rs_Var, calculated with Equation (9).

$$Rs\_var = \frac{G_c - G_0}{SdG_0 - SdG_c} \quad (9)$$

G_c -average genotypic value of the set of F5 derived lines evaluated in cycle 'c'

G_0 -average genotypic value of the founding set of F5 derived lines

SdG_0 - estimated standard deviation of genotypic values of founding set of F5 derived lines

SdG_c - estimated standard deviation of genotypic values of F5 derived lines in cycle 'c'

The numerator term represents the difference in average genotypic values of a population in cycle "c" from cycle "0" normalized to standard deviation of genotypic values in cycle "0." The denominator represents the difference of standard deviation of genotypic values between cycles "0" and "c" normalized to the standard deviation of genotypic values in cycle "0" (Ramasubramanian and Beavis, 2020). For the centralized populations, Rs_Var was estimated by calculating the variance of simulated genotypic values. Standardizing the estimated genotypic variance with respect to the maximum genotypic values in the initial population results in values that range from 0 to 1. For the family island populations, the genotypic variances can be split into within- and between-island genotypic variance. The three measures we used to estimate the global diversity of populations, inter-island diversity, and within-island diversity are provided in the documentation of the R package.

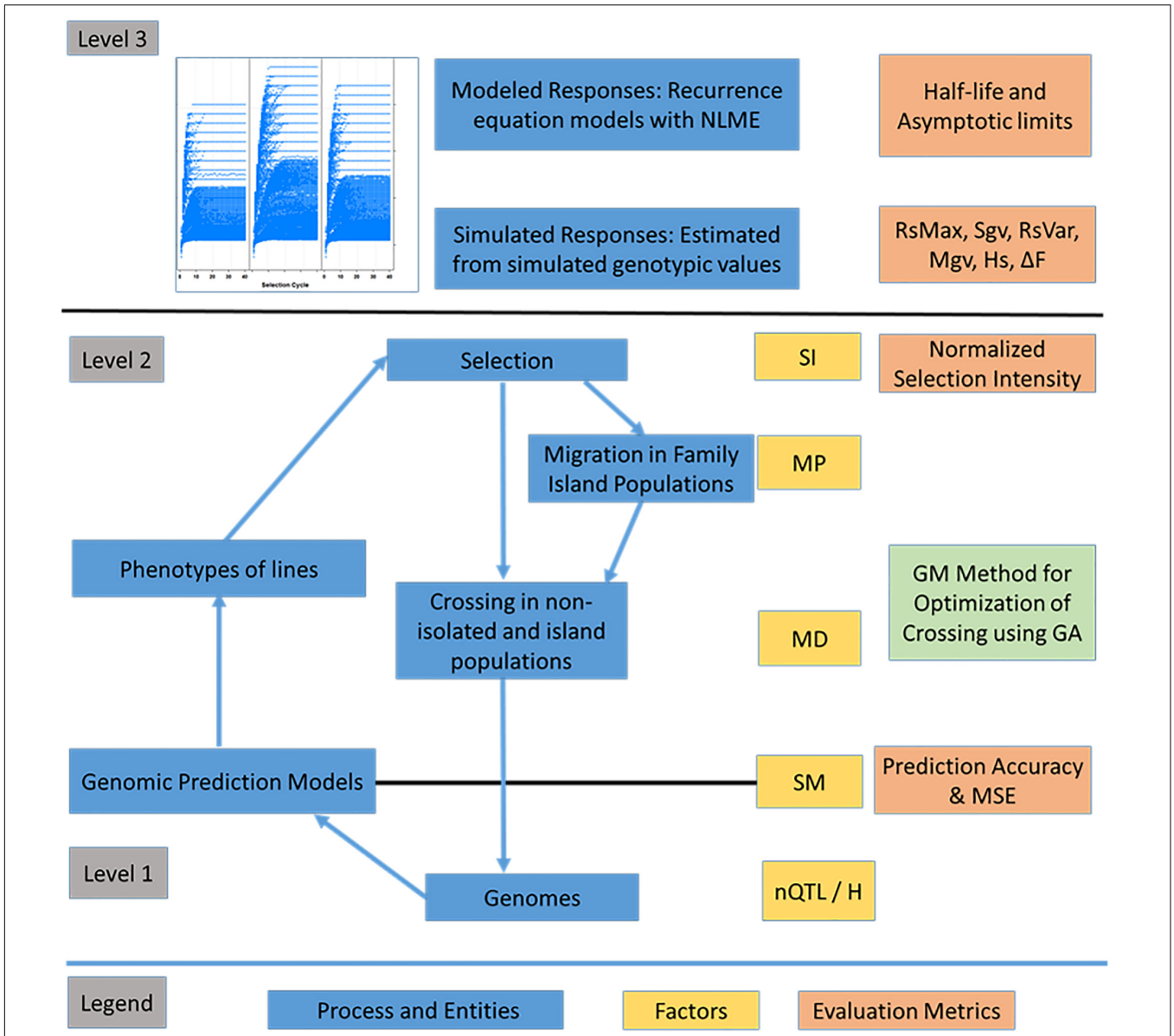

FIGURE 4 | Overview of the recurrent selection process. Representation of entities such as genomes, associated F_5-derived lines and processes such as the estimation of marker effects, selection, migration and crossing. Levels correspond to layers of information with level 1 comprising genomic information, level 2 comprising phenotypes of lines within and across family islands, and level 3 comprising higher-level information including responses across cycles of selection. The factors include nQTL and H at the genome level, selection method (SM) including phenotypic selection (PS), genomic selection (GS), and weighted genomic selection (WGS). The factors at level 2 include selection intensity (SI - top 10% selected fraction); Mating Design (MD), which includes Hub Network, Chain Rule, Random Mating, and Genomic Mating; Migration Policy (MP), which includes: "Isolated Selection," "Best Island," "Random Best," and "Fully Connected" policies. Among the MD levels, GM method involves application of evolutionary multi-objective optimization to minimize inbreeding and maximize gain and usefulness. Level 3 is characterized using evaluation metrics such as half-life and asymptotic limits derived from recurrence equation models and metrics such as Standardized Responses (Rs), Standardized genotypic variance (Sgv), Maximal genotypic values (Mgv), and efficiency of converting loss in genetic variance into gain (RsVar) derived from simulated outcomes. Other metrics include prediction accuracy and MSE for GP models (RR-REML) and expected heterozygosity (Hs).

RESULTS

Analysis of Variance of Modeled Genotypic Values

There is strong evidence from the analyses of variance (**Supplementary File 4**) that the modeled genotypic values across cycles of selection depend on interactions among selection method, mating design, and migration policy. The most parsimonious model included all combinations of factors indicating that interactions among all factors have statistically significant influences on recurrent responses to selection and requires unique estimates of α, and β in (3) for each of the

combinations of factors (M15 in **Supplementary File 4**). For all combinations of factors, we report only migration involving bi-directional migration of two migrants every other cycle. Among the factors that affect only family island populations with migration, migration frequency had significant effects on rate and the asymptotic limits for response to selection, whereas migration direction and size had relatively small effects on rates and no significant effect on the asymptotic limits for response to selection. Rates and genotypic values at the limits of response for a given selection method and mating design also depend on genetic architecture and heritability (data available on request). Rather than belabor the specific outcomes from all possible combinations of factors that affected the modeled responses, the remainder of the reported results are restricted to results from simulations with 400 QTL responsible for 70% of phenotypic variability.

Rates and Limits of Responses to Recurrent Selection

Factors common to centralized and family island populations such as mating design and selection method as well as factors specific to isolated and island model selection had significant effect on estimated population half-lives and asymptotic limits. Half-lives for selection methods on centralized populations ranged from 3.83 to 16.10 cycles with a mean of 9.62 cycles, and asymptotic limits ranged from 71.64 to 160.76 with a mean of 115.97 (58% of the maximum possible potential in the founders). Compared to centralized populations, half-lives for selection on isolated family islands were very low ranging from 1.97 to 2.89 cycles with a mean of 2.43 cycles, and asymptotic limits ranged from 28.42 to 38.30 and a mean of 33.12 (16.5% of the maximum possible potential in the founders) (**Supplementary File 5**; **Supplementary Figure 3**). Estimated half-lives for island model selection methods were on the average greater than selection methods applied to centralized populations ranging from 4.24 to 32.04 cycles with a mean of 13.45 cycles. Asymptotic limits ranged from 47.54 to 198.82 with a mean of 116.8 (58.5% of the maximum possible potential in the founders) (**Supplementary File 5**; **Supplementary Figure 3**).

Responses to Recurrent Selection of Non-island Lines

There were 12 combinations of selection methods and mating designs that were applied to lines in centralized populations. The greatest genotypic values (Rs) were attained with WGS (**Figure 5** and **Supplementary Table 2**). Genomic selection using RRBLUP estimated phenotypic values resulted in greater responses than PS in early cycles while WGS produced greater responses than PS in later cycles (**Figure 5**; **Supplementary Table 2**). Weighted genomic selection followed by the CR mating design resulted in the greatest realization of genetic potential before reaching a limit. Genomic selection using RRBLUP estimated phenotypic values followed by an HN mating design resulted in the greatest rates of response in the first ten cycles and, if followed by RM, provided the greatest responses in the first 20 cycles. When the GM design is applied to selected lines to obtain specified crosses according to optimization criteria, the responses in the first 15 cycles were larger than obtained with RM, whereas responses after the 20th cycle were less than responses for other mating designs (**Figure 5** and **Supplementary Table 2**).

The responses measured as maximum genotypic values (Mgvs) produced response patterns similar to Rs. Use of WGS followed by the CR mating design resulted in an average Mgv of 125 (62.5% of the maximum potential in the founders) followed by PS and GS using RRBLUP estimated phenotypic values in the 40th cycle. Genomic selection followed by the HN mating design (CE-GS-HN) realized greater Mgvs relative to other combinations of factors only in the early cycles (**Supplementary Figure 4**).

The rates at which standardized genotypic variance (Sgv) and expected heterozygosity (Hs) decreased depended on the mating designs (**Figure 6** and **Supplementary Figure 5**). The application of RM and CR mating designs after selection helped maintain genotypic variance and heterozygosity for use in later cycles of recurrent selection. The HN mating design resulted in the fastest loss of Sgv and Hs (heterozygosity), while the GM design demonstrated losses of Sgv and heterozygosity that were intermediate between HN and RM/CR designs.

The rate at which maximum genotypic potential decreased across cycles of selection was reflected in the estimated number of lost favorable alleles. Among the selection methods, GS using RRBLUP-estimated phenotypic values lost genetic potential faster than PS and WGS (**Figure 7**). Among the mating designs, HN resulted in the fastest loss of genetic potential while RM lost genetic potential slower than any of the other mating designs. With the GM method, genetic potential was lost at a rate that was intermediate between RM and HN mating designs. The CR design lost favorable alleles at rates that were similar to GM with GS, whereas after applying CR with PS and WGS, the loss of alleles was similar to RM (**Figure 7**).

Rates of inbreeding are larger for GS compared to PS and WGS in the first 10–15 cycles. The RM and CR mating designs demonstrated the slowest rates of inbreeding, whereas rates of inbreeding with the GM and HN mating had high rates of inbreeding before responses to selection became limited (**Supplementary Figures 6**, **7**). The estimates of genotypic responses, standardized to genotypic variance (Rs_Var), were the greatest in the first 20–30 cycles with CR, RM, and GM mating designs, while the HN mating design lost the greatest amount of Rs_Var with GS, PS, and WGS (**Supplementary Figures 8**, **9**).

Responses to Recurrent Selection of Lines Organized as Family Islands

The genotypic values when the isolated family island populations reached the limits were as much as 67% less than the values when limits were reached in the centralized counterpart populations (**Supplementary Figure 10**). Among the isolated selection methods, GS and WGS with GM design (designated IS-GS-GM and IS-WGS-GM) provided the greatest genotypic values at the response limits. Between 10% and 15% of the maximum potential in the founder populations was realized within the first 10 to 15 cycles, when there was no migration among islands

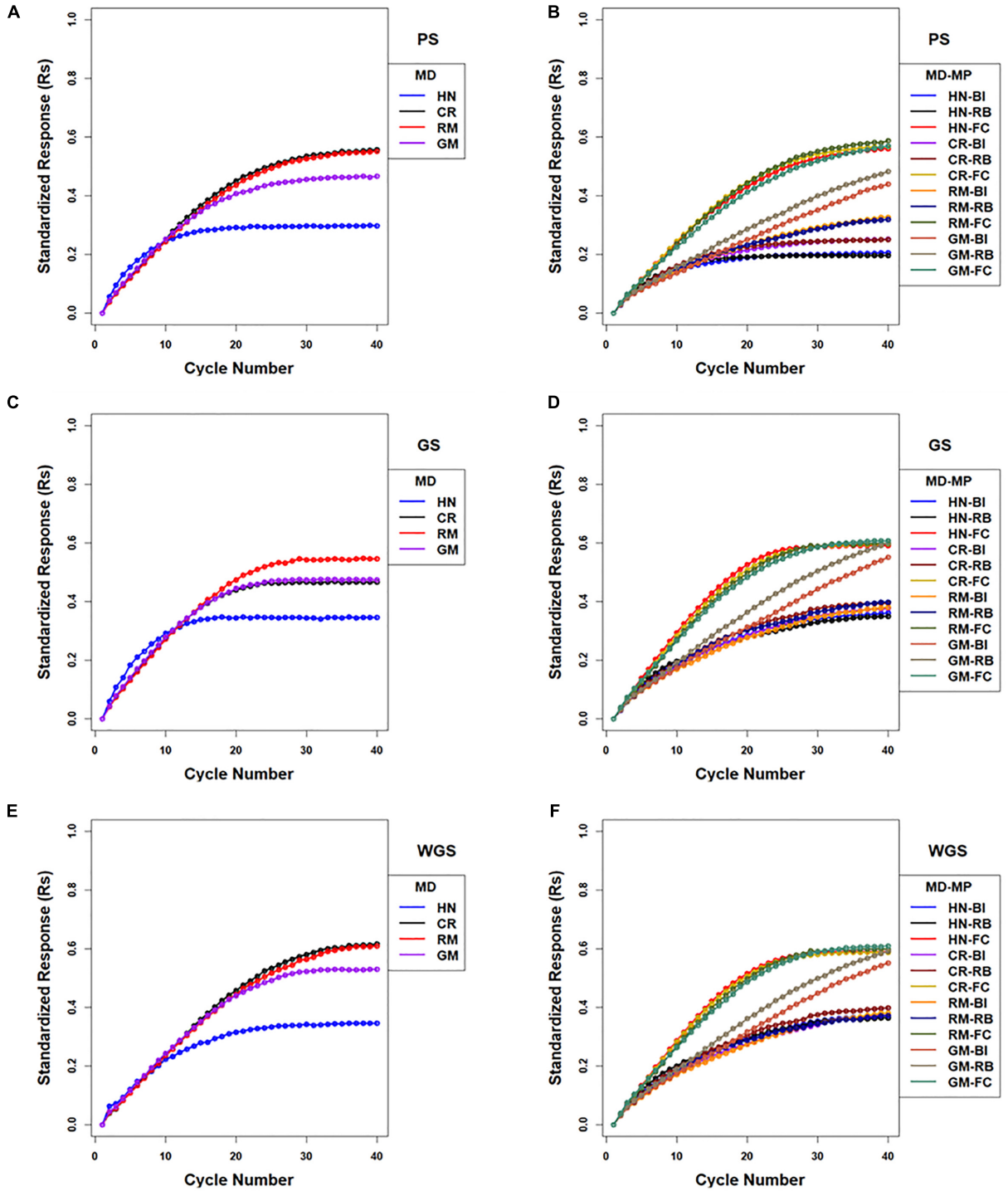

FIGURE 5 | Standardized genotypic responses (Rs) across 40 cycles of recurrent selection on centralized **(A,C,E)** and family island **(B,D,F)** populations, using Phenotypic Selection (PS) **(A,B)**, Genomic Selection (GS) **(C,D)** and Weighted Genomic Selection (WGS) for the four mating designs (MD): Hub Network (HN), Chain Rule (CR), Random Mating (RM), and Genomic Mating (GM). Standardized genotypic responses are represented from a simulated genetic architecture consisting of 400 additive QTL uniformly distributed throughout the genome and responsible for 70% of phenotypic variability. Ten percent of lines are selected to be used in crosses in HN, CR, RM, and GM designs. Migration policies (MP) included the Best Island (BI), Random Best (RB), and Fully Connected (FC) with bi-directional migrations of two migrants every other cycle. GP models are updated every cycle in GS and WGS using training sets with data from all prior cycles of selection.

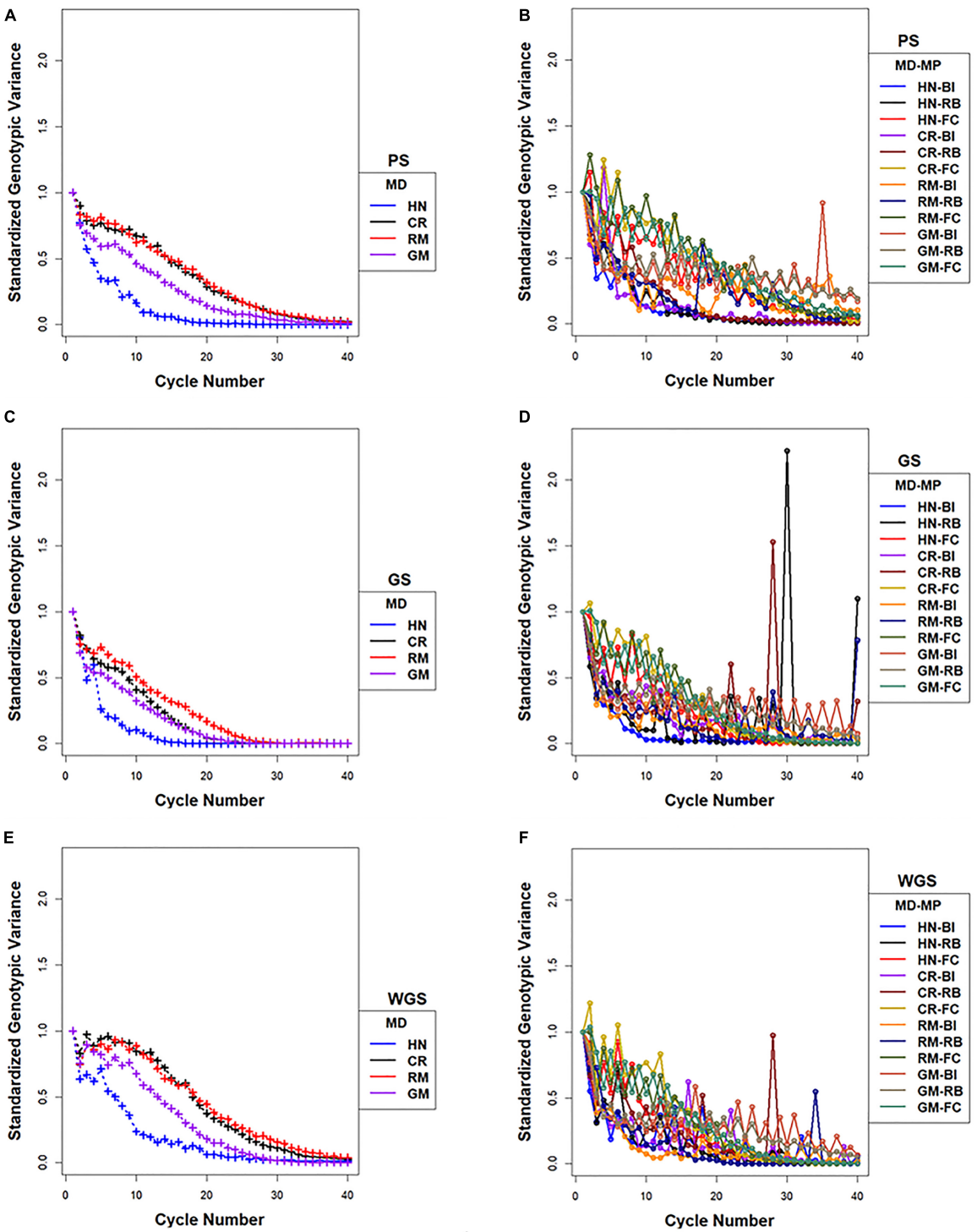

FIGURE 6 | Standardized genotypic variance across 40 cycles of recurrent selection on centralized **(A,C,E)** and family island **(B,D,F)** populations, using Phenotypic Selection (PS) **(A,B)**, Genomic Selection (GS) **(C,D)**, and Weighted Genomic Selection (WGS) **(E,F)** for the four mating designs (MD): Hub Network (HN), Chain Rule (CR), Random Mating (RM), and Genomic Mating (GM). Ten percent of lines are selected for crossing. The genetic architecture in the initial simulated founder lines consisted of 400 additive QTL uniformly distributed throughout the genome and expressed broad sense heritability on an entry mean basis of 0.7. Genetic variance is standardized to the average genotypic variance in founder populations in cycle "0." Average island genetic variance refers to genetic variance within families averaged across 20 families. Migration policy in the island models included "Best Island" (BI), "Random Best" (RB), and "Fully Connected" (FC) with bidirectional exchange of two immigrants and emigrants every other cycle of selection. GP models are updated every cycle in GS and WGS using training sets with data from all prior cycles of selection.

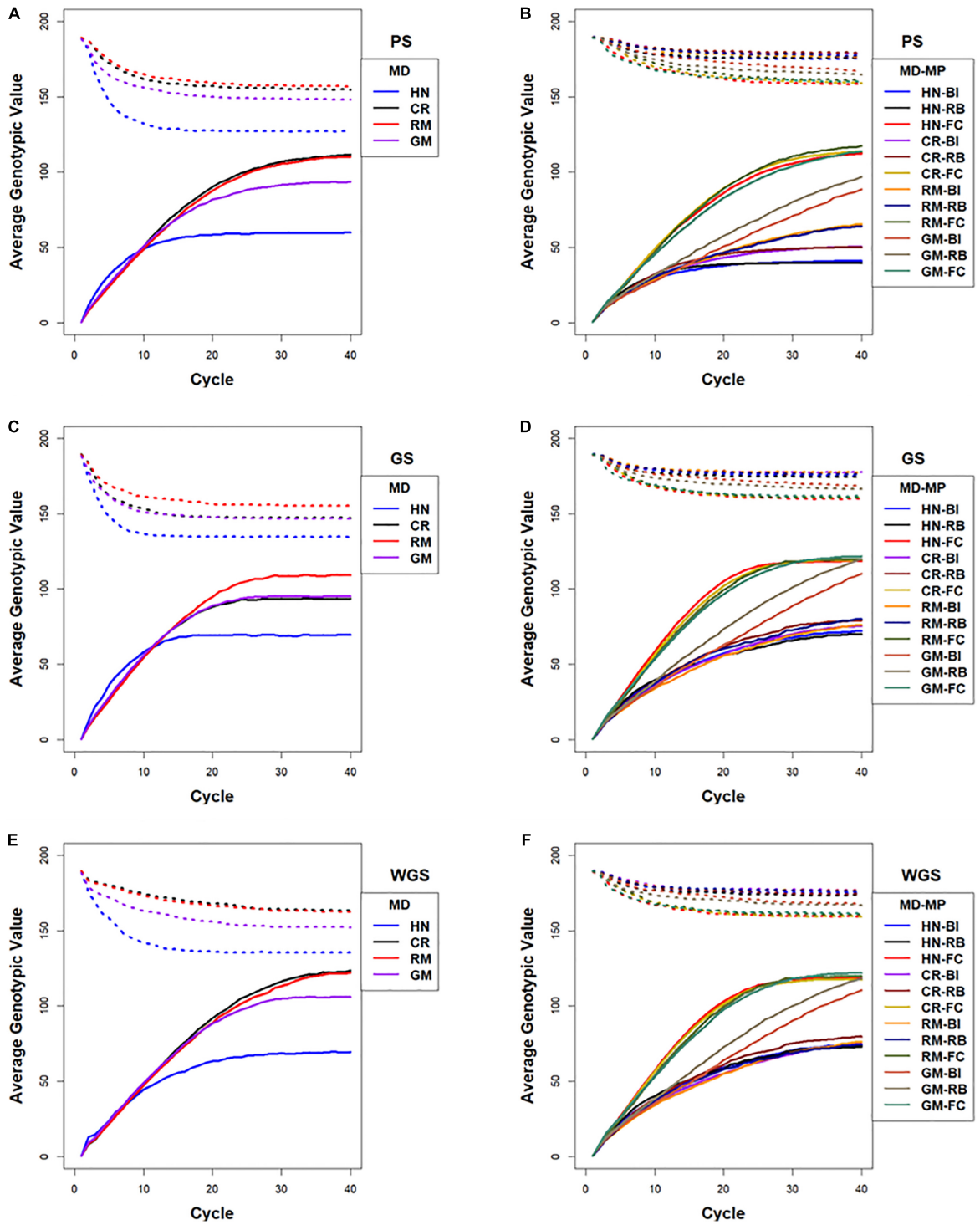

FIGURE 7 | Lost genotypic potential and average genotypic values across 40 cycles of recurrent selection on centralized **(A,C,E)** and family island **(B,D,F)** populations, using Phenotypic Selection (PS) **(A,B)**, Genomic Selection (GS) **(C,D)** and Weighted Genomic Selection (WGS) **(E,F)** and four mating designs (MD): Hub Network (HN), Chain Rule (CR), Random Mating (RM), and Genomic Mating (GM). Ten percent of lines are selected for crosses using HN, CR, RM, and GM mating designs. The genetic architecture in the initial simulated founder lines consisted of 400 additive QTL uniformly distributed throughout the genome and expressed broad sense heritability on an entry mean basis of 0.7. The dotted lines represent maximum genetic potential estimated from favorable alleles that are lost from the population, and solid lines represent increase in average genotypic value of populations due to recurrent selection. Migration policies (MP) in the island models included "Best Island" (BI), "Random Best" (RB), and "Fully Connected" (FC) with bidirectional exchange of two immigrants and emigrants every other cycle of selection. GP models are updated every cycle in GS and WGS using training sets with data from all prior cycles of selection.

(**Supplementary Figure 10**). Maximal genotypic values (Mgvs) followed a pattern similar to Rs, and Sgvs mirrored the response pattern in IS (**Supplementary Figure 10**).

In contrast to selection on isolated family islands, genotypic values at the limits were larger using BI, RB, and FC migration policies among islands, where there is exchange of lines. Among the selection methods applied to the family island populations, GS and WGS realized the greatest genetic potential before reaching limits of responses (**Figures 5**, **7**). The impacts of mating designs on the responses to selection applied to family island populations are distinct from those on centralized populations. In the centralized populations, RM and CR mating designs provided the greatest genotypic values before response to selection became limited, whereas in the family island populations, GM provided the greatest genotypic values when coupled with BI and RB migration policies. The FC migration policy with the largest migration rates produced responses that were similar among the HN, CR, RM, and GM designs (**Figures 5**, **7**).

As noted above, the best responses to selection in the first 10 to 20 cycles on centralized populations were obtained using GS followed with an HN or GM mating design (respectively designated CE-GS-HN and CE-GS-GM in **Figure 5**). The greatest short-term responses to selection in family island populations were obtained using either GS or WGS followed by the HN mating design coupled to a FC migration policy (IM-GS-HN-FC and IM-WGS-HN-FC in **Figure 5** and **Supplementary Table 2**). The gains in the first 10–20 cycles that were obtained using GS and WGS followed by the GM design coupled to a FC migration policy were comparable and showed little difference.

Given the FC migration policy, the largest standardized genotypic responses at the limits to response (0.59–0.61) were obtained using GS or WGS with HN, CR, RM, and GM designs, whereas with RB migration policy, GS and WGS followed by the GM design produced the greatest realization of genetic potential before the 40th cycle (0.59–0.6) compared to (0.3–0.4) with HN, CR, and RM designs (**Figure 5** and **Supplementary Table 2**). The BI policy showed a pattern similar to that of RB, but at a slower rate of response (**Figure 5** and **Supplementary Table 2**).

Maximum genotypic values followed a pattern similar to Rs for most of the island selection methods. In contrast to selection in centralized populations where PS and WGS resulted in the greatest Mgvs in 20–40 cycles, GS in family island populations resulted in larger Mgvs (124.6) than island PS (119.9) by the 40th cycle.

Rates of decrease in maximum available potential are influenced by factors such as selection intensity, selection method, and mating design. Relative to centralized populations, island selection retains allelic diversity in the combined population as selection depletes variance only within islands and not across islands (**Figure 7**). Such loss in maximum potential is not always reflected in rates of responses. Relaxed selection intensity will result in retention of genetic variance with no significant increase in response as it is observed with BI and RB migration policies when combined with RM designs for PS, GS, and WGS.

Island selection with GM design and FC migration policy showed the least rate of decrease of Hs values for PS, GS, and WGS reflecting a greater potential retained in the population followed by island selection with GM design and RB migration policy (IM-GM-RB) as well as island selection with GM design and BI migration policy (IM-GM-BI). Island selection with HN design and BI policy (IM-HN-BI) as well as RB policy (IM-HN-RB) showed the most rapid decrease in Hs across 40 cycles of selection, whereas CR and RM designs with the same RB and BI migration policies showed intermediate rates of decrease in Hs. There is an oscillatory pattern in the decrease of Hs, where Hs increased with every migration event in early cycles. In late cycles, the magnitude of increase in Hs due to a migration event decreased and the oscillatory pattern dampened to a continuous decrease as the populations approached the limits of responses (**Supplementary Figure 5**).

Island PS demonstrated lesser rates of inbreeding compared to island GS and WGS. RM design showed the least rates of inbreeding among the four mating designs for BI, RB, and FC migration policies (**Supplementary Figures 6**, **7**). CR design followed a pattern similar to HN or GM depending on the selection method. Among migration policies, the FC policy demonstrated lesser rates of inbreeding compared to BI and RB policies, whereas the BI policy demonstrated the largest rates of inbreeding. The GM design demonstrated rates of inbreeding that were intermediate between RM and HN/CR designs (**Supplementary Figure 7**).

Rs_Var for island selection with FC migration policies was larger than that observed with centralized populations, demonstrating larger efficiency of converting loss of genetic variance into gain. However, with FC policy, all mating designs showed a similar pattern (**Supplementary Figure 8**), whereas Rs_Var for island selection with BI and RB policies was comparable to that of centralized PS and GS, except for GM design, which showed larger Rs_Var after 10–20 cycles of selection (**Supplementary Figure 9**).

Diversity Within and Among Islands

The average within-island genotypic variance decreased towards zero through 40 cycles of selection, whereas global and inter-island genotypic variance increased before becoming limited. The rates of decrease in average within-island genotypic variance were influenced by the selection method, mating design, and migration policy. Both GS and WGS demonstrated similar patterns of loss of genotypic variance within islands, and rates of loss with both the selection methods were faster than PS (**Figure 6**). The HN mating design demonstrated the fastest loss of within-island genotypic variance followed by RM, CR, and GM designs. The FC migration policy provided the slowest loss of within-island genotypic variance followed by RB and BI migration policies (**Figure 6**). Notice, however, an oscillatory pattern in which within-island genotypic variance increased with every migration event and decreased because of selection in cycles when there were no migrants. For both the within-island genotypic variance and the expected heterozygosity, the magnitudes of oscillations dampened towards zero after 20–30 cycles of selection except for the GM mating designs coupled with BI and RB migration policies (designated IM-GM-BI and IM-GM-RB, respectively, in

Figure 6). The amplitude of increased genetic variance due to migration was greater for RB and BI migration policies with large spikes after 25–30 cycles of selection, while the amplitudes were smaller with the FC migration policies (**Figure 6**).

The largest values for inter-island genotypic variance were obtained with the RM design combined with BI and RB migration policies followed by CR and HN designs with BI and RB migration policies (**Figure 8**). Whereas, the FC migration policies demonstrated the smallest increases in inter-island genotypic variance through 40 cycles of selection (**Figure 8**). Recall that the FC migration policies provide the greatest migration rates among islands. Global genotypic variance in family island populations increased due to increase in inter-island genotypic variance. The BI migration policies demonstrated the largest global genetic variance for RM, HN, and CR mating designs followed by the RB migration policies. The GM design with BI and RB migration policies provided intermediate rates of increase in global genotypic variance while the FC migration policy showed the least increase in global genotypic variance when coupled with the HN, CR, RM, and GM mating designs (**Figure 8**).

Within the classes of migration policies, the migration frequency had significant influence on rates and limits of responses across most combinations of selection methods, mating designs, and migration policies, while numbers of migrants significantly affected responses only for a few combinations of factors. Both rates and limits of response decreased with fewer migrants for the HN mating design. For the RM design, exchange of migrants among family islands once in every three cycles provided the greatest genotypic values at the limits compared to responses with more frequent exchange. Migration size and migration direction had no significant effect on limits of selection responses (data available on request).

Trade-Offs Between Short-Term and Long-Term Gains From Recurrent Selection

There were 12 combinations of selection methods and mating designs applied to centralized populations and 48 combinations of selection methods, mating designs, and migration policies applied to family island populations. From among the 60 methods, GS using a ridge regression model followed by a hub network mating design in centralized populations and WGS followed by crosses using the CR in the centralized populations respectively (designated CE-GS-HN and CE-WGS-CR in **Table 2** and **Figure 4**) demonstrated the greatest responses in the first 20 and last 20 cycles, respectively. However, if the objective for genetic improvement is to maximize gain in the first 5, 10, 30, or 40 cycles, other combinations of the factors are needed to achieve the objective. If the breeding objective is to maximize rates of genetic improvement in five to 10 cycles of recurrent selection then there are two best options: 1. Genomic selection using RRBLUP estimated phenotypic values followed by an HN mating design in family island populations with FC migration policies, or 2. Genomic selection using RRBLUP estimated phenotypic values followed by a GM design in family island populations with FC migration policies (respectively designated as IM-GS-HN-FC and IM-GS-GM-FC in **Table 2**). If the objectives are to maximize both short-term and long-term gains then the best solution was obtained by selecting with RRBLUP estimated phenotypic values followed by an HN/CR/GM in family island populations and applying an FC migration policy (designated IM-GS-HN-FC/ IM-GS-CR-FC/ IM-GS-GM-FC in **Table 2**). Among the combinations applied on centralized populations, WGS followed by the CR mating design or RM resulted in largest long-term gains, while selection using RRBLUP estimated phenotypic values followed by an HN mating design provided the greatest short-term gains. It is important to note that the relative ranking of methods will change with the weights for short-term and long-term objectives.

DISCUSSION

Significance

The challenge of finding optimal trade-offs among competing genetic improvement objectives has usually been approached by combining selection and crossing in a single step without consideration of population structure (Akdemir and Sánchez, 2016; De Beukelaer et al., 2017; Akdemir et al., 2019; Allier et al., 2019a,b, 2020; Ramasubramanian and Beavis, 2020). Akdemir and Sánchez (2016) combined selection and mating in their GM method. De Beukelaer et al. (2017) used weighted selection indices to maximize gain while retaining a threshold level of diversity. Among the three diversity measures they tested, indices that incorporate diversity measures to minimize loss of rare favorable alleles and minimize heterozygosity resulted in responses that were greater than WGS with truncation selection. Including diversity measures in a set offered advantage over truncation selection, as selected mate pairs retained rare favorable alleles better than WGS coupled with RM design. Allier et al., 2019a,b included the impact of within-family selection to maximize genetic gain while minimizing loss of genetic variance, but they did not consider migration among families.

Ramasubramanian and Beavis (2020) investigated GS methods for the genetic improvement of soybean, but only considered the HN mating design applied among F_5-derived lines regardless of their family affiliation. Herein, we approached the challenge by disentangling breeding decisions into four distinct groups: (1) organization of the breeding population, (2) selection methods, (3) mating designs, and (4) migration policies. Each of these were divided into possible alternatives within each group and treated as independent factors in orthogonal treatment combinations.

As with our previous investigation, we found that the fastest rates of genetic improvement resulted when GS followed by the HN mating design is applied to the centralized populations (Ramasubramanian and Beavis, 2020). When combined, these three decisions have reinforcing effects on responses to selection. At the other extreme, when WGS is applied to populations organized as family islands followed by either CR or RM, the tendency of all three to retain genetic diversity reinforce each other resulting in the largest genotypic values, but only after many cycles of selection. Because the slopes of the curves resulting

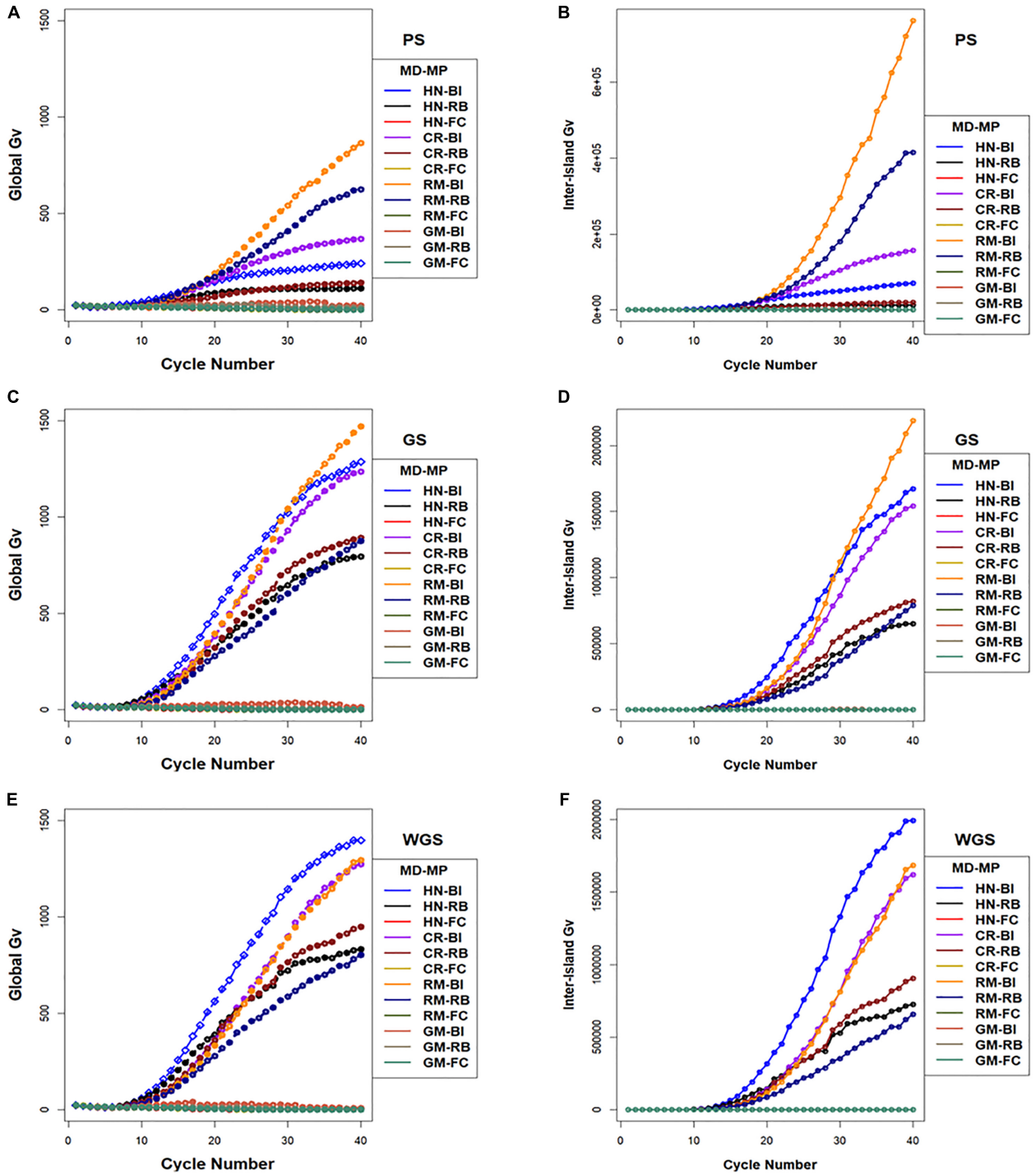

FIGURE 8 | Global and inter-island genotypic variance in island selection. (i) Global genotypic variance **(A,C,E)** and (ii) Inter-island genetic variance **(B,D,F)** for Phenotypic Selection (PS) **(A,B)**, Genomic Selection (GS) **(C,D)**, and Weighted Genomic Selection (WGS) **(E,F)** for the four mating designs including Hub Network (HN), Chain Rule (CR), Random Mating (RM), and Genomic Mating (GM) methods. Migration policy included "Best Island" (BI), "Random Best" (RB), and "Fully Connected" (FC) for 400 simulated QTL and 0.7 H. Migration rules included bidirectional exchange of two immigrants and emigrants every other cycle of selection. Genotypic variance is standardized to the average genotypic variance in founder populations in cycle "0." GP models are updated every cycle in GS and WGS using training sets with data from all prior cycles of selection.

TABLE 2 | Trade-off table for strategies.

Objectives	Objectives weighted equally for gain across 40 cycles	Method				
		IM-FC-GS-HN	**IM-FC-GS-CR**	**IM-FC-WGS-HN**	**IM-FC-GS-GM**	**IM-FC-WGS-GM**
Rs in 5 cycles (Rank)	**0.2**	0.14 **(5)**	0.13 **(13)**	0.13 **(13)**	0.13 **(13)**	0.13 **(13)**
Rs in 10 cycles **(Rank)**	**0.2**	0.29 **(3)**	0.28 **(7)**	0.29 **(3)**	0.27 **(11)**	0.26 **(12)**
Rs in 20 cycles **(Rank)**	**0.2**	0.53 **(1)**	0.51 **(4)**	0.51 **(4)**	0.48 **(8)**	0.49 **(7)**
Rs in 30 cycles **(Rank)**	**0.2**	0.59 **(7)**	0.59 **(7)**	0.59 **(7)**	0.59 **(7)**	0.59 **(7)**
Rs in 40 cycles **(Rank)**	**0.2**	0.59 **(13)**	0.6 **(8)**	0.59 **(13)**	0.61 **(4)**	0.61 **(4)**
Weighted rank		**1**	**2**	**3**	**4**	**4**
Objectives	**Objectives weighted highly for gain in the first 20 cycles**	**Method**				
		IM-FC-GS-HN	**IM-FC-WGS-HN**	**IM-FC-GS-CR**	**CE-GS-CR**	**CE-GS-GM**
Rs in 5 cycles (Rank)	**0.5**	0.14 **(5)**	0.13 **(13)**	0.13 **(13)**	0.14 **(5)**	0.14 **(5)**
Rs in 10 cycles (Rank)	**0.2**	0.29 **(3)**	0.29 **(3)**	0.28 **(7)**	0.28 **(7)**	0.28 **(7)**
Rs in 20 cycles (Rank)	**0.1**	0.53 **(1)**	0.51 **(4)**	0.51 **(4)**	0.44 **(18)**	0.44 **(18)**
Rs in 30 cycles (Rank)	**0.1**	0.59 **(7)**	0.59 **(7)**	0.59 **(7)**	0.47 **(22)**	0.47 **(22)**
Rs in 40 cycles (Rank)	**0.1**	0.59 **(13)**	0.59 **(13)**	0.60 **(8)**	0.47 **(26)**	0.47 **(26)**
Weighted rank		**1**	**2**	**3**	**4**	**4**

Trade-off table to support the decision for selecting the best strategy to achieve objectives including maximum gain in 5, 10, 20, 30, and 40 cycles of recurrent selection. The methods are ranked for each of the objectives based on standardized genetic responses. The absolute genotypic response values for each of the methods are provided along with the ranking of the method for the specific objective in bold numeric in parenthesis. Two sets of objective weights are provided to define the relative importance of the objectives: (i) the weighted rank of methods are estimated with more emphasis on the first 20 cycles (top), (ii) the weighted rank of methods are estimated with equal emphasis on the first and last 20 cycles (bottom). The best five methods among the 60 methods for each of the weighted objectives are presented. The simulations are provided for 400 simulated QTL responsible for 70% of phenotypic variability. Migration policies include "Isolated Selection," "Best Island," "Random Best," and "Fully Connected." Other migration factors are set to constant values: migration frequency - 2, migration direction - 2 (bi-directional), and migration size - 2. Selection methods include PS, Phenotypic Selection; GS, Genomic Selection; and WGS, Weighted Genomic Selection. Mating designs include HN (Hub Network), CR (Chain rule), RM, Random Mating; and GM, Genomic Mating method.

from WGS and PS at 40 cycles are still positive, it is possible that both selection methods could continue to produce greater genetic potential with more cycles of selection. In previous comparative studies, WGS produced long-term responses that are similar to methods such as Optimal Contribution Selection and Expected Maximum – Haploid Value (Daetwyler et al., 2015; Müller et al., 2018). Herein when we applied WGS to centralized lines followed by the GM design, the genotypic values at the limits to response were greater than the genotypic values obtained with PS or GS followed by GM. This combination also retained the largest values for heterozygosity and favorable alleles across more cycles. However, the differences between responses to GS and WGS followed by GM were not significant when applied to the populations organized into family islands with migration among islands.

Between the extreme response curves, it was also possible to find many response curves with intermediate trade-offs between the objectives. For example, applying WGS to lines that were not organized into islands followed by HN provided greater response rates than other combinations of factors involving WGS. Selection among lines organized into family islands resulted in responses that were larger or comparable to responses from centralized populations for only a limited number of combinations of mating design (GM) and migration policies (RB and FC). This may be due to the small numbers of related lines on each island (20 $\times$ smaller than the centralized population). With such a small number, selection can deplete all the genetic variance within the first 10–15 cycles as demonstrated in isolated selection. When there is no migration, which is the major source of new genetic variability, the populations realized only 10%–15% of maximum potential in the founder populations even while optimizing for sustainable gain using the GM method. A relaxed selection intensity, where the top 20% of the lines in each island are selected can sustain responses for longer cycles as demonstrated in centralized and island selection with migration (**Supplementary Table 3**).

As expected, even with small numbers of lines per island, migration had a positive impact on the outcomes. It is known that intermediate levels of migration rate result in optimal trade-offs between gain and diversity (Skolicki, 2007; Skolicki and Jong, 2007: Obolski et al., 2017). However, the range of intermediate parameter values depend on the specific context. In our study, responses in family islands were larger than selection responses in centralized populations only when migration events happened every cycle or once in two cycles. When migration happened once in three cycles of selection, the rates of responses in the early cycles were very low resulting in fewer cycles of response to selection and lower genotypic values as the limits to selection were approached. Migration size and direction did not have any significant impact on response within the small range of parameter values we tested for migration size and direction.

Also, we retained the best line, in terms of the selection metric, within island during migration events and replaced the second best line in the ranked list of selected lines with the immigrant for the BI and RB policies, whereas, for the FC policy, lines that are ranked from 2–6 are replaced. This replacement policy allows crossing between lines that are best within islands and immigrant lines from islands with the highest selection metric resulting in high rates of response within islands. We hypothesize that other policies that replace lines with low selection metric value with high selection metric values from immigrant islands will reduce genetic diversity within islands and result in different outcomes compared to the policy we have implemented.

Nonetheless, we found a very good trade-off among the competing objectives. If GS was applied to lines on FC islands and the selected lines were mated according to the pareto-optimal crosses identified using GM, then the combination preserved genetic variance for long-term gain with little penalty relative to the realized rates of improvement in early cycles by GS and the HN mating design. Yabe et al. (2016) have reported outcomes from recurrent GS in rice populations using the following three migration policies: centralized populations also referred to as bulked GS, discrete GS that corresponds to the fully isolated selection, and island GS which corresponds to the island model selection. In their study, where they used the CR for crossing, GS on centralized populations showed larger responses in the seven to eight cycles compared to isolated and island selection, whereas island GS demonstrated larger responses than the centralized and isolated GS after 12 cycles of selection. Similarly, in this study, GS on centralized populations with CR mating design resulted in larger long-term responses compared to most of the island GS except when an FC migration policy was used, where the responses were roughly similar in the first 10 cycles. Moreover, the responses were larger in the late cycles with Island GS and FC policy than in the centralized policy with CR mating design similar to the outcomes in the Yabe et al. (2016) study.

In another study, Technow et al. (2021) have investigated the impact of breeding program structure in maize hybrid development. They observed that a centralized policy provided the best responses, when the genetic architecture is completely additive. This roughly corresponds to the results we observed, where GS in centralized populations showed the largest short-term responses with the HN mating design, which is similar to the disproportional contributions in Technow et al. (2021). They also noted that, as the genetic complexity increased, the distributed and isolated policies provided larger responses. In summary, motivated by Akdemir and Sánchez (2016) and Yabe et al. (2016), we demonstrate that it is possible to design breeding strategies to produce near maximal rates of genetic improvement while retaining maximal genetic potential for long-term genetic improvement.

Future Research

By framing breeding strategies as orthogonal combinations of population structure, selection methods, mating designs, and migration policies, we illustrated the potential of the approach for a small arbitrary soybean genetic improvement project. We did not consider the relative emphasis of objectives and constraints for any specific genetic improvement project. Consider first the structure of breeding populations. We compared a centralized structure of lines with family islands created by individual crosses among the founders and then we selected within and among islands according to the same criteria. This might make sense within a single soybean genetic improvement project for lines adapted to MGs II and III. Alternatively, individual breeding projects might be considered breeding islands.

There are six public soybean genetic improvement projects adapted to MGs II and III. There are likewise about the same number of commercial soybean genetic improvement projects in the same MGs. All of these projects began at different times and were initiated with unique, albeit overlapping, germplasm resources (Mikel et al., 2010). While all of the projects select lines with greater genotypic values for yield, the yield values are obtained from different, overlapping, environments.

From the perspective of soybean genetic improvement across regions within MGs II and III, each genetic improvement project can be represented as an island where genotypes are exchanged among project islands based on annual evaluations in uniform regional trials and according to legal licensing rules. In practice, breeding projects exchange projects only the best performing lines adapted to similar environmental conditions. Nonetheless, soybean breeders will maintain useful genetic variability by exchanging lines among island projects. An advantage is that diversity among islands increases with selection, even when within-island diversity decreases. Eventually, beyond 40 cycles of recurrent selection, genetic variability among islands will decrease as genetic variability among islands is lost to selection.

Future investigations of breeding strategies to optimize trade-offs between rates of genetic gains and retention of useful genetic variance in soybean adapted to MGs II and III should consider population structures within island projects that more accurately reflect those that currently exist. Also, future investigations should simulate genetic architectures with genotype $\times$ environment effects. It is well known that a line adapted to one environment may not perform well in other environments, and it is possible to define fitness values so that they include environmental effects. Third, future investigations should consider a broader set of migration rules and policies. The FC migration policy is considered the upper bound of island models as all islands are connected to every other island with maximum migration rates among islands. While our results indicate that this policy provided the best long-term genotypic values, it remains to be tested whether it will provide the best results for genetic architectures with genotype by environment interaction effects.

Fourth, we need to recognize islands in time because every cycle of selection discards useful genetic variability. A soybean germplasm resource project was set up (Mikel et al., 2010) to recover useful genetic variability lost during domestication of soybean (Nelson, 2011). Rather than trying to build long bridges to islands located in the distant past, our results suggest that there should be a large amount of useful genetic variability that was discarded in the first few cycles of modern soybean breeding. For that matter, until response to selection reaches the half-life for the population, large amounts of useful

genetic variability can probably be recovered from islands represented by recent cycles of discarded lines. These conjectures should be preceded by simulations to determine the potential benefit and costs associated with sampling lines in recently discarded islands.

Fifth, it should be clear that a predefined mating design does not take advantage of opportunities created by each cycle of progeny to optimize outcomes according to most project objectives. Thus, there continues to be a need for algorithms that efficiently and effectively identify crosses from among genotypes produced by each cycle of selection. It is tempting to adopt and investigate all evolutionary algorithm strategies. However, only a subset is relevant to the practice of plant breeding (Hagan et al., 2012). For example, mutation and recombination rates can be controlled in computational evolutionary algorithms, whereas plant breeders cannot regulate these with current practices. Nonetheless there are many opportunities for cross-disciplinary research between evolutionary computing and plant breeding. There is a large body of literature concerning the properties of evolutionary algorithms and factors and strategies that affect convergence rates and quality of solutions (Goldberg, 1989; Goldberg and Deb, 1992; Whitley et al., 1999; Skolicki, 2007; Skolicki and Jong, 2007; Črepinšek et al., 2013; Obolski et al., 2017), and working with computational scientists should reveal novel methods to maximize the genetic potential of a breeding population in a minimum number of cycles.

Akdemir and Sánchez (2016) proposed only one of many possible GAs to identify pareto-optimal solution pairs. An approach introduced by Gaur and Deb (2016) and Mittal et al. (2020) would use statistical methods such as clustering and machine learning to unravel relationship among pareto-optimal solutions. The statistical knowledge can be used to improve the search for optimal solutions and establish several cycles of optimization. Conceptually, unveiling any relationship among pareto-optimal pairs in a genotypic space is likely to provide new knowledge regarding the characteristics of such complementary pairs. In addition, modeling responses with a first-order recurrence equation or a non-linear mixed effects model to predict the half-life and asymptotic limits of selection have potential to improve the efficiency of GAs by providing repair operators to alter the trajectory of population evolution towards the desired optimal trade-offs.

Lastly, consider the challenge of stating explicit relative emphasis on objectives and definition of constraints for any specific genetic improvement project. As noted previously, this challenge exists because it requires assigning economic and agronomic value of short-term genetic gains vs. the forecasted value of useful genetic variants that may be discarded each cycle of selection. As a thought experiment, note that the trade-off objectives can be reduced to a single "grand" objective of creating a genotype (line) with the genotypic value equal to the full genetic potential of the founders in a single cycle. For a genetic architecture consisting of two alleles at a single locus, achieving the single grand objective is trivial. Also, it is possible to imagine that the grand objective can be achieved for a complex genetic architecture with infinite resources. Clearly, given genetic architectures of complex traits and resource constraints, there are no feasible solutions to the grand objective, but it is a useful reference to serve as the goal.

In summary, we have evaluated and suggested several novel combinations of existing genomic selection methods, mating designs, and migration rules that resulted in improved responses. The study has demonstrated the potential of these new approaches, which integrate the strengths of whole-genome level information, prediction modeling, and optimization methods to contribute to the development of decision support systems for real plant breeding programs.

DATA AVAILABILITY STATEMENT

Simulated data and software codes are available as part of the R package "SoyNAMSelectionMethods" (**Supplementary File 3**). Documentation for downloading and using the package are available at http://gfspopgen.agron.iastate.edu/SoyNAMSelectionMethods_v2_2020.html. The SoyNAM founder genotypic and phenotypic data are available in SoyBase (Grant et al., 2010). The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

VR and WDB contributed to the conception and design of the study. VR wrote the software code for simulations and performed statistical analysis. VR and WDB are responsible for the interpretation of the analysis. VR wrote the first draft of the manuscript. VR and WDB contributed to revisions, preparation of the final draft and approved the submitted version.

FUNDING

Funding for this research was provided by the Department of Agronomy, Iowa State University; the North Central Soybean Research Program; and an NSF grant (1830478). Supplementary funding for large-scale computing was enabled by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation. XSEDE resources consisted of research allocations (DMS180041, DMS190003, DMS190015, and DMS190018) on PSC-Bridges Large Memory nodes for the simulations involving island model and genomic mating simulations.

ACKNOWLEDGMENTS

We want to thank Deniz Akdemir for discussions on implementing "genomic mating" and Lizhi Wang for efficient

programs to simulate meiosis. We also want to thank Alencar Xavier for sharing an efficient expectation maximization method for fitting ridge regression GP models. A preprint of this article is available on bioRxiv at https://www.biorxiv.org/content/10.1101/2021.02.19.431938v1.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.675500/full#supplementary-material

REFERENCES

Akdemir, D., and Sánchez, J. I. (2016). Efficient breeding by genomic mating. *Front. Genet.* 7:210. doi: 10.3389/fgene.2016.00210

Akdemir, D., Beavis, W., Fritsche-Neto, R., Singh, A. K., and Isidro-Sánchez, J. (2019). Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity* 122, 672–683. doi: 10.1038/s41437-018-0147-1

Akdemir, D., Sanchez, J. I., Haikka, H., and Brum, I. B. (2018). *GenomicMating: Efficient Breeding by Genomic Mating. R package version 2.0*.

Allier, A., Lehermeier, C., Charcosset, A., Moreau, L., and Teyssèdre, S. (2019a). Improving short- and long-term genetic gain by accounting for within-family variance in optimal cross-selection. *Front. Genet.* 10:1006. doi: 10.3389/fgene.2019.01006

Allier, A., Moreau, L., Charcosset, A., Teyssèdre, S., and Lehermeier, C. (2019b). Usefulness criterion and post-selection parental contributions in multi-parental crosses: application to polygenic trait introgression. *G3* 9, 1469–1479. doi: 10.1534/g3.119.400129

Allier, A., Teyssèdre, S., Lehermeier, C., Moreau, L., and Charcosset, A. (2020). Optimized breeding strategies to harness genetic resources with different performance levels. *BMC Genomics* 21:349.

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., and Jannink, J.-L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome* 4, 132–144. doi: 10.3835/plantgenome2011.02.0007

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., and Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci.* 242, 23–36. doi: 10.1016/j.plantsci.2015.08.021

Baty, F., Ritz, C., Charles, S., Brutsche, M., Flandrois, J.-P., and Delignette-Müller, M.-L. (2015). A toolbox for nonlinear regression in R: The package nlstools. *J. Stat. Softw.* 5, 1–21. doi: 10.18637/jss.v066.i05

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci.* 48, 1649–1664. doi: 10.2135/cropsci2008.03.0131

Bernardo, R., and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47, 1082–1090. doi: 10.2135/cropsci2006.11.0690

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., et al. (2015). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci.* 55, 154–163. doi: 10.2135/cropsci2014.07.0460

Brisbane, J., and Gibson, J. (1995). Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions. *Theor. Appl. Genet.* 91, 421–431. doi: 10.1007/BF00222969

Bulmer, M. G. (1971). The effect of selection on genetic variability. *Am. Nat.* 105, 201–211.

Byrum, J., Beavis, B., Davis, C., Doonan, G., Doubler, T., Kaster, V., et al. (2017). Genetic gain performance metric accelerates agricultural productivity. *Interfaces* 47, 442–453. doi: 10.1287/inte.2017.0909

Cantú-Paz, E. (2000). *Efficient and accurate parallel genetic algorithms. Boston, Mass.* Boston: Kluwer Academic Publishers.

Carvalheiro, R., de Queiroz, S. A., and Kinghorn, B. (2010). Optimum contribution selection using differential evolution. *R Bras Zootec* 39, 1429–1436. doi: 10.1590/S1516-35982010000700005

Clark, S. A., Kinghorn, B. P., Hickey, J. M., and van der Werf, J. H. J. (2013). The effect of genomic information on optimal contribution selection in livestock breeding programs. *Genet. Sel. Evol.* 45, 44–44. doi: 10.1186/1297-9686-45-44

Cockerham, C. C., and Burrows, P. M. (1980). Selection limits and strategies. *Proc. Natl. Acad. Sci. U.S.A.* 77, 546–549. doi: 10.1073/pnas.77.1.546

Combs, E., and Bernado, R. (2013). Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *Plant Genome* 6, 1–7. doi: 10.3835/plantgenome2012.11.0030

Cooper, M., Messina, C. D., Podlich, D., Totir, L. R., Baumgarten, A., Hausmann, N. J., et al. (2014). Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop Pasture Sci.* 65, 311–336. doi: 10.1071/CP14007

Cooper, M., Podlich, D., Micallef, K., Smith, O., Jensen, N., Chapman, S., et al. (2002). "Quantitative genetics, genomics and plant breeding," in *Complexity, quantitative traits and plant breeding: a role for simulation modelling in the genetic improvement of crops*, ed. M. S. Kang (Wallingford: CABI Publishing), 143–166. doi: 10.1079/9780851996011.0143

Cowling, W. A., Li, L., Siddique, K. H. M., Henryon, M., Berg, P., Banks, R. G., et al. (2017). Evolving gene banks: improving diverse populations of crop and exotic germplasm with optimal contribution selection. *J. Exp. Bot.* 68, 1927–1939.

Črepinšek, M., Liu, S.-H., and Mernik, M. (2013). Exploration and exploitation in evolutionary algorithms: a survey. *ACM Comput. Surv.* 45:35. doi: 10.1145/2480741.2480752

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., et al. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* 112, 48–60. doi: 10.1038/hdy.2013.16

Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., and Hayes, B. J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics* 200, 1341–1348. doi: 10.1534/genetics.115.178038

De Beukelaer, H. D., Badke, Y., Fack, V., and Meyer, G. D. (2017). Moving beyond managing realized genomic relationship in long-term genomic selection. (Author abstract). *Genetics* 206, 1127–1138. doi: 10.1534/genetics.116.194449

Deb, K. (2003). Unveiling innovative design principles by means of multiple conflicting objectives. *Eng. Optim.* 35, 445–470. doi: 10.1080/0305215031000151256

Deb, K. (2011). "Multi-objective optimisation using evolutionary algorithms: an introduction," in *Multi-Objective Evolutionary Optimisation for Product Design and Manufacturing*, eds L. Wang, A. Ng and K. Deb (London: Springer).

Dempfle, L. (1974). A note on increasing the limit of selection through selection within families. *Genet. Res.* 24, 127–135. doi: 10.1017/S0016672300015160

Diers, B. W., Specht, J., Rainey, K. M., Cregan, P., Song, Q., Ramasubramanian, V., et al. (2018). Genetic architecture of soybean yield and agronomic traits. *G3* 8, 3367–3375. doi: 10.1534/g3.118.200332

Frank, M., and Wolfe, P. (1956). An algorithm for quadratic programming. *Nav. Res. Logist. Q.* 3, 95–110. doi: 10.1002/nav.3800030109

Gaur, A., and Deb, K. (2016). Adaptive use of innovization principles for a faster convergence of evolutionary multi-objective optimization algorithms. *GECCO* 75–76.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Goiffon, M., Kusmec, A., Wang, L., Hu, G., and Schnable, P. S. (2017). Improving response in genomic selection with a population-based selection strategy: optimal population value selection. *Genetics* 206, 1675–1682. doi: 10.1534/genetics.116.197103

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Boston, MA: Addison-Wesley Pub. Co.

Goldberg, D. E., and Deb, K. (1992). Massive multimodality, deception, and genetic algorithms. in Parallel Problem Solving from Nature eds R. Manner and B. Manderick (Berlin: Springer-Verlag).

Goldberg, S. (1958). *Introduction to difference equations, with illustrative examples from economics, psychology, and sociology*. New York: Wiley.

Gorjanc, G., Gaynor, R. C., and Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent

genomic selection. *Theor. Appl. Genet.* 131, 1953–1966. doi: 10.1007/s00122-018-3125-3

Goudet, J. (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* 5, 184–186. doi: 10.1111/j.1471-8286.2004.00828.x

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, D843–D846. doi: 10.1093/nar/gkp798

Grundy, B., Villanueva, B., and Woolliams, J. A. (1998). Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. *Genet. Res.* 72, 159–168. doi: 10.1017/S0016672398003474

Guo, B., Sleper, D. A., and Beavis, W. D. (2010). Nested association mapping for identification of functional markers. *Genetics* 186, 373–383. doi: 10.1534/genetics.110.115782

Guo, B., Wang, D., Guo, Z., and Beavis, W. D. (2013). Family-based association mapping in crop species. *Theor. Appl. Genet.* 126, 1419–1430. doi: 10.1007/s00122-013-2100-2

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., et al. (2014). The impact of population structure on genomic prediction in stratified populations. *Theor. Appl. Genet.* 127, 749–762. doi: 10.1007/s00122-013-2255-x

Hagan, S., Knowles, J., and Kell, D. B. (2012). Exploiting genomic knowledge in optimising molecular breeding programmes: algorithms from evolutionary computing. *PLoS One* 7:e48862. doi: 10.1371/journal.pone.0048862

Haimes, Y. Y., Lasdon, L. and Wismer, D. (1971). On a bicriterion formation of the problems of integrated system identification and system optimization. *IEEE Trans. Syst. Man Cybern.* 1, 296–297.

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52, 146–160. doi: 10.2135/cropsci2011.06.0297

Hickey, J. M., Chiurugwi, T., Mackay, I., and Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* 49, 1297–1303. doi: 10.1038/ng.3920

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. *Crop Sci.* 54, 1476–1488. doi: 10.2135/cropsci2013.03.0195

Hill, W. G., and Robertson, A. (2008). The effect of linkage on limits to artificial selection. *Genet. Res.* 89, 311–336. (First published in 1968)

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42:35. doi: 10.1186/1297-9686-42-35

Johnson, B., Gardner, C. O., and Wrede, K. C. (1988). Application of an optimization model to multi-trait selection programs. *Crop Sci.* 28, 723–728. doi: 10.2135/cropsci1988.0011183X002800050001x

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521

Jonas, E., and de Koning, D. J. (2016). Goals and hurdles for a successful implementation of genomic selection in breeding programme for selected annual and perennial crops. *Biotechnol. Genet. Eng. Rev.* 32, 18–42. doi: 10.1080/02648725.2016.1177377

Kang, H. (1983). Limits of artificial selection under balanced mating systems with family selection. *Silvae Genet.* 32, 188–195.

Kang, H., and Namkoong, G. (1980). Limits of artificial selection under unbalanced mating systems. *Theor. Appl. Genet.* 58, 181–191. doi: 10.1007/BF00263115

Kang, H., and Nienstaedt, H. (1987). Managing long-term tree breeding stock. *Silvae Genet.* 1987, 30–39. doi: 10.1016/j.scitotenv.2020.143695

Karush, W. (1939). *Minima of functions of several variables with inequalities as side constraints*. Chicago: Univ of Chicago. M Sc Dissertation.

Kinghorn, B. P. (2011). An algorithm for efficient constrained mate selection. *Genet. Sel. Evol.* 43:4. doi: 10.1186/1297-9686-43-4

Konak, A., Coit, D. W., and Smith, A. E. (2006). Multi-objective optimization using genetic algorithms: a tutorial. *Reliab. Eng. Syst.* 91, 992–1007. doi: 10.1016/j.ress.2005.11.018

Kuhn, H., and Tucker, A. (1951). *Nonlinear programming In Proceedings of 2nd Berkeley symposium*. Berkeley: University of California Press, 481–492.

Lazimy, R. (1982). Mixed-integer quadratic programming. *Math. Program.* 22, 332–349. doi: 10.1007/BF01581047

Lin, Z., Shi, F., Hayes, B. J., and Daetwyler, H. D. (2017). Mitigation of inbreeding while preserving genetic gain in genomic breeding programs for outbred plants. *Theor. Appl. Genet.* 130, 969–980. doi: 10.1007/s00122-017-2863-y

Liu, H., Meuwissen, T. H., Sorensen, A. C., and Berg, P. (2015). Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs. *Genet. Sel. Evol.* 47:19. doi: 10.1186/s12711-015-0101-0

Luque, G. (2011). *Parallel Genetic Algorithms: Theory and Real World Applications*. Heidelberg: Springer.

Marulanda, J., Mi, X., Melchinger, A., Xu, J.-L., Würschum, T., and Longin, C. (2016). Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. *Theor. Appl. Genet.* 129, 1901–1913. doi: 10.1007/s00122-016-2748-5

McCarl, B. A., Moskowitz, H., and Furtan, H. (1977). Quadratic programming applications. *Omega* 5, 43–55. doi: 10.1016/0305-0483(77)90020-2

Melchinger, A. E., Schmidt, W., and Geiger, H. H. (1988). Comparison of testcrosses produced from F2 and first backcross populations in maize. *Crop Sci.* 28, 743–749. doi: 10.2135/cropsci1988.0011183X002800050004x

Meuwissen, T., Hayes, B. and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Meuwissen, T. H. (1997). Maximizing the response of selection with a predefined rate of inbreeding. *J. Anim. Sci.* 75, 934–940. doi: 10.2527/1997.754934x

Mikel, M. A., Diers, B. W., Nelson, R. L., and Smith, H. H. (2010). Genetic diversity and agronomic improvement of North American soybean germplasm. *Crop Sci.* 50, 1219–1229. doi: 10.2135/cropsci2009.08.0456

Mittal, S., Saxena, D. K., Deb, K., and Goodman, E. (2020). Enhanced innovized repair operator for evolutionary multi-and many-objective optimization. *arXiv* [Preprint]. arXiv:2011.10760

Müller, D., Schopp, P., and Melchinger, A. E. (2018). Selection on expected maximum haploid breeding values can increase genetic gain in recurrent genomic selection. *G3* 8, 1173–1181. doi: 10.1534/g3.118.200091

Nakaya, A., and Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? *Ann. Bot.* 110, 1303–1316. doi: 10.1093/aob/mcs109

Nelson, R. L. (2011). Managing self-pollinated germplasm collections to maximize utilization. *Plant Genet. Resour.* 9, 123–133. doi: 10.1017/S147926211000047X

Obolski, U., Lewin-Epstein, O., Even-Tov, E., Ram, Y., and Hadany, L. (2017). With a little help from my friends: cooperation can accelerate the rate of adaptive valley crossing. *BMC Evol. Biol.* 17:143. doi: 10.1186/s12862-017-0983-2

Oddi, F. J., Miguez, F. E., Ghermandi, L., Bianchi, L. O., and Garibaldi, L. A. (2019). A nonlinear mixed-effects modeling approach for ecological data: using temporal dynamics of vegetation moisture as an example. *Ecol. Evol.* 9, 10225–10240. doi: 10.1002/ece3.5543

Pinheiro, J. C., and Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. New York,NY. Springer. doi: 10.1007/978-1-4419-0318-1

Pinheiro, J. C., Bates, D. J., DebRoy, S., Sarkar, D., and R Core Team (2021). *nlme: Linear and Nonlinear Mixed Effects Models. R Package Version 3.1-152.* Available online at: https://CRAN.R-project.org/package=nlme

Podlich, D. W., and Cooper, M. (1999). Modelling plant breeding programs as search strategies on a complex response surface. *Lect. Notes Comput. Sci.* 1585, 171–178. doi: 10.1007/3-540-48873-1_23

Pryce, J. E., Hayes, B. J., and Goddard, M. E. (2012). Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *J. Dairy Sci.* 95, 377–388. doi: 10.3168/jds.2011-4254

Ramasubramanian, V., and Beavis, W. D. (2020). Factors affecting response to recurrent genomic selection in soybeans. *bioRxiv* [Preprint]. doi: 10.1101/2020.02.14.949008

Rardin, R. L. (2017). *Optimization in Operations Research*, 2nd Edn. Boston: Pearson.

Robertson, A. (1960). A theory of limits in artificial selection. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 153, 234–249.

Ryman, N., and Leimar, O. (2009). GST is still a useful measure of genetic differentiation — a comment on Jost's D. *Mol. Ecol.* 18, 2084–2087. doi: 10.1111/j.1365-294X.2009.04187.x

Takuno, S., Terauchi, R., and Innan, H. (2012). The power of QTL mapping with RILs. *PLoS One* 7:e46545. doi: 10.1371/journal.pone.0046545

Technow, F., Podlich, D., and Cooper, M. (2021). Back to the future: implications of genetic complexity for the structure of hybrid breeding programs. *G3* 11:jkab153. doi: 10.1093/g3journal/jkab153

Saeki, Y., Tudari, M., and Crowley, P. H. (2014). Allocation trade-offs and life histories: a conceptual and graphical framework. *OIKOS* 123, 786–793. doi: 10.1111/oik.00956

Schierenbeck, S., Pimentel, E. C. G., Tietze, M., Körte, J., Reents, R., Reinhardt, F., et al. (2011). Controlling inbreeding and maximizing genetic gain using semi-definite programming with pedigree-based and genomic relationships. *J. Dairy Sci.* 94, 6143–6152. doi: 10.3168/jds.2011-4574

Schnell, F. W. (1983). *Problème der Elternwahl-Ein Überblick. Arbeitstagung der Arbeitsgemeinschaft der Saatzuchleiter*. Gumpenstein: Verlag und Druck der Bundesanstalt für alpenländische Landwirtschaft, 1–1.

Seada, H., and Deb, K. (2018). "Non-dominated sorting based multi/many-objective optimization: Two decades of research and application," in *Multi-Objective Optimization*, eds J. Mandal, S. Mukhopadhyay, and P. Dutta (Singapore: Springer), 1–24.

Sheftel, H., Shoval, O., Mayo, A., and Alon, U. (2013). The geometry of the Pareto front in biological phenotype space. *Ecol. Evol.* 3, 1471–1483. doi: 10.1002/ece3.528

Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., et al. (2012). Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* 336, 1157–1160. doi: 10.1126/science.1217405

Skolicki, Z. (2007). *An analysis of island models in evolutionary computation*. Fairfax, VA: George Mason University.

Skolicki, Z., and Jong, K. D. (2007). The importance of a two-level perspective for island model design. *IEEE Congr. Evol. Compu.* 2007, 4623–4630. doi: 10.1109/CEC.2007.4425078

Sonesson, A., Woolliams, J., and Meuwissen, T. (2010). "Maximising genetic gain whilst controlling rates of genomic inbreeding using genomic optimum contribution selection," in *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Germany: German Society for Animal Science.

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2015). Fingerprinting soybean germplasm and its utility in genomic research. *G3* 28, 1999–2006. doi: 10.1534/g3.115.019000

Song, Q., Yan, L., Quigley, C., Jordan, B. D., Fickus, E., Schroeder, S., et al. (2017). Genetic characterization of the soybean nested association mapping population. *Plant Genome* 10. doi: 10.3835/plantgenome2016.10.0109

Specht, J. E., Diers, B. W., Nelson, R. L., de Toledo, J. F. F., Torrion, J. A., and Grassini, P. (2014). "Soybean," in *Yield gains in major US field crops* eds S. Smith, B. Diers, J. Specht and B. Carver (Madison, WI: American Society of Agronomy), 311–355. doi: 10.2135/cssaspecpub33.c12

Sun, C., and VanRaden, P. M. (2014). Increasing long-term response by selecting for favorable minor alleles. *PLoS One* 9:e88510. doi: 10.1371/journal.pone.0088510

USDA-ERS. (2020). *Commodity Costs and Returns*. Available online at: https://www.ers.usda.gov/data-products/commodity-costs-and-returns/commodity-costs-and-returns/#Recent%20Cost%20and%20Returns (accessed November 16, 2020).

Whitley, D., Rana, S., and Heckendorn, R. B. (1999). The island model genetic algorithm: on separability, population size and convergence. *CIT J. Comput. Inf. Technol.* 7, 33–47.

Woolliams, J. A., Berg, P., Dagnachew, B. S., and Meuwissen, T. H. (2015). Genetic contributions and their optimization. *J. Anim. Breed Genet.* 132, 89–99. doi: 10.1111/jbg.12148

Wray, N. R. and Goddard, M. E. (1994). Increasing long-term response to selection. *Genet. Sel. Evol.* 26, 431–451. doi: 10.1186/1297-9686-26-5-431

Wright, S. (1967). "Surfaces" of selective value. *Proc. Natl. Acad. Sci. U.S.A.* 58, 165–172. doi: 10.1073/pnas.58.1.165

Wright, S. (1988). Surfaces of selective value revisited. *Am. Natur.* 131, 115–123. doi: 10.1086/284777

Xavier, A. (2019). Efficient estimation of marker effects in plant breeding. *G3* 9, 3855–3866. doi: 10.1534/g3.119.400728

Xavier, A., Jarquin, D., Howard, R., Ramasubramanian, V., Specht, J. E., Graef, G. L., et al. (2017). Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3* 8, 519–529. doi: 10.1534/g3.117.300300

Xavier, A., Muir, W. M., and Rainey, K. M. (2016). Assessing predictive properties of genome-wide selection in soybeans. *G3* 6, 2611–2616. doi: 10.1534/g3.116.032268

Xavier, A., Thapa, R., Muir, W., and Rainey, K. (2018). Population and quantitative genomic properties of the USDA soybean germplasm collection. *Plant Genet. Resour.* 16, 513–523. doi: 10.1017/S1479262118000102

Yabe, S., Yamasaki, M., Ebana, K., Hayashi, H, and Iwata, H. (2016). Island-model genomic selection for long-term genetic improvement of autogamous crops. *PLoS One* 11:e0153945. doi: 10.1371/journal.pone.0153945

Zadeh, L. (1963). Optimality and non-scalar-valued performance criteria. *TAC* 8, 59–60. doi: 10.1109/TAC.1963.1105511

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*, 1st Edn. New York, NY: Springer. doi: 10.1007/978-0-387-87458-6_1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 December 2021
doi: 10.3389/fgene.2021.707754

Harnessing Genetic Diversity in the USDA Pea Germplasm Collection Through Genomic Prediction

*Md. Abdullah Al Bari[1], Ping Zheng[2], Indalecio Viera[1], Hannah Worral[3], Stephen Szwiec[3], Yu Ma[2], Dorrie Main[2], Clarice J. Coyne[4], Rebecca J. McGee[5] and Nonoy Bandillo[1]**

[1]Department of Plant Sciences, North Dakota State University, Fargo, ND, United States, [2]Department of Horticulture, Washington State University, Pullman, WA, United States, [3]NDSU North Central Research Extension Center, Minot, ND, United States, [4]USDA-ARS Plant Germplasm Introduction and Testing, Washington State University, Pullman, WA, United States, [5]USDA-ARS Grain Legume Genetics and Physiology Research, Pullman, WA, United States

Phenotypic evaluation and efficient utilization of germplasm collections can be time-intensive, laborious, and expensive. However, with the plummeting costs of next-generation sequencing and the addition of genomic selection to the plant breeder's toolbox, we now can more efficiently tap the genetic diversity within large germplasm collections. In this study, we applied and evaluated genomic prediction's potential to a set of 482 pea (*Pisum sativum* L.) accessions—genotyped with 30,600 single nucleotide polymorphic (SNP) markers and phenotyped for seed yield and yield-related components—for enhancing selection of accessions from the USDA Pea Germplasm Collection. Genomic prediction models and several factors affecting predictive ability were evaluated in a series of cross-validation schemes across complex traits. Different genomic prediction models gave similar results, with predictive ability across traits ranging from 0.23 to 0.60, with no model working best across all traits. Increasing the training population size improved the predictive ability of most traits, including seed yield. Predictive abilities increased and reached a plateau with increasing number of markers presumably due to extensive linkage disequilibrium in the pea genome. Accounting for population structure effects did not significantly boost predictive ability, but we observed a slight improvement in seed yield. By applying the best genomic prediction model (e.g., RR-BLUP), we then examined the distribution of genotyped but nonphenotyped accessions and the reliability of genomic estimated breeding values (GEBV). The distribution of GEBV suggested that none of the nonphenotyped accessions were expected to perform outside the range of the phenotyped accessions. Desirable breeding values with higher reliability can be used to identify and screen favorable germplasm accessions. Expanding the training set and incorporating additional orthogonal information (e.g., transcriptomics, metabolomics, physiological traits, etc.) into the genomic prediction framework can enhance prediction accuracy.

Keywords: genomic selection, genomic prediction, reliability criteria, germplasm accessions, pea (*Pisum sativum* L), next-generation sequencing

OPEN ACCESS

Edited by:
Diego Jarquin, University of Nebraska-Lincoln, United States

Reviewed by:
Rounak Dey, Harvard University, United States
Reka Howard, University of Nebraska System, United States

****Correspondence:***
Nonoy Bandillo
nonoy.bandillo@ndsu.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *10 May 2021*
Accepted: *26 October 2021*
Published: *24 December 2021*

Citation:
Bari MAA, Zheng P, Viera I, Worral H, Szwiec S, Ma Y, Main D, Coyne CJ, McGee RJ and Bandillo N (2021) Harnessing Genetic Diversity in the USDA Pea Germplasm Collection Through Genomic Prediction. Front. Genet. 12:707754. doi: 10.3389/fgene.2021.707754

INTRODUCTION

Pea (*Pisum sativum* L.) is a vitally important pulse crop that provides protein (15.8–32.1%), vitamins, minerals, and fibers. Pea consumption has cardiovascular benefits as it is rich in potassium, folate, and digestible fibers, which promote gut health and prevent certain cancers (Mudryj et al., 2014; Tayeh et al., 2015). Considering the health benefits of pulse crop, the US Department of Agriculture recommends regular pulses consumption, including peas, to promote human health and wellbeing (http://www.choosemyplate.gov/). In 2019, more than 446,000 hectares of edible dry pea were planted with production totaling 1,013,600 tonnes in the USA, making it the fourth largest legume crop (http://www.fao.org) (USDA, 2020). Growing peas also help maintain soil health and productivity by fixing atmospheric nitrogen (Burstin et al., 2015). Recently, the pea protein has emerged as a frontrunner and showed the most promise in the growing alternative protein market. The Beyond Meat burger is a perfect example of a pea protein product gaining traction in the growing market. About 20-g protein (17.5%) in each burger comes from pea (https://www.nasdaq.com/articles/heres-why-nows-thetime-to-buy-beyond-meat-stock-2019-12-05). Another product made from pea, Ripptein, is a non-dairy milk product of pea protein that is gaining tremendous interest as an alternative dairy product (https://www.ripplefoods.com/ripptein/). Additionally, peas are gaining attention in the pet food market as it is grain-free and an excellent source of essential amino acids required by cats and dogs (PetfoodIndustry.com; Facciolongo et al., 2014). As the demand for pea increases, particularly in the growing alternative protein market, genetic diversity expansion is needed to hasten the current rate of genetic gain in pea (Vandemark et al., 2014).

Germplasm collections serve as an essential source of variation for germplasm enhancement that can sustain long-term genetic gain in breeding programs. The USDA *Pisum* collection, held at the Western Regional Plant Introduction Station at Washington State University, is a good starting point to investigate functional genetic variation useful for applied breeding efforts. To date, this collection consists of 6,192 accessions plus a Pea Genetic Stocks collection of 712 accessions. From this collection, the USDA core collection, comprised of 504 accessions, was assembled to represent ~18% of all USDA pea accessions at the time of construction (Simon and Hannan 1995; Coyne et al., 2005). Subsequently, single-seed descent derived homozygous accessions were developed from a subset of the core to form the 'Pea Single Plant'-derived (PSP) collection. The PSP was used to facilitate the collection's genetic analysis (Cheng et al., 2015). The USDA Pea Single Plant Plus Collection (Pea PSP) was assembled as well as included the PSP and Chinese accessions and field, snap and snow peas from US public pea breeding programs (Holdsworth et al., 2017).

Genomic selection (GS) takes advantage of high-density genomic data that holds a promise to increase the rate of genetic gain (Meuwissen et al., 2001). As genotyping costs have significantly declined relative to current phenotyping costs, GS has become an attractive option as a selection decision tool to evaluate accessions in extensive germplasm collections. A genomic prediction approach could use only genomic data to predict each accession's breeding value in the collection (Meuwissen et al., 2001; Habier et al., 2007; VanRaden, 2008). The predicted values would significantly increase the value of accessions in germplasm collections by giving breeders a means to identify those favorable accessions meriting their attention from the thousand available accessions in germplasm collections (Longin et al., 2014; Crossa et al., 2016; Jarquin et al., 2016). Several studies used the genomic prediction approach to harness diversity in germplasm collections, including lentil (Haile et al., 2020), soybean (Jarquin et al., 2016), wheat (Crossa et al., 2016), rice (Spindel et al., 2015), sorghum (Yu et al., 2016), maize (Gorjanc et al., 2016), and potato (Bethke et al., 2019). A pea genomic selection study for drought-prone Italian environment revealed increased selection accuracy of pea lines (Annicchiarico et al., 2019; Annicchiarico et al., 2020). To the best of our knowledge, no such studies have been performed using the USDA Pea Germplasm Collection, but a relevant study has been conducted using a diverse pea germplasm set comprised of more than 370 accessions genotyped with a limited number of markers (Burstin et al., 2015; Tayeh et al., 2015).

To date, methods to sample and utilize an extensive genetic resource like germplasm collections remain a challenge. In this study, a genomic prediction approach targeting complex traits, including seed yield and phenology, was evaluated to exploit diversity contained in the USDA Pea Germplasm Collection. No research has been conducted before on genomic prediction for the genetic exploration of the USDA Pea Germplasm Collection. Different cross-validation schemes were used to answer essential questions surrounding the efficient implementation of genomic prediction and selection, including determining best prediction models, optimum population size and number of markers, and impact of accounting population structure into genomic prediction framework. We then examined the distribution of all nonphenotyped accessions using SNP information in the collection by applying genomic prediction models and estimated reliability criteria of genomic estimated breeding values for the assessed traits.

MATERIALS AND METHODS

Plant Materials

A total of 482 USDA germplasm accessions were used in this study, including the Pea Single Plant Plus Collection (Pea PSP) comprised of 292 accessions (Cheng et al., 2015). The USDA Pea Core Collection contains accessions from different parts of the world and represents the entire collection's morphological, geographic, and taxonomic diversity. These accessions were initially acquired from 64 different countries and are conserved at the Western Regional Plant Introduction Station, USDA, Agricultural Research Service (ARS), Pullman, WA (Cheng et al., 2015).

DNA Extraction, Sequencing, SNP Calling

Green leaves were collected from seedlings of each accession grown in the greenhouse with the DNeasy 96 Plant Kit (Qiagen,

Valencia, CA, USA). Genomic libraries for the Single Plant Plus Collection were prepped at the University of Minnesota Genomics Center (UMGC) using genotyping-by-sequencing (GBS). Four hundred eighty-two (482) dual-indexed GBS libraries were created using restriction enzyme *Ape*KI (Elshire et al., 2011). A NovaSeq S1 1 × 100 Illumina Sequencing System (Illumina Inc., San Diego, CA, USA) was then used to sequence the GBS libraries. Preprocessing was performed by the UMGC that generated the GBS sequence reads. An initial quality check was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequencing adapter remnants were clipped from all raw reads. Reads with final length <50 bases were discarded. The high-quality reads were aligned to the reference genome of *Pisum sativum* (Pulse Crop Database https://www.pulsedb.org/, Kreplak et al., 2019) using the Burrow Wheelers Alignment tool (Version 0.7.17) (Li and Durbin, 2009) with default alignment parameters, and the alignment data was processed with SAMtools (version 1.10) (Li et al., 2009). Sequence variants, including single and multiple nucleotide polymorphisms (SNPs and MNPs, respectively), were identified using FreeBayes (Version 1.3.2) (Garrison and Marth, 2012). The combined read depth of 10 was used across samples for identifying an alternative allele as a variant, with the minimum base quality filters of 20. The putative SNPs from freeBayes were filtered across the entire population to maintain the SNPs for biallelic with minor allele frequency (MAF) < 5%. The putative SNP discovery resulted in biallelic sites of 380,527 SNP markers. The QUAL estimate was used for estimating the Phred-scaled probability. Sites with a QUAL value less than 20 and more than 80% missing values were removed from the marker matrix. The rest of the markers were further filtered out so that heterozygosity was less than 20%. The filters were applied using VCFtools (version 0.1.16) (Danecek et al., 2011) and in-house Perl scripts. The SNP data were uploaded in a public repository and is available at this link: https://www.ncbi.nlm.nih.gov/sra/PRJNA730349 (Submission ID: SUB9608236). Missing data were imputed using a *k*-nearest neighbor genotype imputation method (Money et al., 2015) implemented in TASSEL (Bradbury et al., 2007). SNP data were converted to a numeric format where 1 denotes homozygous for a major allele, -1 denotes homozygous for an alternate allele, and 0 refers to heterozygous loci. Finally, 30,646 clean, curated SNP markers were identified and used for downstream analyses.

Phenotyping

Pea germplasm collections (Pea PSP) were planted following augmented design with standard checks ("Hampton," "Arargorn," "Columbian," and "1,022") at the USDA Central Ferry Farm in 2016, 2017, and 2018 (planting dates were March 14, March 28, and April 03, respectively). The central Ferry farm is located at Central Ferry, WA at 46°39′5.1″N; 117°45′45.4″ W, and elevation of 198 m. The Central Ferry farm has a Chard silt loam soil (coarse-loamy, mixed, superactive, mesic Calcic Haploxerolls) and was irrigated with subsurface drip irrigation at 10 min d^{-1}. All seeds were treated with fungicides; mefenoxam (13.3 ml a.i. 45 kg-1), fludioxonil (2.4 ml a.i. 45 kg -1), and thiabendazole (82.9 ml a.i.45 kg -1), insecticide; thiamethoxam (14.3 ml a.i. 45 kg -1), and sodium molybdate (16 g 45 kg -1) prior to planting. Thirty seeds were planted per plot; each plot was 152 cm long, having double rows with 30 cm center spacing. The dimensions of each plot were 152 × 60 cm. Standard fertilization and cultural practices were used.

The following traits were recorded and are presented in this manuscript. Days to first flowering are the number of days from planting to when 10% of the plot's plants start flowering. The number of seeds per pod is the number of seeds in each pod. Plant height (cm) is defined as when all plants in a plot obtained full maturity and were measured in centimeters from the collar region at soil level to the plants' top. Pods per plant is the number of recorded pods per plant. Days to maturity referred to physiological maturity when plots were hand-harvested, mechanically threshed, cleaned with a blower, and weighed. Plot weight (gm) is the weight of each plot in grams after each harvest. Seed yield (kg ha^{-1}) is the plot weight converted to seed yield in kg per hectare.

Phenotypic Data Analysis

A mixed linear model was used to extract best linear unbiased predictors (BLUPs) for all traits evaluated using the following model:

$$y_{ij} = \mu + G_i + E_j + (G*E)_{ij} + e_{ij} \quad (1)$$

where y_{ij} is the observed phenotype of *i*th genotypes and *j*th environment which is the number of years, μ is the overall mean, G_i is the random genetic effect (*i* is number of genotypes), E_j is the random environments (*j* is number of years), $(G*E)_{ij}$ is the genotype by environment interaction, and e_{ij} is the residual error.

For the purpose of estimating heritability, we fit the same model above. The heritability in broad sense (H^2) on an entry-mean basis for each assessed trait was calculated to evaluate the quality of trait measurements following the equation (Hallauer et al., 2010):

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2/j + \sigma_e^2/jr} \quad (2)$$

where σ_G^2 is the genetic variance, σ_{GE}^2 is variance due to the genotype by year interaction, σ_e^2 is the error variance, *j* is number of years considered as environments, and *r* is the relative number of occurrences of each genotype in a trial (harmonic mean of the entries). We also calculated heritability proposed by Cullis et al. (2006) implemented in Sommer package in R (Covarrubias-Pazaran, 2016).

$$H^2_{\text{Cullis}} = 1 - \left(\frac{\text{PEV}}{\text{md*Vg}}\right) \quad (3)$$

where PEV is the predicted error variance for the genotype, V_g refers to the genotypic variance, md is the mean values from the diagonal of the relationship matrix, which is an identity matrix.

The R package, lme4 (Bates et al., 2015), was used to analyze the data. The trait values derived from the BLUPs were used to measure correlation with the ggcorrplot using ggplot2 package (Wickham 2016). All phenotypic and genomic prediction models were analyzed in the R environment (R Core Team, 2020).

Genomic Prediction Models

The genomic prediction models were fitted as follows:

$$y = \mu + Zu + \varepsilon \quad (4)$$

where y is a vector of the genotype BLUPs obtained from **Eq. 1**, μ is the intercept of the model used for the study, Z is the SNP marker matrix, u is the vector of marker effects, and ε is a residual vector.

Five genomic prediction models were evaluated including ridge regression ridge regression best linear unbiased prediction approach (RR-BLUP), partial least squares regression model (PLSR), random forest (RF), BayesCpi, and Reproducing Kernel Hilbert Space (RKHS).

The RR-BLUP model assumes all markers have an equal contribution to the genetic variance. One of the most widely used methods for predicting breeding values is RR-BLUP, comparable to the best linear unbiased predictor (BLUP) used to predict the worth of entries in the context of mixed models (Meuwissen et al., 2001). The RR-BLUP basic frame model is:

$$y = Zu + \varepsilon \quad (5)$$

where $u \sim N(0, I\sigma_u^2)$ is a vector of marker effects and Z is the genotype matrix e.g., (aa, Aa, AA) = (0, 1, 2) for biallelic single nucleotide polymorphisms (SNPs) that relates to phenotype y (Endelman, 2011). The RR-BLUP genomic prediction was implemented using the "RR-BLUP" package (Endelman, 2011).

Partial least square regression is a reduction dimension technique that aims to find independent latent components that maximize the covariance between the observed phenotypes and the markers (predictor variables) (Colombani et al., 2012). The number of components (also known as latent variables) should be less than the number of observations to avoid multicollinearity issues and commonly the number of components are chosen by cross validation. PLSR was executed using the "pls" package (Mevik and Wehrens, 2007).

Random forest is a machine learning model for genomic prediction that uses an average of multiple decision trees to determine the predicted values. This regression model was implemented using the "randomForest" package (Breiman, 2001). The number of latent components for PLSR and decision trees for random forest was determined by a five-fold cross-validation to have a minimum prediction error.

BayesCpi was used to verify the influence of distinct genetic architectures of different traits on prediction accuracy. The BayesCpi assumes that each marker has a probability π of being included in the model, and this parameter is estimated at each Markov Chain Monte Carlo (MCMC) iteration. The vector of marker effects u is assumed to be a mixture of distributions having the probability π of being null effect and (1- π) of being a realization of a normal distribution, so that, $u_j|\pi, \sigma_g^2 \sim N(0, \sigma_g^2)$. The vector of residual effects was considered as $e \sim N(0, \sigma_e^2)$. The marker and residual variances were assumed to follow a chi-square distribution $\sigma_g^2 \sim \chi^2(S_b, \nu_0)$and $\sigma_e^2 \sim \chi^2(S_b, \nu_0)$, respectively, with $\nu_0 = 5$ degrees of freedom as prior and S_b shape parameters assuming a heritability of 0.5 (Pérez and de los Campos, 2014).

The last model used was the RKHS. The method is a regression where the estimated parameters are a linear function of the basis provided by the reproducing kernel (RK). RKHS considers both additive and non-additive genetic effects (de los Campos et al., 2013). In this work, the multi-kernel approach was used by averaging three kernels with distinct bandwidth values. In this implementation the averaged kernel, $\bar{K}$ was given by:

$\bar{K} = \sum_r K_r \sigma^2_{\beta_r} \tilde{\sigma}_\beta^{-2}$, where $\tilde{\sigma}^2_\beta = \sum \sigma^2_{\beta_r}$. Here r = 3 and $\sigma^2_{\beta_r}$ are interpretable as variance parameters associated with each kernel. Therefore, for each r^{th} kernel the proportion of sharing alleles between pairs of individuals (ii′) was given by $K_r = \exp\{-h_k d^2_{ii'}\}$, where h_k is a bandwidth parameter associated with r^{th} reproducing kernel and d^2_{iiC} is the genetic distance between individuals i and i′ computed as follows: $d^2_{ii'} = \sum_{j=1}^{p} (x_{ij} - x_{iCj})^2$, where j = 1, . . . , p markers stated as above. The bandwidth parameter values for the three kernels were h = 0.5{1/5,1,5, as suggested by (Pérez and de los Campos, 2014). Those values were chosen using the rule proposed by de los Campos et al. (2010).

Genomic prediction models RR-BLUP, PLSR, RF were carried out using "GSwGBS" package (Gaynor, 2015) while the BayesianCpi and RKHS were executed with the BGLR package (de los Campos et al., 2010). We calculated each genomic prediction model's predictive ability as the Pearson correlation between the estimated breeding values from model 1) (obtained using the full data set) and those of validation set predicted from the respective model. For that, we used a cross-validation scheme considering 80% of the observations, randomly selected, as training and the remaining 20% as validation set. The process was repeated 20 times for each model. From the predictive ability values, we estimated the confidence interval for this parameter using the bootstrap method considering 10,000 samples (James et al., 2013).

Determining Optimal Training Population Size

The influence of training population size on predictive ability was evaluated using a validation set comprising of 50 randomly selected lines and training sets of variable sizes. The validation set was formed by randomly sampling 50 lines without replacement. The training population of size n was formed sequentially by adding 25 accessions from the remaining accessions such that its size ranged between 50 and 175. We subset the collection into subgroups of 50, 75, 100, 125, 150, and 175 individuals each. The RR-BLUP model was used to predict each trait. This procedure was repeated 20 times, and accuracies of each training population size were averaged across 20 replicates. To predict a particular subpopulation with increasing population size, a similar procedure was followed to using variable training population size ranged from 50 to 175 with an increment of 25.

Determining Optimal Marker Density

To evaluate the effects of GBS marker selection on predictive ability, we randomly sampled markers five times with the following subset: one thousand (1 K), five thousand (5 K), ten

thousand (10 K), fifteen thousand (15 K), twenty thousand (20 K), twenty-five thousand (25 K), and thirty thousand (30 K). A random sampling of SNP was implemented to minimize or avoid any possible biases on sampling towards a particular distribution. Using the RR-BLUP model, a five-fold cross validation approach was used to obtain predictive ability in each marker subset. This procedure was repeated 20 times and predictive ability for each subset of SNPs were averaged across 20 replicates.

Accounting for Population Structure Into the Genomic Prediction Framework

We explored the confounding effect due to population structure on predictive ability. We investigated subpopulation structure on 482 accessions genotyped with 30,600 SNP markers using the ADMIXTURE clustering-based algorithm (Alexander et al., 2009). ADMIXTURE identifies K genetic clusters, where K is specified by the user, from the provided SNP data. For each individual, the ADMIXTURE method estimates the probability of membership to each cluster. An analysis was performed in multiple runs by inputting successive values of K from 2 to 10. The optimal K value was determined using ADMIXTURE's cross-validation (CV) error values. Based on >60% ancestry, each accession was classified into seven subpopulations (K = 7). Accessions within a subpopulation with membership coefficients of <60% were considered admixed. A total of eight subpopulations were used in this study, including admixed as a separate subpopulation. Principal component (PC) analysis was also conducted to summarize the genetic structure and variation present in the collection.

To account for the effect of population structure, we included the top 10 PC, or the Q-matrix from ADMIXTURE into the RR-BLUP model and performed five-fold cross-validation repeated 20 times. Alternatively, we also used the subpopulation (SP) designation identified by ADMIXTURE as a factor in the RR-BLUP model. Albeit a smaller population size, we also performed a within-subpopulation prediction. As stated above, a subpopulation was defined based on >60% ancestry cut-off. Only three subpopulations with this cut-off were identified and used with reasonable number of entries (e.g., N > 40): SP5 (N = 51), SP7 (N = 58), and SP8 (N = 41). A leave-one-SP-out was used to predict individuals within the subpopulation with the RR-BLUP model. We also used increasing population sizes to predict specific subpopulation (e.g. SP8) using the RR-BLUP model.

Estimating Reliability Criteria and Predicting Unknown Phenotypes

Nonphenotyped entries were predicted based on the RR-BLUP model using SNP markers only. The reliability criteria for each of the nonphenotyped lines were then calculated using the formula (Hayes et al., 2009; Clark et al., 2012) as follows:

$$r\,(\mathrm{PEV}) = \sqrt{\left(1 - \left(\frac{\mathrm{PEV}}{\sigma_G^2}\right)\right)} \qquad (6)$$

TABLE 1 | Heritability and summary statistics for seed yield and other agronomic traits.

Trait	Mean	Range	SD	CV(%)	H^2	H^2_{Cullis}
DFF (days)	71	60–84	4.8	6.7	0.90	0.80
NoSeedsPod (Nos.)	5.7	4.4–6.9	0.5	8.5	0.84	0.66
PH (cm)	74	37.6–108.3	11.5	15.5	0.81	0.68
PodsPlant (Nos.)	18	15–23	1.5	8.3	0.50	0.27
DM (days)	104	99–112	2.4	2.3	0.51	0.38
SeedYield (kg ha^{-1})	2,918	1734–4,463	451	15.4	0.67	0.46

DFF is days to first flowering; NoSeedsPod is the number of seeds per pod, PH is plant height, PodsPlant is the number of pods per plant, DM is days to physiological maturity, SeedYield is seed yield per hectare, SD is the standard deviation, CV is coefficient of variance, H^2 is heritability in the broad sense.

where PEV is the predicted error variance, and σ_G^2 is the genetic variance.

RESULTS

Phenotypic Heritability and Correlation

Recorded days to first flowering had a wide range of variability from 60 to 84 days with a mean of 71 days. The estimated heritability for days to first flowering was 0.90 using **Eq. 2** and 0.80 as per Cullis heritability using **Eq. 3** (**Table 1**). For the number of seeds per pod, the mean was 5.7 with a heritability estimate of 0.84 (H^2_{Cullis} = 0.66). The heritability for plant height was 0.81 (H^2_{Cullis} = 0.68), with an average height of 74 cm. The number of pods per plant had a heritability estimate of 0.50 (H^2_{Cullis} = 0.27) with a mean of 18 number of pods per plant and ranged from 15 to 23 pods. Days to physiological maturity had a mean of 104 days with an estimated heritability of 0.51 (H^2_{Cullis} = 0.38). Seed yield per hectare ranged widely from 1734 to 4,463 kg ha^{-1} with a mean yield of 2,918 kg ha^{-1} and a heritability value of 0.67 (H^2_{Cullis} = 0.46). The number of pods per plant was highly and positively correlated with seed yield. Correlation estimation also suggested seed yield was positively correlated with plant height, days to physiological maturity, and days to first flowering (**Supplementary Figure S1**).

Predictive Ability of Different Genomic Prediction Models

No single model consistently performed best across all traits that we evaluated (**Table 2**), however Bayesian model BayesCpi, RKHS, and RR-BLUP, in general, tended to generate better results. Roughly the predictive abilities from different models were similar, although slight observed differences were likely due to variations on genetic architecture and the model's assumptions underlying them. For days to first flowering, the highest predictive ability was obtained from the RR-BLUP (0.60). RR-BLUP, RF, and RKHS models generated the highest predictive ability for number of pods per plant (0.28). Number of seeds per pod was better predicted by RR-BLUP and Bayes Cpi (0.42). For plant height highest prediction accuracies were obtained from RF and BayesCpi (0.45). BaysCpi also gave the highest prediction accuracies for days to physiological maturity (0.47). For seed

TABLE 2 | Predictive ability for seed yield and agronomic traits using five genomic prediction models.

Traits	RR-BLUP	PLSR	RF	BayesCpi	RKHS
DFF (days)	0.60 (0.57–0.63)	0.57 (0.53–0.61)	0.55 (0.52–0.58)	0.59 (0.55–0.63)	0.54 (0.5–0.58)
NoSeedsPod	0.42 (0.37–0.48)	0.41 (0.36–0.46)	0.40 (0.35–0.45)	0.42 (0.38–0.46)	0.40 (0.34–0.48)
PH (cm)	0.39 (0.33–0.44)	0.42 (0.38–0.48)	0.45 (0.4–0.5)	0.45 (0.41–0.48)	0.43 (0.39–0.48)
PodsPlant	0.28 (0.22–0.33)	0.25 (0.2–0.31)	0.28 (0.22–0.34)	0.23 (0.17–0.29)	0.28 (0.23–0.34)
DM (days)	0.42 (0.36–0.47)	0.44 (0.39–0.5)	0.41 (0.35–0.46)	0.47 (0.43–0.5)	0.45 (0.4–0.48)
SeedYield (kg ha^{-1})	0.38 (0.34–0.42)	0.31 (0.27–0.36)	0.39 (0.35–0.44)	0.35 (0.31–0.39)	0.42 (0.37–0.48)

DFF is days to first flowering; NoSeedsPod is the number of seeds per pod; PH is Plant height in cm, PodsPlant is the number of pods per plant; DM is days to physiological maturity; within parentheses are ranges of predictive ability.

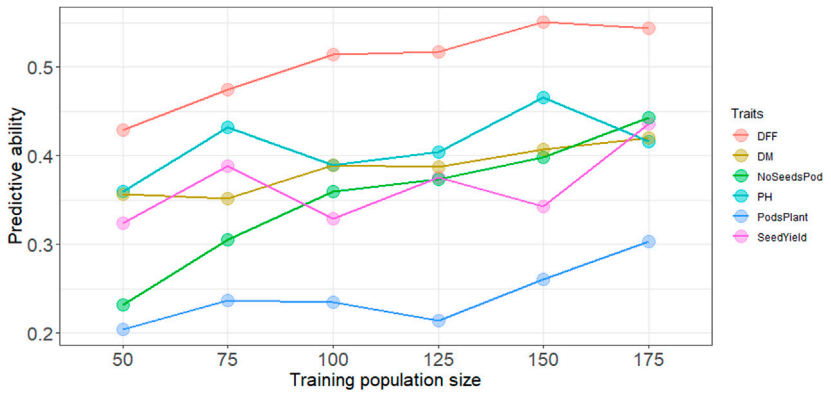

FIGURE 1 | Predictive ability with increasing training population size using the RR-BLUP model, DFF is days to first flowering, DM, is days to physiological maturity, NoSeedsPod is number of seeds per pod, PH is plant height in cm, PodsPlant is the number of pods per plant, SeedYield is seed yield in kg ha^{-1}.

yield, RKHS had slight advantages over other models (0.42). As mentioned above, some differences between the model's accuracy were only marginal and cannot be a criterion for choosing one model (**Table 2**). For example, among the tested models, the highest difference in predictive accuracy, considering number of seeds per pod, had a magnitude of 0.02, a marginal value. The lack of significant differences among genomic prediction models can be interpreted as either a good approximation to the optimal model by all methods or there may be a need for further research (Yu et al., 2016). Unless indicated otherwise, the rest of our results focused on findings from the RR-BLUP model.

Determining Optimal Number of Individuals

Increasing the training population size led to a slight increase in the predictive ability overall for all traits. Across all traits except days to first flowering and plant height, predictive ability reached a maximum with the largest training population size of N = 175 (**Figure 1**). A training population comprised of 50 individuals had the lowest predictive ability across all traits. For days to first flowering, and plant height predictive ability did steadily increase up at N = 150, and prediction ability reached the maximum for most traits at highest training population size with N = 175. Regardless of population size, predictive ability was consistently higher for days to first flowering, whereas predictive ability was consistently lower for pods per plant (**Figure 1**). However, while predicting subpopulation 5 highest predictive ability was obtained for plant height (**Supplementary Figure S3**).

Determining Optimal Marker Density

The different marker subsets had insignificant differences on predictive ability for all the traits evaluated in this study. In general, however, predictive abilities were higher between 5K and 15K SNPs and reached a plateau with increasing number of SNPs (**Supplementary Figure S2**). For seed yield, plant height, and days to maturity, highest predictive ability were 0.38, 0.39, and 0.42 respectively. The highest predictive ability for days to first flowering was 0.61 using a SNP subset of 15K.

Accounting for Population Structure in the Genomic Prediction Model

Population structure explained some portion of the phenotypic variance, ranging from 9 to 19%, with the highest percentages observed for plant height (19%) and seed yield (17%). Using either ADMIXTURE or PCA to account for the effect due to population structure, we improved the predictive ability. We observed a 6% improvement for days to first flowering and 32% for seed yield compared over models that did not account for population structure.

We also performed within subpopulation predictions. Presented here are the predictive abilities for subpopulations 5, 7, and 8, as they had at least 40 entries. Subpopulation 8 had the

TABLE 3 | Predictive ability within and across subpopulations using RR-BLUP and all SNP markers.

Sub pops	DFF	NoSeedsPod	PH	PodsPlant	DM	SeedYield
Sub pop 5 (51)	0.27	0.26	0.08	-0.01	0.02	0.18
Sub pop 7 (58)	0.34	0.40	0.22	0.12	-0.01	0.01
Sub pop 8 (41)	0.68	0.35	0.33	0.07	0.43	0.37
SP-	0.50	0.45	0.47	0.25	0.51	0.34
SP+	0.53	0.35	0.42	0.25	0.48	0.45
SP PC10	0.51	0.41	0.44	0.18	0.20	0.43
Var exp (R^2)	0.13	0.09	0.19	0.15	0.15	0.17

DFF is days to first flowering, NoSeedsPod is the number of seeds per pod, PH is plant height, PodsPlant is the number of pods per plant, DM is days to physiological maturity, SP- does not account for population structure, SP+, refers to the population structure addressed in the model, SP PC10 addresses population structure with 10 PC, Var exp (R^2) refers the variance explained by population structure after fitting a regression model, within parenthesis represent the number of entries in each subpopulation.

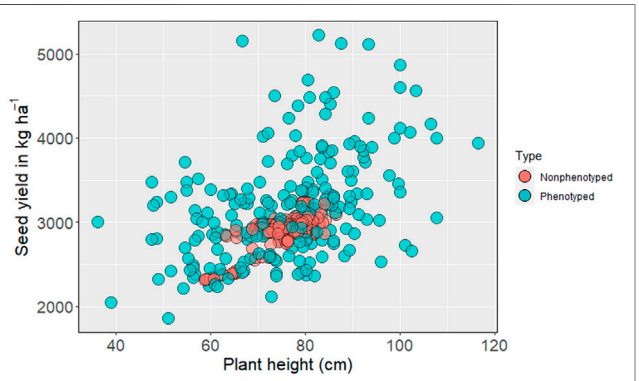

FIGURE 2 | Distribution of phenotyped and predicted nonphenotyped accessions within the USDA pea germplasm collection for seed yield and plant height.

highest predictive ability for days to first flowering (0.68), plant height (0.33), days to maturity (0.43), and seed yield (0.37). The highest predictive abilities for the number of seeds per pod (0.40) and pods per plant (0.12) were obtained from subpopulation 7 (**Table 3**). Notably, predictive ability was generally higher when all germplasm sets or subpopulations were included in the model compared to when predictions were made using a subset of germplasm.

Predicting Genotyped but Nonphenotyped Accessions

The genomic prediction model was then used to predict nonphenotyped entries based on their SNP information. Based on the distribution of GEBV, none of the predicted phenotypes for nonphenotyped accessions exceeded the top-performing observed phenotypes for seed yield (**Figure 2**). The mean seed yield of predicted entries (2,914 kg ha^{-1}) was not significantly different from the mean seed yield of observed genotypes (2,918 kg ha^{-1}). The mean of observed and predicted entries were non-significant for the other five traits (**Supplementary Table S1**). The GEBV for number of pods per plant, number of seeds per pod (**Supplementary Figures S4, S5**), days to first flowering, and days to maturity all fall within the range of observed phenotypes (Similar Figures not added).

Reliability Estimation

We obtained reliability criteria for all traits, including seed yield and phenology, for 244 nonphenotyped accessions. The average reliability values ranged from 0.30 to 0.35, while the highest values for evaluated traits ranged from 0.75 to 0.78. The higher reliability values were distributed in the top, bottom, and intermediate predicted breeding values (**Supplementary Tables S2–S7**). For seed yield (kg ha^{-1}), the highest reliability was obtained from the bottom 50 (**Figure 3**). Higher reliability criteria are primarily distributed among the intermediate and top GEBV for days to first flowering. Predicted intermediate plant height showed the highest reliability, as presented in **Figure 3**.

DISCUSSION

Widely utilized plant genetic resources collections, such as the USDA pea germplasm collection, hold immense potential as diverse genetic resources to help guard against genetic erosion and serve as unique sources of genetic diversity from which we could enhance genetic gain, boost crop production, and help reduce crop losses due to disease, pests, and abiotic stresses (Jarquin et al., 2016; Crossa et al., 2017; Holdsworth et al., 2017; Mascher et al., 2019). As the costs associated with genotyping on a broader and more accurate scale continue to decrease, opportunities increase to evaluate and utilize these collections in plant breeding. Relying on phenotypic evaluation alone can be costly, rigorous, and time-intensive. However, by incorporating high-density marker coverage and efficient computational algorithms, we can better realize the potential for utilizing these germplasm stocks by reducing the time and cost associated with their evaluation (Yu et al., 2016; H. Li et al., 2018; Yu et al., 2020). In this study, we evaluated the potential of genotyping-by-sequencing derived SNPs for genomic prediction. We found that it holds promises for extracting useful diversity from germplasm collections for applied breeding efforts.

In this study, predictive ability was generally similar among methods, and there was no single model that worked across traits, consistent with results obtained by other authors (Burstin et al., 2015; Spindel et al., 2015; Yu et al., 2016; Azodi et al., 2019). For example, considering only the punctual estimates, RR-BLUP model was the best for days to first flowering, however for plant height, days to physiological maturity, and seed yield, the best models were BayesCpi and RF, BayesCpi and RKHS,

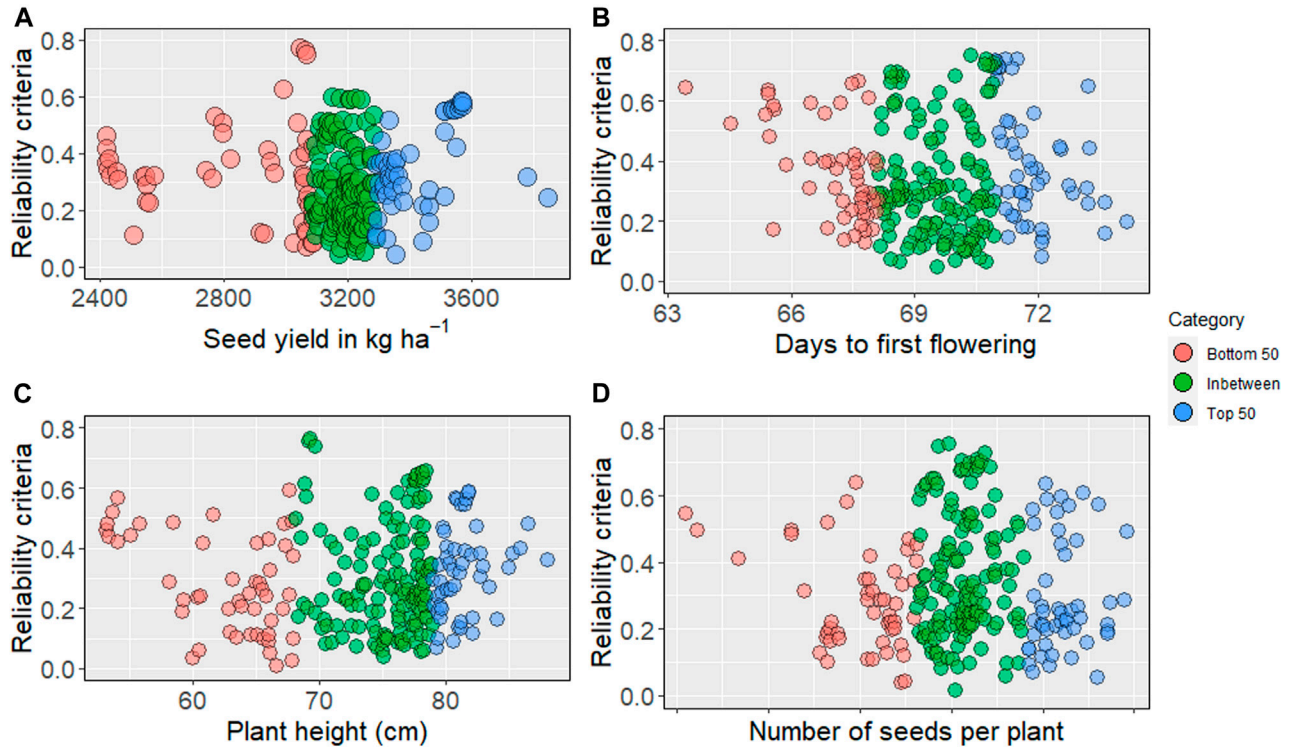

FIGURE 3 | Reliability criteria for nonphenotyped lines: the top 50 of genomic estimated breeding values are blue, and bottom 50 are in red, intermediates are in green. **(A)** reliability estimates for seed yield (kg ha^{-1}), **(B)** days to first flowering, **(C)** plant height, **(D)** number of seeds per plant.

respectively. In recent work, Azodi et al. (2019) compared 12 models (6 linear and 6 non-linear) considering 3 traits in 6 different plant species, and they did not find any best algorithm for all traits across all species. Newer statistical methods are expected to boost prediction accuracy; however, the biological complexity and unique genetic architecture of traits can be regarded as the root cause for getting zero or slight improvement on prediction accuracy (Valluru et al., 2019; Yu et al., 2020). As data collection accelerates in at different levels of biological organization (Kremling et al., 2019), genomic prediction models will expand and nonparametric models, including machine learning, may play an essential role for boosting prediction accuracy (Azodi et al., 2019; Yu et al., 2020).

A related work in pea has been published but only based on a limited number of markers (Burstin et al., 2015). This work assessed genomic prediction models in a diverse collection of 373 pea accessions with 331 SNP markers and found no single best model across traits, which is consistent with our findings. In this work, the authors reported that traits with higher heritability, such as thousand seed weight and flowering date, had higher prediction accuracy. We also verified days to first flowering as having the highest heritability and predictive accuracies through all the models. Interestingly, yield components like the number of seeds per pod and pods per plant showed lower predictive accuracy, regardless of prediction models used. Consistent with our results, Burstin et al. (2015) also found yield components like seed number per plant as having lower predictive accuracy and higher standard deviation for prediction. These traits are highly complex and largely influenced by the environment.

The predictive ability increased for all traits except plant height when we increased the model's training population size, suggesting that adding more entries in the study can boost predictive ability. By accounting population structure into genomic prediction framework, we observed an improved prediction accuracy for some traits—seed yield and days to first flowering—but not for other traits. Although the population structure explained 9–19% of the phenotypic variance, we cannot fully and conclusively answer the effect of population structure in prediction accuracy due to smaller population size. In addition, accounting for the relatedness among individuals in the training and testing sets can potentially boost prediction accuracy (Riedelsheimer et al., 2013; Lorenz and Smith, 2015; Rutkoshi et al., 2015); it was outside the scope of this research but deserves further study. Adding more environments (year-by-location combination) can also potentially improve prediction accuracy using genomic prediction frameworks that account for genotype-by-environment interactions and/or phenotypic plasticity (Jarquin et al., 2014; Crossa et al., 2017; X. Li et al., 2018; Guo et al., 2020). In general, we observed that predictive ability slightly increased and plateaued after reaching certain subset of SNPs. Such a plateau on prediction ability maybe due to overfitting of models (Hickey et al., 2014; Norman et al., 2018), presumably due to extensive linkage disequilibrium in the pea genome (Kreplak et al., 2019).

Previous studies have indicated the importance of considering reliability values when using predictive ability values to select

genotypes (Yu et al., 2016). We found higher reliability estimates were spread across all GEBVs rather than clustering around higher or lower extreme of GEBVs. Those accessions with top predicted values and high reliability estimates maybe selected as candidate parents for increasing seed yield and/or germplasm enhancement. However, for a trait such as days to flowering in pea, even low or intermediate predicted values maybe suitable candidates when paired with high reliability values. We found the means of GEBV for nonphenotyped entries were non-significantly different with phenotyped accessions, and almost none of nonphenotyped accessions were expected to exceed seed yield of phenotyped accessions. Several accessions in the USDA pea germplasm collection can be readily incorporated into breeding programs for germplasm enhancement by incorporating above-average accessions with high or moderately high reliability values (Yu et al., 2020).

CONCLUSIONS AND RESEARCH DIRECTIONS

The research findings demonstrated that the wealth of genetic diversity available in a germplasm collection could be assessed efficiently and quickly using genomic prediction to identify valuable germplasm accessions that can be used for applied breeding efforts. With the integration of more orthogonal information (e.g., expression, metabolomics, proteomics, etc.) into genomic prediction framework (Kremling et al., 2019; Valluru et al., 2019) coupled with the implementation of more complex genomic selection models like a multivariate genomic selection approach (Rutkoski et al., 2015), we can considerably enhance predictive ability. This research framework could greatly contribute to help discover and extract useful diversity targeting high-value quality traits such as protein and mineral concentrations from a large germplasm collection in the future.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. The SNP dataset can be accessed here: https://www.ncbi.nlm.nih.gov/sra/PRJNA730349.

AUTHOR CONTRIBUTIONS

NB, CJC, and MAB conceived and designed the manuscript. CJC, DM, and RJM designed and executed the field and genotyping experiments. YM and PZ performed DNA extraction, constructed the library, and called SNPs. MAB, IV, and SS analyzed data, curated SNPs, and ran genomic prediction models. NB oversaw statistical analyses. MAB, HW, IV, and NB wrote and edited the overall manuscript. All authors edited, reviewed, and approved the manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge the funding provided by USDA Plant Genetic Resource Evaluation Grant for the GS analysis, USA Dry Pea and Lentil Council Research Committee for the field phenotyping, and USDA ARS Pulse Crop Health Initiative for the SNP genotyping and support from USDA ARS Project: 5348-21000-017-00D (CC), and 5348-21000-024-00D (RJM). We also acknowledge the support from USDA-NIFA (Hatch Project ND01513 for NB) and the North Dakota Department of Agriculture through the Specialty Crop Block Grant Program (19–429). Technical assistance from Britton Bourland, Lydia Fields, Kurt Tetrick, and Jennifer Morris is gratefully acknowledged. This work used resources of the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University, Fargo, ND, USA which were made possible in part by NSF MRI Award No. 2019077.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.707754/full#supplementary-material

Supplementary Figure S1 | Phenotypic correlation among seed yield and agronomic traits evaluated in this study, DFF is days to first flowering, PH is plant height in cm, SeedYield is seed yield in kg ha^{-1}, DM is the days to physiological maturity

Supplementary Figure S2 | Predictive ability with increasing SNP markers RR-BLUP model, DFF is days to first flowering, DM, is days to physiological maturity, NoSeedsPod is number of seeds per pod, PH is plant height in cm, PodsPlant is pods per plant, SeedYield is seed yield in kg ha^{-1}

Supplementary Figure S3 | Predictive ability of subpopulation 5 with increasing training population size

Supplementary Figure S4 | Distribution of phenotyped and predicted non phenotyped accessions for seed yield and number of pods per plant in the USDA germplasm collections

Supplementary Figure S5 | Distribution of phenotyped and predicted non phenotyped accessions for seed yield and number of seeds per pod in the USDA germplasm collections

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Annicchiarico, P., Nazzicari, N., Laouar, M., Thami-Alami, I., Romani, M., and Pecetti, L. (2020). Development and Proof-Of-Concept Application of Genome-Enabled Selection for Pea Grain Yield under Severe Terminal Drought. *Ijms* 21 (7), 2414–2420. doi:10.3390/ijms21072414

Annicchiarico, P., Nazzicari, N., Pecetti, L., Romani, M., and Russi, L. (2019). Pea Genomic Selection for Italian Environments. *BMC Genomics* 20 (1), 1–18. doi:10.1186/s12864-019-5920-x

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., and Shiu, S.-H. (2019). Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3: Genes, Genomes, Genet.* 9 (11), 3691–3702. doi:10.1534/g3.119.400498

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. *J. Stat. Soft.* 67 (1), 1–48. doi:10.18637/jss.v067.i01

Bethke, P. C., Halterman, D. A., and Jansky, S. H. (2019). Potato Germplasm Enhancement Enters the Genomics Era. *Agronomy* 9 (10), 575. doi:10.3390/agronomy9100575

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples. *Bioinformatics* 23 (19), 2633–2635. doi:10.1093/bioinformatics/btm308

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/A:1010933404324

Burstin, J., Salloignon, P., Chabert-Martinello, M., Magnin-Robert, Jean. Bernard., Magnin-Robert, J.-B., Siol, M., et al. (2015). Genetic Diversity and Trait Genomic Prediction in a Pea Diversity Panel. *BMC Genomics* 16 (1), 1–17. doi:10.1186/s12864-015-1266-1

Cheng, P., Holdsworth, W., Ma, Y., Coyne, C. J., Mazourek, M., Grusak, M. A., et al. (2015). Association Mapping of Agronomic and Quality Traits in USDA Pea Single-Plant Collection. *Mol. Breed.* 35 (2). doi:10.1007/s11032-015-0277-6

Clark, S. A., Hickey, J. M., Daetwyler, Hans. D., Daetwyler, H. D., and van der Werf, J. H. (2012). The Importance of Information on Relatives for the Prediction of Genomic Breeding Values and the Implications for the Makeup of Reference Data Sets in Livestock Breeding Schemes. *Genet. Sel. Evol.* 44 (1), 4. doi:10.1186/1297-9686-44-4

Colombani, C., Croiseau, P., Fritz, S., Guillaume, F., Legarra, A., Ducrocq, V., et al. (2012). A Comparison of Partial Least Squares (PLS) and Sparse PLS Regressions in Genomic Selection in French Dairy Cattle. *J. Dairy Sci.* 95 (4), 2120–2131. doi:10.3168/jds.2011-4647

Covarrubias-Pazaran, G. (2016). Genome-Assisted Prediction of Quantitative Traits Using the R Package Sommer. *PLoS ONE* 11 (6), e0156744–15. doi:10.1371/journal.pone.0156744

Coyne, C. J., Brown, A. F., Timmerman-Vaughan, G. M., McPhee, K. E., and Grusak, M. A. (2005). USDA-ARS Refined Pea Core Collection for 26 Quantitative Traits. *Pisum Genet.* 37 (11), 1–4.

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., et al. (2016). Genomic Prediction of Gene Bank Wheat Landraces. *G3: Genes, Genomes, Genet.* 6 (7), 1819–1834. doi:10.1534/g3.116.029637

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* 22, 961–975. doi:10.1016/j.tplants.2017.08.011

Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006). On the Design of Early Generation Variety Trials with Correlated Data. *Jabes* 11 (4), 381–393. doi:10.1198/108571106X154443

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The Variant Call Format and VCFtools. *Bioinformatics* 27 (15), 2156–2158. doi:10.1093/bioinformatics/btr330

de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., Crossa, J., Weigel, Kent. A., et al. (2010). Semi-Parametric Genomic-Enabled Prediction of Genetic Values Using Reproducing Kernel Hilbert Spaces Methods. *Genet. Res.* 92 (4), 295–308. doi:10.1017/S0016672310000285

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole-genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* 193, 327–345. doi:10.1534/genetics.112.143313

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A Robust, Simple Genotyping-By-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6 (5), e19379–10. doi:10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* 4 (3), 250–255. doi:10.3835/plantgenome2011.08.0024

Facciolongo, Anna. Maria., Rubino, Giuseppe., Zarrilli, Antonia., Vicenti, Arcangelo., Ragni, Marco., and Toteda, Francesco. (2014). Alternative Protein Sources in Lamb Feeding 1. Effects on Productive Performances, Carcass Characteristics and Energy and Protein Metabolism. *Prog. Nutr.* 16 (2), 105–115.

Garrison, E., and Marth, G. (2012). Haplotype-based Variant Detection from Short-Read Sequencing. ArXiv: 1207.3907 [q-Bio]Retrieved from: http://arxiv.org/abs/1207.3907.

Gaynor, R. C. (2015). "GSwGBS: An R Package Genomic Selection with Genotyping-By-Sequencing," in *Genomic Selection for Kansas Wheat* (Manhattan, KS: K-State Research Exchange).

Gorjanc, G., Jenko, J., Hearne, S. J., and Hickey, J. M. (2016). Initiating Maize Pre-breeding Programs Using Genomic Selection to Harness Polygenic Variation from Landrace Populations. *BMC Genomics* 17 (1), 1–15. doi:10.1186/s12864-015-2345-z

Guo, J., Pradhan, S., Shahi, D., Khan, J., Mcbreen, J., Bai, G., et al. (2020). Increased Prediction Accuracy Using Combined Genomic Information and Physiological Traits in A Soft Wheat Panel Evaluated in Multi-Environments. *Sci. Rep.* 10 (1), 1–12. doi:10.1038/s41598-020-63919-3

Habier, D., Fernando, R. L., and Dekkers, J. C. M. 2007. "The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values, 177, 2389, 2397. doi:10.1534/genetics.107.081190

Haile, T. A., Heidecker, T., Wright, D., Neupane, S., Ramsay, L., Vandenberg, A., et al. (2020). Genomic Selection for Lentil Breeding: Empirical Evidence. *Plant Genome* 13 (1), 1–15. doi:10.1002/tpg2.20002

Hallauer, A. R., Carena, M. J., and Miranda Fo, J. B. (2010). *Hand Book of Plant Breeding: Quantitative Genetics in maize Breeding*. 3rd ed. New York: Springer.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited Review: Genomic Selection in Dairy Cattle: Progress and Challenges. *J. Dairy Sci.* 92 (2), 433–443. doi:10.3168/jds.2008-1646

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. *Crop Sci.* 54 (4), 1476–1488. doi:10.2135/cropsci2013.03.0195

Holdsworth, W. L., Gazave, E., Cheng, P., Myers, J. R., Gore, M. A., Coyne, C. J., et al. (2017). A Community Resource for Exploring and Utilizing Genetic Diversity in the USDA Pea Single Plant Plus Collection. *Hortic. Res.* 4 (January), 1–13. doi:10.1038/hortres.2017.17

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

Jarquin, D., Specht, J., and Lorenz, A. (2016). Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions. *G3: Genes, Genomes, Genet.* 6 (8), 2329–2341. doi:10.1534/g3.116.031443

Kremling, K. A. G., Diepenbrock, C. H., Gore, M. A., Buckler, E. S., and Bandillo, N. B. (2019). Transcriptome-Wide Association Supplements Genome-wide Association in Zea mays. *G* 9, 3023–3033. doi:10.1534/g3.119.400549

Kreplak, J., Madoui, M.-A., Cápal, P., Novák, P., Labadie, K., Aubert, G., et al. (2019). A Reference Genome for Pea Provides Insight into Legume Genome Evolution. *Nat. Genet.* 51 (9), 1411–1422. doi:10.1038/s41588-019-0480-1

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/map Format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi:10.1093/bioinformatics/btp352

Li, H., Rasheed, A., Hickey, L. T., and He, Z. (2018). Fast-Forwarding Genetic Gain. *Trends Plant Sci.* 23 (3), 184–186. doi:10.1016/j.tplants.2018.01.007

Li, X., Guo, T., Mu, Q., Li, X., and Yu, J. (2018). Genomic and Environmental Determinants and Their Interplay Underlying Phenotypic Plasticity. *Proc. Natl. Acad. Sci. USA* 115 (26), 6679–6684. doi:10.1073/pnas.1718326115

Longin, C. F. H., Friedrich, H., and Reif., J. C. (2014). Redesigning the Exploitation of Wheat Genetic Resources. *Trends Plant Sci.* 19 (10), 631–636. doi:10.1016/j.tplants.2014.06.012

Lorenz, A. J., and Smith, K. P. (2015). Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. *Crop Sci.* 55, 2657–2667. doi:10.2135/cropsci2014.12.0827

Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., and Stein, N. (2019). Genebank Genomics Bridges the Gap between the Conservation of Crop Diversity and Plant Breeding. *Nat. Genet.* 51 (7), 1076–1081. doi:10.1038/s41588-019-0443-6

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M E. 2001. "Prediction of Total Genetic Value Using Genome-wide Dense Marker Maps." doi:10.1093/genetics/157.4.1819

Mevik, B.-H., and Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *J. Stat. Soft.* 18 (2), 1–23. doi:10.18637/jss.v018.i02

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., and Myles, S. (2015). LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3: Genes, Genomes, Genet.* 5 (11), 2383–2390. doi:10.1534/g3.115.021667

Mudryj, A. N., Yu, N., and Aukema, H. M. (2014). Nutritional and Health Benefits of Pulses. *Appl. Physiol. Nutr. Metab.* 39 (11), 1197–1204. doi:10.1139/apnm-2013-0557

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3: Genes, Genomes, Genet.* 8 (9), 2889–2899. doi:10.1534/g3.118.200311

Pérez, P., and de los Campos, G. (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* 198 (2), 483–495. doi:10.1534/genetics.114.164442

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Riedelsheimer, C., Brotman, Y., Méret, M., Melchinger, A. E., and Willmitzer, L. (2013a). The Maize Leaf Lipidome Shows Multilevel Genetic Control and High Predictive Value for Agronomic Traits. *Sci. Rep.* 3, 1–7. doi:10.1038/srep02479

Riedelsheimer, C., Yariv, B., Michaël, M., Albrecht, E. M., and Lothar, W. 2013b. "The Maize Leaf Lipidome Shows Multilevel Genetic Control and High Predictive Value for Agronomic Traits." *Scientific Rep.* 3: 1–7. doi:10.1038/srep02479

Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L., et al. (2015). Efficient Use of Historical Data for Genomic Selection: A Case Study of Stem Rust Resistance in Wheat. *Plant Genome* 8 (1), 1–10. doi:10.3835/plantgenome2014.09.0046

Simson, C. J., and Hannan, R. M. (1995). Development and Use of Core Subsets of Cool-Season Food Legume Germplasm Collections. *HortScience* 30, 907. doi:10.21273/HORTSCI.30.4.907C

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., et al. (2015). Genomic Selection and Association Mapping in Rice (*Oryza Sativa*): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. *Plos Genet.* 11 (2), e1004982–25. doi:10.1371/journal.pgen.1004982

Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., et al. (2015). Genomic Prediction in Pea: Effect of Marker Density and Training Population Size and Composition on Prediction Accuracy. *Front. Plant Sci.* 6 (NOVEMBER), 1–11. doi:10.3389/fpls.2015.00941

USDA (2020). *United States Acreage*, 1–50. Available at: https://www.nass.usda.gov/Publications/Todays_Reports/reports/acrg0620.pdf.

Valluru, R., Gazave, E. E., Fernandes, S. B., Ferguson, J. N., Lozano, R., Hirannaiah, P., et al. (2019). Deleterious Mutation burden and its Association with Complex Traits in Sorghum (*Sorghum Bicolor*). *Genetics* 211 (3), 1075–1087. doi:10.1534/genetics.118.301742

Vandemark, G. J., Brick, M., Osorno, J. M., Kelly, D. J., and Urrea, C. A. (2014). "Edible Grain Legumes," in *Yield Grains in Major U.S. Field Crops*. Editors S Smith, B. Diers, J. Speecht, and B. Carver (Madison, WI: CSSA), 87–123. doi:10.3390/cli6020041

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi:10.3168/jds.2007-0980

Wickham, H. (2016). *ggplot2:* Elegant Graphics For Data Analysis. New York: Springer-Verlag.

Yu, X., Leiboff, S., Li, X., Guo, T., Ronning, N., Zhang, X., et al. (2020). Genomic Prediction of Maize Microphenotypes Provides Insights for Optimizing Selection and Mining Diversity. *Plant Biotechnol. J.* 18, 2456–2465. doi:10.1111/pbi.13420

Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S. E., et al. (2016). Genomic Prediction Contributing to a Promising Global Strategy to Turbocharge Gene Banks. *Nat. Plants* 2 (10). doi:10.1038/nplants.2016.150

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership