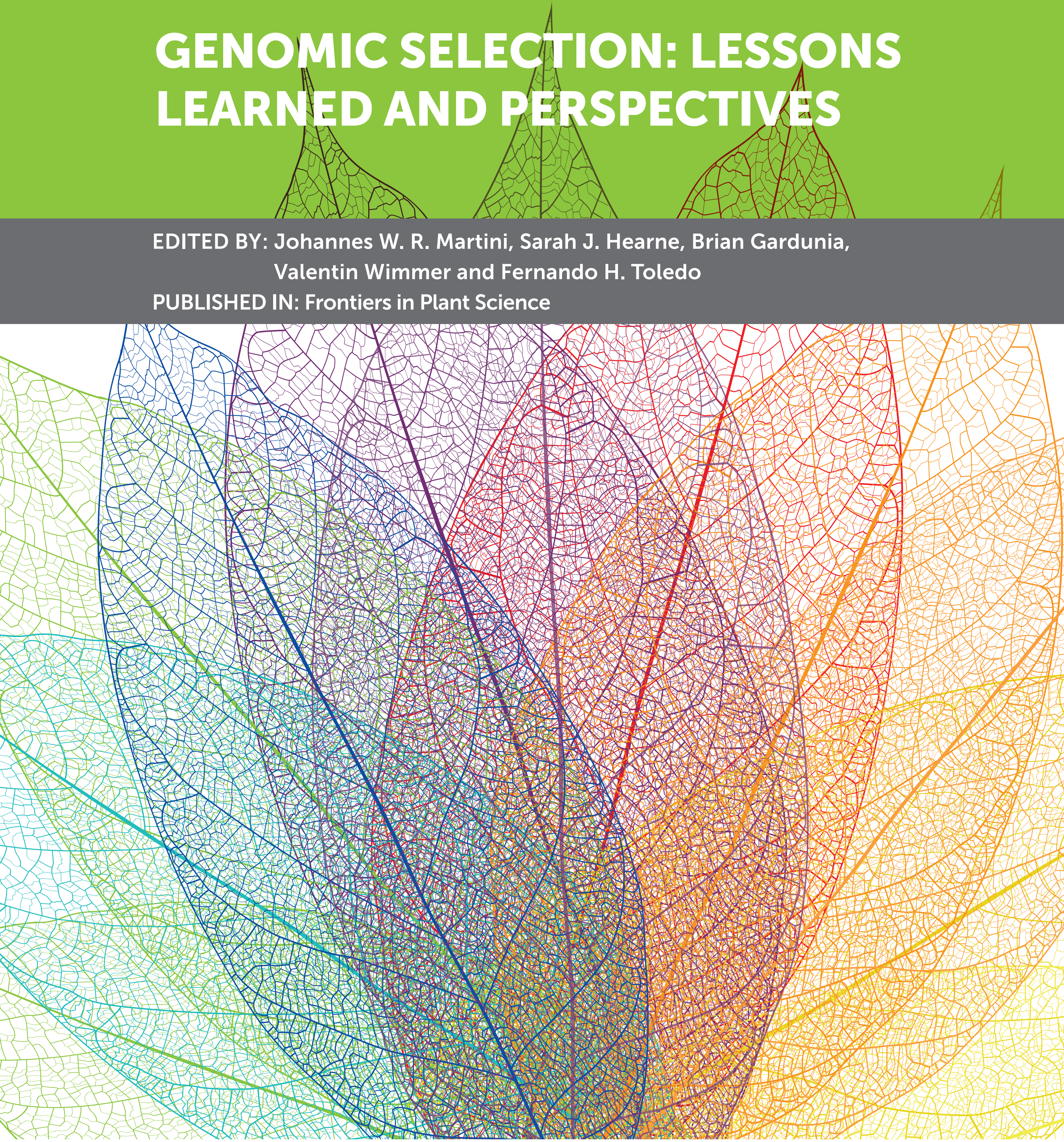
GENOMIC SELECTION: LESSONS LEARNED AND PERSPECTIVES

EDITED BY: Johannes W. R. Martini, Sarah J. Hearne, Brian Gardunia,
Valentin Wimmer and Fernando H. Toledo

PUBLISHED IN: Frontiers in Plant Science



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-674-3
DOI 10.3389/978-2-88974-674-3

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

GENOMIC SELECTION: LESSONS LEARNED AND PERSPECTIVES

Topic Editors:
Johannes W. R. Martini, International Maize and Wheat Improvement Center (Mexico), Mexico
Sarah J. Hearne, International Maize and Wheat Improvement Center (Mexico), Mexico
Brian Gardunia, Bayer Crop Science (United States), United States
Valentin Wimmer, KWS Saat (Germany), Germany
Fernando H. Toledo, International Maize and Wheat Improvement Center (Mexico), Mexico

Topic Editor Valentin Wimmer is affiliated to KWS SAAT SE & Co. KGaA, Germany. Topic Editor Brian Gardunia is affiliated to Bayer Crop Sciences and has a collaboration with AbacusBio, and is an author on patents with Bayer Crop Sciences.

The other Topic Editors did not disclose any conflicts of interest.

Citation: Martini, J. W. R., Hearne, S. J., Gardunia, B., Wimmer, V., Toledo, F. H., eds. (2022). Genomic Selection: Lessons Learned and Perspectives. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-674-3

Table of Contents

EDITORIAL
published: 27 May 2022
doi: 10.3389/fpls.2022.890434

Editorial: Genomic Selection: Lessons Learned and Perspectives

Johannes W. R. Martini[1]*, *Sarah J. Hearne*[1,2], *Brian Gardunia*[3], *Valentin Wimmer*[4] and *Fernando H. Toledo*[1]

[1] *International Maize and Wheat Improvement Center, Texcoco, Mexico,* [2] *Excellence in Breeding Platform, Texcoco, Mexico,* [3] *Bayer Crop Science, St. Louis, MO, United States,* [4] *KWS SAAT SE & Co., KGaA, Einbeck, Germany*

Keywords: genomic selection (GS), plant breeding, selection gain, breeding schemes, genotype-by-environment interaction

Editorial on the Research Topic

Genomic Selection: Lessons Learned and Perspectives

Genomic selection (GS) has been one of the most prominent Research Topics in breeding science in the last two decades after the milestone paper by Meuwissen et al. (2001). Its huge potential for increasing the efficiency of breeding programs attracted scientific curiosity and research funding. Many different statistical prediction methods have been tested, and different use cases have been explored.

We organized this Research Topic to look both back and forward. The objectives were to review the developments of the last 20 years, to provide a snapshot of current hot topics, and potentially also to define areas on which more (or less) focus should be put in the future, thereby supporting readers with formulating and prioritizing their ideas for future research.

Several questions were brought up when organizing this Research Topic including: How did GS change breeding schemes? Which impact did GS have on realized selection gain? What, considering the context of particularities of different crops, may be optimal breeding schemes to leverage the full potential of GS? What has been the impact of and what is the potential of hybrid prediction, statistical epistasis models, deep learning and other methods? What are the long-term effects of GS? Can predictive breeding approaches also be used to harness genetic resources from germplasm banks in a more efficient way?

Having closed our Research Topic, we are happy to present a solid collection of 21 contributions from 149 authors which reviews the past work around GS, presents new insights, and points at topics with potential for future research. The 21 contributions consist of 12 original research articles, a method paper, two review contributions, five opinion articles and a perspective.

Concerning original research, the main topics that have been addressed were "genetic architecture" and "genetic architecture enhanced prediction methods," "shortening the breeding cycle," "genotype x environment interaction," "sparse-testing," and "genomic selection in polyploids."

Additionally to considerations around GS for major staple crops, Ferrão et al. "propose a strategy for using genomic selection in blueberry, with the potential to be applied to other polyploid species of a similar background." In particular, the authors highlight that "the use of additive effects under a linear mixed model framework (GBLUP) showed the best balance between efficiency and accuracy." The topic of GS in tetraploids has also been considered by Wilson et al. for the case of potato. Moreover, Liu et al. investigated prediction methods based on genes known to be relevant for fiber length in cotton. Pégard et al. considered GS for poplar in the context of forest tree

Edited by:
Mallikarjuna Swamy,
International Rice Research Institute (IRRI), Philippines

Reviewed by:
Krishnamurthy S. L.,
Central Soil Salinity Research Institute (ICAR), India
Balram Marathi,
Professor Jayashankar Telangana State Agricultural University, India

****Correspondence:***
Johannes W. R. Martini
jwrmartini@gmail.com

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *05 March 2022*
Accepted: *08 April 2022*
Published: *27 May 2022*

Citation:
Martini JWR, Hearne SJ, Gardunia B, Wimmer V and Toledo FH (2022) Editorial: Genomic Selection: Lessons Learned and Perspectives. Front. Plant Sci. 13:890434. doi: 10.3389/fpls.2022.890434

breeding and highlight "that genomic evaluation performance could be comparable to the already well-optimized pedigree-based evaluation under certain conditions [...] Genome-based methods showed advantages over pedigree counterparts when ranking candidates at the within-family levels, for most of the families."

The other eight original research contributions were related to wheat, maize and rice.

Bonnett et al. addressed the application of GS in a wheat breeding pipeline. In particular, the authors considered the performance of selected material when applying genomic selection with different prediction methods in an early generation.

The topic of modeling environmental effects and genotype-by-environment interactions (GEI) was addressed by several authors. Westhues et al. included environmental predictors in GS using gradient boosting. Based on "data collected by the Maize Genomes to Fields" initiative, the authors found that "Accuracy in forecasting grain yield performance of new genotypes in a new year was improved by up to 20% over the baseline model by including environmental predictors with gradient boosting methods." Genotype-by-environment interactions were also considered by Tomar et al. who investigated the predictive ability of a multi-environment genomic prediction model for yield in spring wheat. Atanda et al. and He et al. considered the modeling of GEI with the focus on applications in sparse-testing, and Rembe et al. investigated the impact of GEI on reciprocal recurrent genomic selection.

Ma and Cao addressed the dissection of grain yield of maize and compared the predictive ability of different approaches, in particular when incorporating markers associated with the traits of interest as a fixed effect in the statistical model. Finally, Cao et al. addressed genomic prediction of resistance to Tar Spot.

As a contribution of a method article, Schrauf et al. discussed how to compare different genomic prediction models by cross validations. The authors "emphasize the importance of paired comparisons to achieve high power in the comparison between candidate models, as well as the need to define notions of relevance in the difference between their performances. Regarding the latter," the authors "borrow the idea of equivalence margins from clinical research and introduce new statistical tests."

As review contributions, Fritsche-Neto et al. reviewed GS in small scale maize hybrid programs and Simeão et al. described the current status and future application of GS in tropical forage grasses.

Concerning opinion articles, Crossa et al. presented their view on the "Modern Plant Breeding Triangle," comprising genomics, phenomics, and environomics. Martini et al. highlighted the challenges that prediction approaches face when aiming at harnessing genetic resources, that is predicting diverse material which may not be sufficiently represented in the training set. Covarrubias-Pazaran et al. outlined how public breeding programs could be strengthened by focusing on quantitative genetics principles, and by sharing data resources including genomic data and breeding values predicted from experimental evaluations from different organizations. Another opinion contribution was provided by Gholami et al. who compared the adoption of GS across different breeding institutions, in more detail dairy cattle breeding and public and private plant breeding programs. The authors highlight that differences in the organizational structure of plant and animal breeding institutions, as well as differences in the cost-benefit structures of the use of GS in private and public plant breeding may have been the cause for differences in the adoption of GS. Gianola contributed with his reflections on trends and developments in statistical genetics addressing for instance the "deconstruction of genetic architecture" and highlighting that "quantitative genetics provides just a linear (local) approximation to complexity with little (if any) mechanistic value." Moreover, the author emphasized the principal of parsimony in genetic models and that a bias of a statistical method does not need to be a problem but that "practically all machine learning methods (e.g., random forests) provide biased predictions that, on average, will be better than unbiased machines."

In the direction of what Gianola called the "linear (local) approximation," Powell et al. argue that "The implicit capture of non-stationary effects of alleles requires the G2P map to be re-estimated across different contexts" and discuss the "development and application of hierarchical G2P maps that explicitly capture non-stationary effects of alleles."

The rough outline of the content of our Research Topic emphasizes that GS is now well-established across many plant species. Moreover, five out of 12 research articles were related to GEI indicating the relevance of this topic in current research. Plant breeding programs may have more need to estimate GEI because a program's purpose is to develop improved varieties which is inherently tied to the target environments. Was our Research Topic able to answer all the questions originally formulated? We do not think so. For instance, additional contributions on the optimal use of GS for different crops, but also a more detailed retrospective analysis of realized selection gain after the introduction of GS, or the relevance of epistasis models, hybrid prediction and new machine learning models would have been desirable.

We hope that our Research Topic supports readers with the priorization of their own ideas for future investigation, and we look forward to a potential second volume, maybe 25 years after the milestone paper by Meuwissen et al. (2001).

AUTHOR CONTRIBUTIONS

JM drafted the Editorial. SH, BG, VW, and FT checked the manuscript and suggested modifications. All authors contributed to the article and approved the submitted version.

REFERENCES

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

Conflict of Interest: VW was employed by KWS SAAT SE & Co., KGaA, Einbeck, Germany. BG was employed by Bayer Crop Science, St. Louis, MO, United States.

The remaining 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.

ORIGINAL RESEARCH
published: 28 October 2020
doi: 10.3389/fpls.2020.581954

Favorable Conditions for Genomic Evaluation to Outperform Classical Pedigree Evaluation Highlighted by a Proof-of-Concept Study in Poplar

Marie Pégard[1], *Vincent Segura*[1,2], *Facundo Muñoz*[1], *Catherine Bastien*[1], *Véronique Jorge*[1] *and Leopoldo Sanchez*[1]*

[1] BioForA, INRA, ONF, Orléans, France, [2] AGAP, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France

Edited by:
Fernando H. Toledo,
International Maize and Wheat
Improvement Center, Mexico

Reviewed by:
Evandro Novaes,
Universidade Federal de Lavras, Brazil
Jaroslav Klápšte,
New Zealand Forest Research
Institute Limited (Scion), New Zealand

***Correspondence:**
Leopoldo Sanchez
leopoldo.sanchez-rodriguez@inrae.fr

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *10 July 2020*
Accepted: *22 September 2020*
Published: *28 October 2020*

Citation:
Pégard M, Segura V, Muñoz F, Bastien C, Jorge V and Sanchez L (2020) Favorable Conditions for Genomic Evaluation to Outperform Classical Pedigree Evaluation Highlighted by a Proof-of-Concept Study in Poplar.
Front. Plant Sci. 11:581954.
doi: 10.3389/fpls.2020.581954

Forest trees like poplar are particular in many ways compared to other domesticated species. They have long juvenile phases, ongoing crop-wild gene flow, extensive outcrossing, and slow growth. All these particularities tend to make the conduction of breeding programs and evaluation stages costly both in time and resources. Perennials like trees are therefore good candidates for the implementation of genomic selection (GS) which is a good way to accelerate the breeding process, by unchaining selection from phenotypic evaluation without affecting precision. In this study, we tried to compare GS to pedigree-based traditional evaluation, and evaluated under which conditions genomic evaluation outperforms classical pedigree evaluation. Several conditions were evaluated as the constitution of the training population by cross-validation, the implementation of multi-trait, single trait, additive and non-additive models with different estimation methods (G-BLUP or weighted G-BLUP). Finally, the impact of the marker densification was tested through four marker density sets. The population under study corresponds to a pedigree of 24 parents and 1,011 offspring, structured into 35 full-sib families. Four evaluation batches were planted in the same location and seven traits were evaluated on 1 and 2 years old trees. The quality of prediction was reported by the accuracy, the Spearman rank correlation and prediction bias and tested with a cross-validation and an independent individual test set. Our results show that genomic evaluation performance could be comparable to the already well-optimized pedigree-based evaluation under certain conditions. Genomic evaluation appeared to be advantageous when using an independent test set and a set of less precise phenotypes. Genome-based methods showed advantages over pedigree counterparts when ranking candidates at the within-family levels, for most of the families. Our study also showed that looking at ranking criteria as Spearman rank correlation can reveal benefits to genomic selection hidden by biased predictions.

Keywords: black poplar, genomic evaluation, marker density, degraded phenotypes, non-additive effects, multi-trait, intra-family selection, breeding scheme

1. BACKGROUND

Forest tree species of interest for domestication like poplar are particular in many ways compared to other domesticated species, notably when it comes to breeding. Among the various particularities, forest trees have long juvenile phases, ongoing crop-wild gene flow, and extensive outcrossing (Miller and Gross, 2011). All of these hamper the process of "*controlled*" recombination by the breeder. Slow growth and cumbersomeness typical of trees do not facilitate either the conduction of breeding programs, notably with evaluation stages being costly both in time and resources. One of the poplar's particularities is clonality or the possibility of asexual reproduction, which is a powerful tool in evaluation and operational breeding (Bisognin, 2011). However, benefits rarely go hand in hand with simplicity. Typically for developing a new poplar variety, a first year is used for mating and seedling growth in nurseries. A second year is used to propagate the cuttings and install the experiments using a statistical design to do evaluations in different environments, and many subsequent years pass before we can assess genotype-by-environment (G × E) interactions, or late maturation traits like wood quality. Selection in poplars proceeds typically via independent level stages (independent culling levels), with early stages involving screening for fast-growing, disease-resistant individuals from large numbers of candidates. Late stages focus on a reduced remainder to select on final growth, architecture, disease resistance, and wood properties. This has been so far operationally efficient considering the constraints imposed by the particularities of trees, but it remains time consuming and lacks precision at the early stages.

For previous and additional reasons, perennials like trees are good candidates for the implementation of genomic selection (GS) (Muranty et al., 2014). GS can potentially accelerate the breeding process, by unchaining selection from phenotypic evaluation without affecting precision (Meuwissen et al., 2001). When applied early at the seedling stage, GS could potentially save evaluation resources and reduce the time required for evaluation of late maturation traits. GS involves ranking and selecting individuals by using a genome-wide marker set and prediction models calibrated previously in a training set. GS has been made possible thanks to easy access to cheap genotyping data, and to recent developments in evaluation methodology (de los Campos et al., 2009). Recent studies of GS in forest trees were conducted on several species: eucalypts (Resende et al., 2012b; Müller et al., 2017; Tan et al., 2017, 2018; Cappa et al., 2019; Ballesta et al., 2020), spruce (Gamal El-Dien et al., 2015, 2016; Ratcliffe et al., 2015; Lenz et al., 2017, 2020; Chen et al., 2018; Chamberland et al., 2020), pines (Resende et al., 2012a; de Almeida Filho et al., 2016; Ratcliffe et al., 2017; Gianola and Fernando, 2020; Ukrainetz and Mansfield, 2020), and rubber trees (Cros et al., 2019; Souza et al., 2019). Given the differences among forest species in general, and between their breeding programs in particular, assessments of GS feasibility at a case-by-case basis are often desirable.

According to Hayes et al. (2009), several parameters are involved in genomic evaluation accuracy. First, the extent of linkage disequilibrium in the population, which is linked to the effective population size, affects the accuracy of genomic prediction. Linkage facilitates the use of markers as *proxies* of unknown QTLs in estimating genetic effects. The required marker density is directly dictated by the extent of linkage disequilibrium: the lower the linkage disequilibrium, the higher the number of required markers (Grattapaglia and Resende, 2011; Wientjes et al., 2013). The second parameter of importance for accuracy is the composition of the training set. Such a set must be representative of the candidates for which a prediction is required. Several studies developed methods to optimize the composition of the training set (Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015). The third parameter is trait genetic architecture, usually unknown or poorly understood, but that has an influence on the performances of the different evaluation methods (Wimmer et al., 2013). Some evaluation methods, such as those using some efficient strategy to focus only on relevant variables like the family of bayesian methods, appear to be more efficient with traits with fairly uneven distributions of gene effects. Other methods with less stringent *a priori* on the distribution of gene effects work generally well with highly polygenic traits, like G-BLUP. Other modeling approaches intent to capture the underlying complexity of genetic architectures, by including non-additive effects like dominance and epistatic interactions (Toro and Varona, 2010; Su et al., 2012; Vitezica et al., 2013, 2017; Muñoz et al., 2014; Martini et al., 2017), and by considering multiple correlated traits. The latter have not been often used, despite some promising simulation studies (Calus and Veerkamp, 2011; Guo et al., 2014), empirical studies (Jia and Jannink, 2012), and the known fact from classical evaluation that genetic correlations can back accuracies of poorly heritable traits or those harboring many missing values in the dataset (Gilmour et al., 2009).

In the present work, we intended to benefit from the large corpus of knowledge already established around the concept of GS to carry out a proof-of-concept study on the feasibility of the methodology in the context of the black poplar breeding program in France. Black poplar is the leading Eurasian species of riparian forest, with a wide distribution area, and contributing as a parent together with *Populus deltoides* to one of the most widely used hybrid (*Populus* × *canadensis*) tree in the wood industry. This study is the first GS study for a *Populus* species. One of the main objectives of the study was to compare GS to pedigree-based traditional evaluation, by assessing different modeling options including non-additive genetic effects and multiple-trait evaluation. The study also considered the role of marker densification in the performance of GS, by benefiting from a recent imputation study (Pegard et al., 2018). The potential benefits of shortening the breeding cycle, although of importance, were not evaluated but only discussed in present work, because of the relatively late sexual maturity in the species. Finally, the design of the calibration and validation sets was taken into account as an additional factor in the comparison. Globally, the study intended to identify the situations in which GS could be a feasible option for poplar, and also the assessments required to reveal any eventual advantage.

TABLE 1 | Description of the pedigree and distribution of family sizes after correction of the pedigree from the marker information.

Father	SAN-GIORGIO												
Mother	**SRZ**	**BDG**	**71077-308**	**92510-1**	**72145-007**	**72131-017**	**73182-009**	**73193-056**	**72131-036**	**3824-3**	**71034-2-406**	**72146-11**	**Total**
VGN	55	57	54	32		34	14	30					276
71041-3-402		28	11	17	30		25						111
71072-501	25	28	29										82
SSC	15	20	20	22									77
71040					24	20							44
662200037					25	32	118	31					206
73193-089						22	20	18					60
662200216							31		19				50
71069-914										22			22
73193-091										21		30	51
H480										13	19		32
Total	95	133	114	71	79	108	208	79	19	56	19	30	1,011

Colors correspond to experimental field trials: green for the trial conducted in 2000/2001, pink for the trial conducted in 2012/2013, blue for the trial conducted in 2014/2016, and orange for the trial conducted in 2017/2018. These latter individuals represent two parental females (underlined codes) are progenies of VGN and BDG and were subsequently used as parent for the multiple pair mating.

2. MATERIALS AND METHODS

2.1. Plant Material

The population under study corresponds to a pedigree of 24 parents and 1,011 offspring, structured into 35 full-sib cohorts, and involving a 4 by 4 factorial mating design together with a series of multiple pair-mating designs (Pegard et al., 2018). Most of the parents were sampled from natural populations or were high-performance trees already used in the breeding program. The population corresponds therefore to the offsprings of these individuals obtained by controlled crosses. The effective population size was estimated to be 12 from coancestry matrices (Caballero, 2000) computed from pedigree corrected by marker information. Family size ranged from 10 to 118, with an average of 26 individuals per family. Pedigree description and distribution of family sizes are available in **Table 1**.

2.2. Phenotyping

Field evaluations corresponded to four different experimental trials. All four experimental trials were planted in the same location (47° 37’59” N, 1° 49’59” W, Guéméné-Penfao, France) with small variations in plot orientation and with common genotypes as controls across experimental trials (**Supplementary Table 1**). The first experimental trial (2000 and 2001) involved the factorial mating design with a total of 14 families and 413 offspring phenotyped. In second and third experimental trials, 126 individuals in 6 families and 105 from 5 families were phenotyped (2012/2013 and 2014/2016, respectively). Finally, in order to reinforce the connectivity between the different experimental trials, 10 additional full-sib families with some parents already in use in previous experimental trials were added in 2015 and phenotyped in 2017/2018. In total, 367 individuals were phenotyped in this last batch. At their respective time-frames, all 1,011 offspring and the 24 parents were vegetatively propagated, and field evaluated in separate experiments according to the same six randomized complete block design.

Phenotyping involved seven different measurements over different years (2000/2001, 2012/2013, 2014/2016, and 2017/2018), and for five different traits. Growth was assessed as stem circumference and tree height. Stem circumference at 1 m was considered for the second year (circ2). Height was assessed with a graduated rod after 1 (height1) and 2 years of growth (height2). Mean branching angle was scored on proleptic branches at the age of 2 years with a 1–4 scoring scale (angbranch; score 1 : 0–30° from the horizontal; score 2: 30–40°; score 3: 40–55°; score 4: >55°). The scale for angbranch was calibrated in such a way that resulting measures in the same population of reference resulted in phenotyping distributions being close to normality. Rust resistance was assessed with a 1 (no symptom) to 9 (generalized symptoms) scale (Legionnet et al., 1999) at year 1 (rust1) and year 2 (rust2). Budburst phenology of the stem terminal bud was evaluated by measuring its kinetics (every 3 or 5 days from March to April) with a 0–5 scale, where stage 0 corresponded to a completely closed bud while stage 5 corresponded to the initiation of stem internode elongation (Castellani et al., 1967). A local polynomial regression model was fitted between stages and dates for each individual and this model was further used to predict the date in Julian days at which the terminal bud was at stage 3 and in order to assess individual susceptibility to late frosts (Howe et al., 2000). As a result of such fitting for budburst, distributions were continuous and close to normality.

All seven phenotypes were independently adjusted to field micro-environmental heterogeneity with the breedR package [Muñoz and Sanchez, 2018, implemented in R3.3.1 platform (R Core Team, 2018)]. We used an individual-tree mixed model over all four experimental trials, comprising all available information with genotyped and non-genotyped individuals (not included in this study) according to a single-step formulation (mixture of pedigree relationship matrix for non-genotyped individuals and genomic equivalent for genotyped individuals) (Legarra et al., 2009). A random effect capturing spatial heterogeneity at individual level within trials was fitted thanks to the use of a bi-splines surface covering row and column axes (Cappa and Cantet, 2007; Cappa et al., 2015). Such a surface was nested within each evaluation experimental trial. Bi-splines were anchored at a given number of knots for rows and columns, with higher numbers increasing the roughness of the surfaces and lower numbers giving extra smoothness. Knot numbers were optimized by an automated grid search based on the Akaike information criterion (Akaike 1974) provided by breedR. The use of all available information in field trials, including non-genotyped individuals, minimized the occurrence of gaps in the surfaces and facilitated the prediction of accurate micro-environmental individual effects across the experiment. The fact of using common genotypes across trials (see **Supplementary Table 1**) and the use of genomic and pedigree relatedness in the mixed model facilitated the adjustment across trials. The micro-environmental individual effect was subtracted from the observed phenotype to obtain a spatially adjusted individual phenotype. A clonal mean of spatially adjusted phenotypes was calculated for each trait and used as raw phenotypes for the rest of the study (hereafter adjusted clonal mean). As a default option, data from all blocks (6 Blocs) were used as input to the model. The same model was fitted to data from only three of the blocks (blocks 1, 3, and 5 called 3 Blocs model afterwards), to assess prediction quality with a less precise phenotype. All measurements were tested for deviations from normality by a randomized Q-Q plot.

2.3. Genotyping

All 1,033 individuals in this population (22 parents, 2 being both parents and offspring, and 1,009 offspring) were genotyped using the Populus nigra 12K custom Infinium Bead-Chip (Illumina, San Diego, CA) (Faivre-Rampant et al., 2016). Additionally, 43 individuals were sequenced, including an extra founder that was identified as one of the grandparents in the pedigree. Among the remaining 42 sequenced, there were 22 parents, 14 progenies, and six unrelated individuals from natural populations. Progenies were chosen in such a way that all parents had at least one offspring with its genome sequenced. The set of unrelated individuals were used to assess the imputation ability under challenging conditions. In a previous study (Pegard et al., 2018) genotype imputation from 7K (effective SNPs out of 12K in array) to 1,466,586 SNPs was performed attaining imputation qualities higher than 0.84 per individual, and evaluated by a leave-one-out cross-validation scheme (CV). Resulting imputation was used in the present study to constitute alternative sets of selected markers for genotyping. For quality assessment and selection of the marker sets, we used the proportion of alleles correctly imputed by genomic position across individuals (Props), and Props corrected by the probability of correct imputation by chance (Badke et al., 2013; cProps). Among the imputed SNPs, we selected those with Props higher than 0.90, with cProps higher than 0.60 and a minor allele frequency (MAF) higher than 0.05, to obtain a set of 249,805 SNPs (250K). That latter set comprised the total of the 7K from the chip. We selected two alternative smaller marker sets: 50K (with 50,565 SNPs), and 7K_homo (with 7,048 SNPs) where coverage and homogeneity of density was optimized over the original 7K array. These two sets were composed by selecting, respectively 1 SNPs every 1,000 or 50,000 bp out of the 250K set. Whenever more than one candidate SNPs were available for the same window, we selected the one that had the highest values of Props and cProps.

2.4. Models

We estimated variance components and heritabilities with the complete data set and single trait models, and genetic correlations with a genomic multiple-trait model (GBLUP). The Akaike Information Criterion (AIC) was used to assess for each given trait the quality of each model. Two alternative methods were used to calculate genomic estimated breeding values for each trait: the best linear unbiased prediction based on genomic information (GBLUP) (Whittaker et al., 2000; Meuwissen et al., 2001), and the weighted GBLUP (wGBLUP; Legarra et al., 2009; Zhang et al., 2016). They were all compared to the best linear unbiased prediction based on pedigree information (PBLUP) (Henderson, 1975). The models for GBLUP (and PBLUP) using matrix notation for additive and non-additive effects were given by:

$$y = B\beta + Zu + \varepsilon \tag{1}$$

$$y = B\beta + Zu + Wd + \varepsilon \tag{2}$$

where y was the adjusted clonal mean, β a vector of fixed effects, u the vector of random additive effects following $N(0,G\sigma_a^2)$ with σ_a^2 the additive variance and G (or A in PBLUP) the relationship matrix, d was the vector of random dominance effects following $N(0,D\sigma_d^2)$ with σ_d^2 the dominance variance and D the dominance relationship matrix, ε the vector of residual effects following $N(0,I\sigma_e^2)$ with σ_e^2 the residual variance. The design matrix B contains the values of the covariables with fixed effects and Z, W, and I are indicator matrices relating the clonal mean to the random effects. The methods used to obtain the relationship matrices are explained in the next section. The PBLUP and GBLUP single-trait models as well as the multi-trait models were fitted with the R package breedR (Muñoz and Sanchez, 2018). All the analyses are summarized in **Table 2**.

2.5. Relationship Matrix Estimation

The ARM (additive relationship matrix) was built from the known pedigree at the moment of the controlled crossings, and denoted hereafter as A. However, a preliminary marker assessment in this study showed that there were errors in the pedigree. Pedigree was corrected based on these results and a new reconstructed ARM was obtained, denoted hereafter as A_{cor}. Pedigree errors involved in most cases a wrong paternity attribution and, less frequently, individuals supposed to be different genetically. The total number of parents after correction did not change, with an added father and a removed mother.

TABLE 2 | Combination of models and marker sets tested.

Methods	ADD	ADD + DOM	MultiTrait	SNP set
P-BLUP	Yes	Yes	Yes	None
P-BLUPcor	Yes	Yes	Yes	None
GBLUP	Yes	Yes	Yes	7K
	Yes	Yes	No	50K
	Yes	Yes	No	100K
	Yes	Yes	No	250K
wGBLUP	Yes	Yes	No	7K
	Yes	Yes	No	50K
	Yes	Yes	No	100K
	Yes	Yes	No	250K
BayesCpi	Yes	No	No	7K
	Yes	No	No	50K
	Yes	No	No	100K
	No	No	No	250K

The main change concerned the number of families that went from 39 to 35. Both A and A_{cor} were calculated with the R package nadiv (Wolak, 2012), and kept for the comparison in order to show the potential loss due to pedigree errors and the maximum performance attainable by pedigree. Concerning the genomic relationship, we used a normalized matrix (G Equation 3) calculated following VanRaden's formulation (Habier et al., 2007; VanRaden, 2007) and the scaling proposed by Forni et al. (2011) to assure compatibility with A, for each genotyping set (7K, 50K, 100K, and 250K) :

$$G = \frac{(M - P_1)W_a(M - P_1)'}{trace[(M - P_1)W_a(M - P_1)']/n} \tag{3}$$

where M was a genotyping matrix with m markers in columns and n individuals rows, P_1 was a matrix $(n \times p)$ containing the minor allele frequency $(2p_i)$, at the marker i, and W_a was a matrix of weights described below. *Ad hoc* scripts in R were used to make the computations for G (R3.3.1 platform). To assess dominance effects, a dominance matrix based on the pedigree information was calculated with the R package nadiv (Wolak, 2012) with expected and observed pedigree information (D and D_{cor}). The genomic dominance matrix was calculated as:

$$D = \frac{(X - P_2)W_d(X - P_2)'}{trace[(X - P_2)W_d(X - P_2)']/n} \tag{4}$$

where X was the genotyping $(n \times p)$ matrix containing code "0" for the homozygous and "1" for the heterozygous, P_2 the $(n \times p)$ matrix containing the heterozygous frequency $(2p_iq_i)$ according to Vitezica et al. (2013) and normalized in the same way as for G in Equation 3, and W_d the matrix of weights as described below. We used one of the procedures of Wang et al. (2012) for calculating weights in wGBLUP. Unlike GBLUP, where all markers have the same variance and therefore the same weight, the derivative wGBLUP uses a transformed G according to marker weights to select markers. The weights were calculated as $w_j = \hat{u}_j^2$ where w_j was the weight for the SNP j and $\hat{u}_j$ was the estimated marker effect obtained as

$$\hat{u}_a = W_aX'G^{-1}\hat{g} \tag{5}$$

$$\hat{u}_d = W_dX'D^{-1}\hat{d}, \tag{6}$$

where $W_{a,d}$ was a diagonal of weights, either a identity matrix (GBLUP) or a diagonal of w weights (wGBLUP) for additive (W_a) or dominance (W_d) relationship matrices, $\hat{g}$ the genomic estimated breeding values (GEBV) and $\hat{d}$ the estimated dominance effects. Several iterations of recomputed $\hat{u}_a$, $\hat{u}_d$, $\hat{g}$, and $\hat{d}$ were performed to update G, following recommendation by Wang et al. (2012), and according to the following steps:

1. Define $i = 1$, $W_{(a,d)i} = I$ and G_i as Equation (3)
2. Compute $\hat{g}_i$ using GBLUP approach
3. Compute additive SNP effects with Equation (5) and dominance SNP effects with Equation (6)
4. Calculate SNP weights as $w_{aj+1} = \hat{u}_{ai}^2$ and $w_{dj+1} = \hat{u}_{di}^2$
5. Scale w_{aj+1} and w_{dj+1}
6. Calculate G_{i+1} with Equation (3)
7. Calculate D_{i+1} with Equation (4)
8. $i = i + 1$
9. Iterate from 2 until $i = 3$.

A weighted relationship matrix was obtained from each subsequent iteration, giving respectively, Gw1, Gw2, and Gw3, as three distinct matrices leading to separate evaluation methods. In this study, therefore, eight relationship matrices were tested (A, Acor, G, Gw1, Gw2, Gw3, D, Dcor), and the resulting predictions were compared via cross-validation and by an independent data set.

2.6. Prediction Accuracy and Cross-Validation

We assessed the impact of the composition of the training (TS) and validation sets (VS) on the performance of the genomic evaluation by trying two TS/VS sizes and two different TS/VS compositions in a 10-fold cross-validation scheme. The two sizes were 50% (T50) and 25% (T25) of the individuals evaluated in the 2000/2001, 2012/2013, and 2014/2016 experimental trials. The last field evaluation trial of 2017/2018 did not contribute to TS and was used as an extra independent validation set (TestSet) for each of the four TS, as it represented a sample of the next generation of selection candidates. Such Testset represents an independent validation experiment without the risk of eventual overfitting that is typical of cross-validation schemes, and it is the result of a mating campaign involving a sample of parents from the breeding population (see **Table 1**). The two composition scenarios for TS and VS involved: a sampling of individuals independently of their family membership and a sampling of different family sets. Both size and composition were combined to obtain the desired percentage (50 or 25%) of individuals or the desired percentage (50 or 25%) of families. The performance of the models was evaluated following different criteria. Firstly, predictive ability, which was defined as the Pearson correlation coefficient between the adjusted clonal means and the GEBVs of the samples in the VS, or in the TestSet. The accuracy (Accuracy and Accuracy test) of the models were estimated by dividing each predictive ability by the square root of the heritability of the corresponding A model for the given trait. Additionally, the Spearman rank correlation between the adjusted clonal means and the GEBVs of the individuals in the VS was calculated (Spearman). We estimated the Spearman and Pearson correlation of the top 5% of the trait, for the section between 5 and 10%, and between 10 and 50% within the VS. Finally, we assessed potential bias in genomic predictions by estimating the intercept and the slope of the linear regression between the adjusted clonal means and the GEBVs of each model, in the VS and in the TestSet. Predictive abilities were also calculated at the within family level. The prediction ability obtained within families following PBLUP with the corrected pedigree were subtracted to the equivalent prediction ability obtained from the genomic model, for given cross-validation scenario and trait. A weighted average was then calculated according to the size of the family. given cross-validation scenario and trait. A weighted average was then calculated according to the size of the family.

2.7. Testing Factor Importance

In order to assess the main factors accounting for genomic evaluation performance, we applied the Random Forest algorithm (Liaw and Wiener, 2002) implemented in the Boruta R package (Kursa and Rudnicki, 2010). The main factors (or features) were: Trait, Matrix (A, Acor, G, Gw1, Gw2, Gw3, D, Dcor, Dw1, Dw2, Dw3), GeneticEffect (Additive, Additive, and Dominance), ST_MT (Single-Trait, Multiple-Trait), GenoSet (none, 7K,7K_homo, 50K, 250K), Type (Individual, Family), Perc (T50, T25), and PhenoSet (6 Blocs, 3 Blocs). Classification of features was done for each of the performance variables available: predicting ability, Accuracy, Spearman correlation, and slope.

3. RESULTS

3.1. Heritabilities

Heritabilities with their corresponding variance components and Akaike Information Criterion (AIC) are shown for all models and traits in the **Supplementary Table 2**. In general, most traits showed intermediate to high heritabilities (average of 0.73) as illustrated by **Figure 1A**, with height and rust showing the highest average values, and budburst correspondingly the lowest. The fact that we used adjusted clonal means as phenotypes to be explained in the models induced a low residual term, which in turn raised the heritability estimates. In terms of models, G and weighted G resulted in higher heritabilities across traits (**Figure 1B**), with an advantage to the latter under additive models, and to the former under models comprising also dominance (**Figure 1C**). Most of the genomic scenarios (G and weighted G) resulted in higher heritabilities than the pedigree-based counterparts, with uncorrected pedigree resulting in the lowest heritabilities overall. Another factor increasing heritability across traits was marker density, with highest values observed with the 250K SNPs set, followed by the 50k and the 7K_homo sets, with 7K resulting in the lowest values among genomic alternatives. On the contrary, using a phenotype adjusted with less information had little effect on heritabilities.

3.2. Accuracies Estimated by Cross-Validation With Different Training Sets

Three out of seven traits (budburst, height1, and rust1) were selected to show the cross-validation accuracies in **Figure 2** (the remaining traits are shown in **Supplementary Figure 1**), assuming different relationship matrices and four different training scenarios (size and composition). Results correspond to single-trait additive models with a relationship matrix based on the 7K SNP panel. Accuracies varied between 0.17 and 1.01 across all scenarios and traits. It is important to note that, because of the choice of a particular model of reference to provide a basis heritability (pedigree-based model with the A matrix), accuracies larger than one were obtained.

Accuracies responded greatly to changes in the way the training set was constituted (percentage and composition). The fact of using different families for training than for validation had a large impact on the accuracy when compared to the alternative scenario where the splitting between training and validation occurred mostly within families. Basically, as expected, predicting different families was less accurate than predicting different individuals within the same cohort, with losses in accuracy averaging 13%. This pattern was found for all traits, except for one training scenario for angbranch, where differences between the two compositions were also the weakest. Concerning the percentage, the effect of reducing the training set from T50 to T25 had also an impact on accuracy, although mostly when training and validation involved different families. On average, reduction in accuracy with decreasing training set size was around 4.2% for the training composition based on individuals, and vary depending on the trait (from −18 to 8%) for that based on families.

3.3. Challenging Prediction Models With New Individuals

We used a completely independent set of individuals representing the next generation of selection candidates to evaluate the different prediction models with 7K SNP and across two different training scenarios (T25 and T50). Results of accuracies from this independent set are presented for three traits in **Figure 3**.

Accuracies were substantially lower under the new more challenging testing scenario than those already shown for the cross-validation scheme for the same traits (see **Figure 2**). In general, marker-based models resulted in a less affected level of accuracy compared to the pedigree-based counterparts: the G-based and Gw1 models were the best performers, notably for rust1 and budburst. For height1, however, A and Acor models obtained comparable performances to those from genomic based models. Otherwise, the model based on uncorrected A had generally poorer accuracies than those shown by the corrected A. The behavior of the different models in terms of accuracies depended greatly on traits and, to a much lower extent, on the

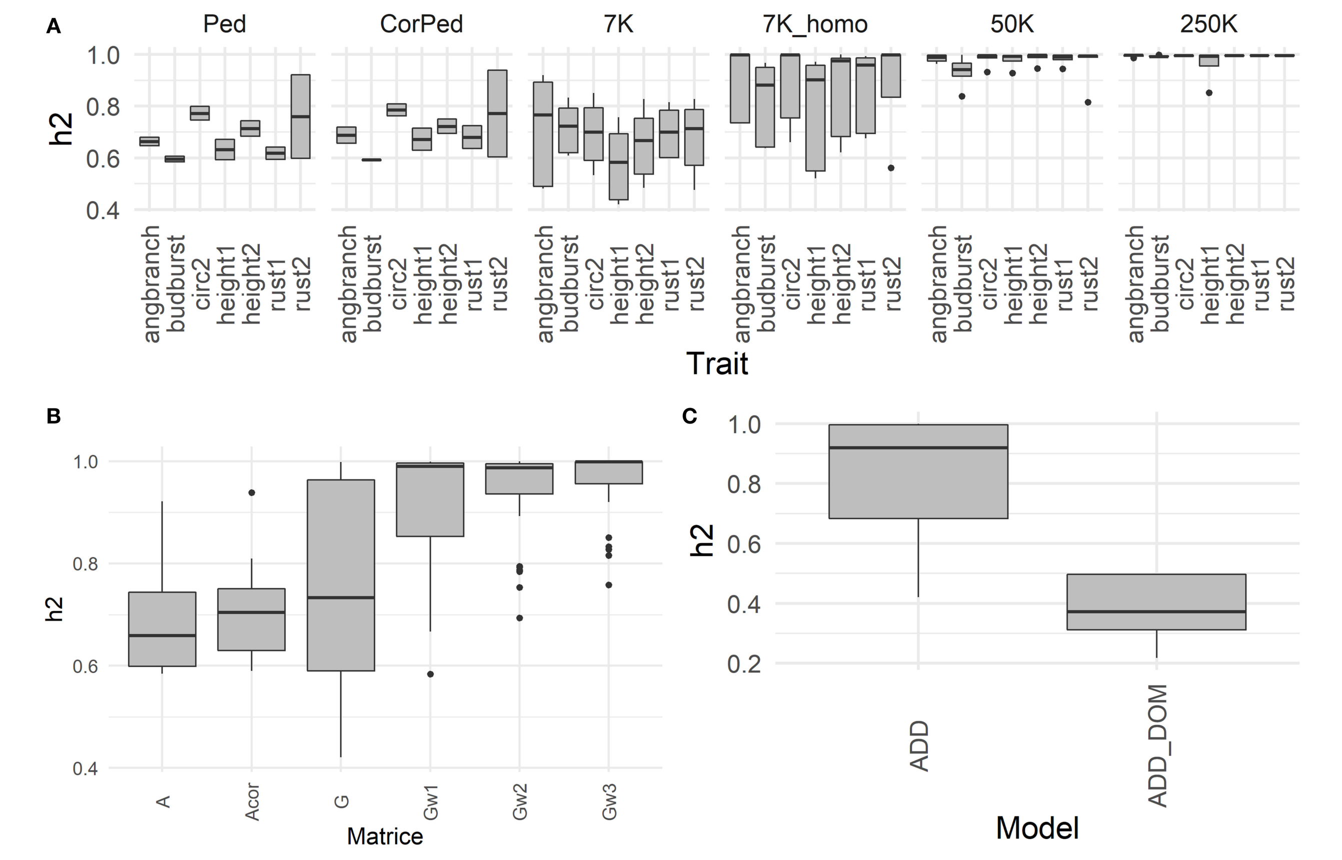


FIGURE 1 | Heritability obtained with a global model using all the available data. **(A)** Heritabilities derived from an additive model and the 6-block adjusted dataset, the boxplots represent the heritabilities per trait (angbranch, budburst, circ2, height1, height2, rust1, rust2), according to the marker density (Ped, CorPed, 7K,7K_homo, 50K, and 250K), across matrix. **(B)** Heritabilities derived from an additive model and the 6-block adjusted dataset, the boxplots represent the heritabilities per matrix (A, Acor, G, Gw1, Gw2, Gw3), across traits and marker density. **(C)** Heritabilities derived from the 6-block adjusted dataset, the boxplots represent the heritabilities per model (ADD: additive; ADD_DOM), across traits, marker density, and matrix.

training scenario. Concerning the training scenario, it is to be noted that family sampling obtained slightly higher accuracies than individual sampling, although differences were not of significance. These results give an idea of the performance obtained in a real candidate selection test. In doing so, we decided to look at the impact of other factors on the independent dataset rather than on cross-validation. The results obtained for the cross-validation are in **Supplementary Material**.

3.4. Prediction Performance in the Test Set With More Complex Models

By adding a dominance effect to the single trait model for each trait with the 7K SNP panel, we did not observe significant changes in accuracy with respect to the purely additive model (**Figure 4**, upper part, and **Supplementary Figure 2**). Overall, dominance did not lead to losses in accuracy, with similar performances to that of additive counterparts across traits.

Another added complexity were the multi-trait additive models, which were also evaluated in terms of accuracies (**Figure 4** lower part, and **Supplementary Figure 3**). The advantages of a multiple-trait approach over the single-trait counterpart were trait-dependent and generally very small. For instance, rust1 showed clearly no benefit in using a multiple-trait prediction, while for height1 the multiple-trait prediction had a small advantage when training over different families. For budburst, however, the multiple-trait approach brought a loss with the G-based model in both training scenario. Moreover, the multiple-trait approach did not seem to benefit from the use of marker-based G matrices over pedigrees. Therefore, the multiple-trait prediction did not bring a clear-cut advantage across traits and training scenarios. Genetic correlations between the traits involved in the multiple-trait analysis are shown in **Supplementary Figure 4** as supplementary data.

In summary for the TestSet, the accuracy of unweighted G-based models appeared to be slightly better than with pedigree-based models, although in most cases the Acor model obtained comparable levels of performance to the best G-based method (data not shown). The cross-validation sampling strategy (individual/family) impacted the accuracy in all cases and for all traits, with individual scenarios having, in general, higher accuracy than family scenarios. The percentage of individuals

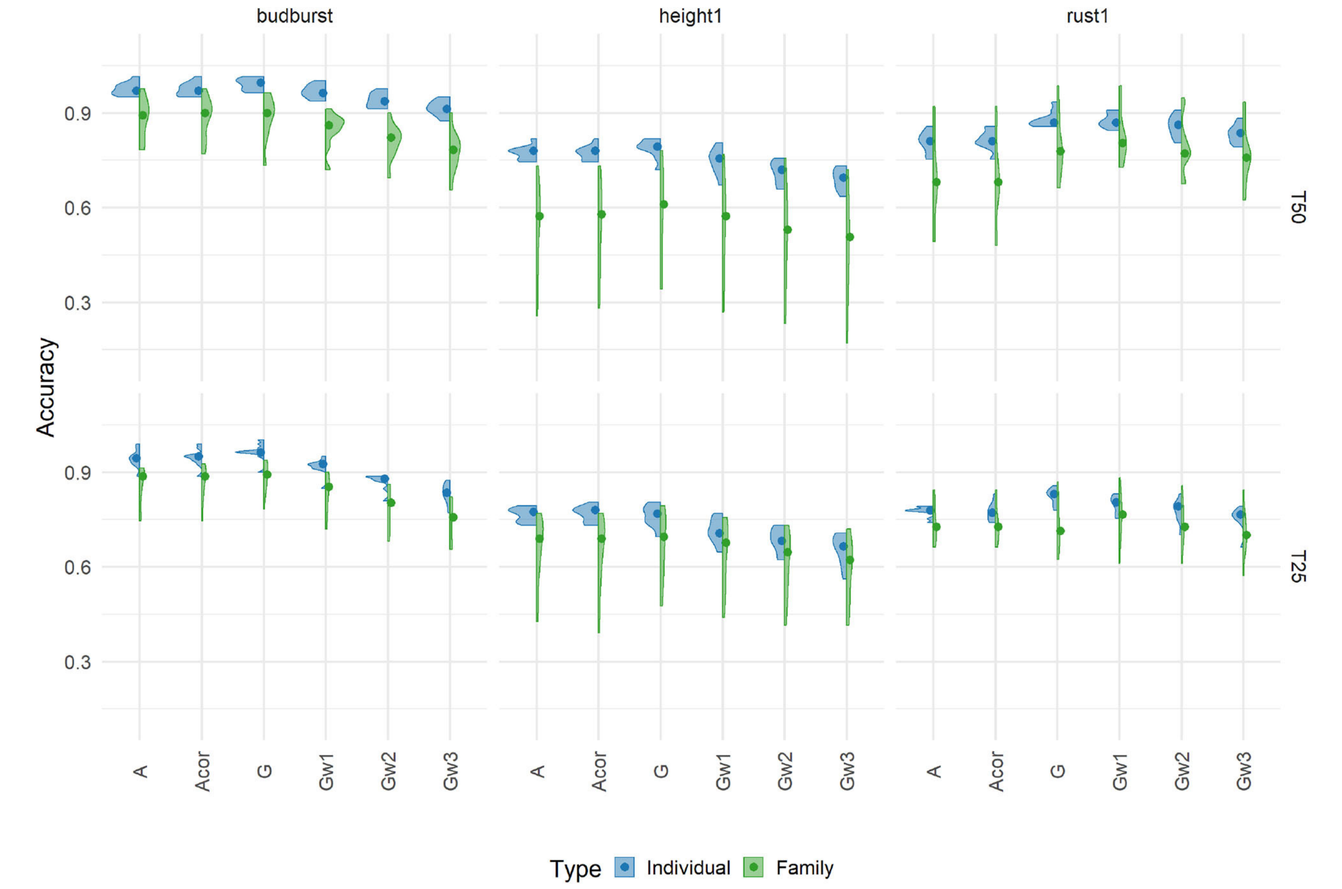


FIGURE 2 | Cross-validation prediction accuracies using an additive model with 7K SNP for three traits, grouped by the proportion of individuals (Individual, in blue) or families (Family, in green), in training sets 50% (T50) and 25% (T25). Each violin plot represented the accuracy of 10 repetitions for each scenario, and the dot represented the median of each distribution.

in the training population (T50/T25) showed a less important impact on accuracy than that of composition. More advanced models involving dominance effects and multiple-traits did not improve the performance of genomic predictions.

3.5. Effect of Marker Density on Accuracies

The same three traits (budburst, height1, and rust1) were used to show the effect of an increase in marker density on prediction accuracy over different modeling approaches on the TestSet in **Figure 5** (**Supplementary Figures 5**, **6**). We compared the accuracies obtained with four marker sets of increasing density with a single-trait additive model, and T50/Individual sampling scheme.

The effects of density were clearly trait-dependent, and the choice of traits illustrated here cover well these differences in behavior. Such densities were also differently exploited according to traits by the different G matrices used in the modeling. For traits like height1 and rust1, densification in the number of SNPs had no clear benefit in terms of accuracy, and the use of weighted G matrices did not exploit the extra density to bring additional accuracy. For traits like budburst, however, densification brought some benefits in accuracy when combined with some weighting in the G matrices, notably after one step of weighting and using the highest densities of 250K.

Besides the number of markers, their distribution over the genome seemed also of relevance for accuracy. This is particularly illustrated in the comparison between the 7K and 7K_homo sets, where the latter represents an even distribution sample over the genome. Such even distribution was not beneficial for accuracies across traits compared to the original 7K set. This latter set was seemingly richer for some relevant genes, as the array design from which the 7K set results favored certain regions linked to important traits over a homogeneous distribution.

3.6. Challenging Prediction Models With Degraded Phenotypes

Phenotypes used as dependent variables in the models resulted from averaging six field replicates that were previously spatially adjusted. To test whether the number of replicates could have an effect on the difference in performance between pedigree and genomic-based evaluations, new evaluations were produced based only on 3 out of 6 replicates. The adjusted clonal means produced were compared to those obtained with 6 Blocks. The correlation between the two clonal mean sets was close but

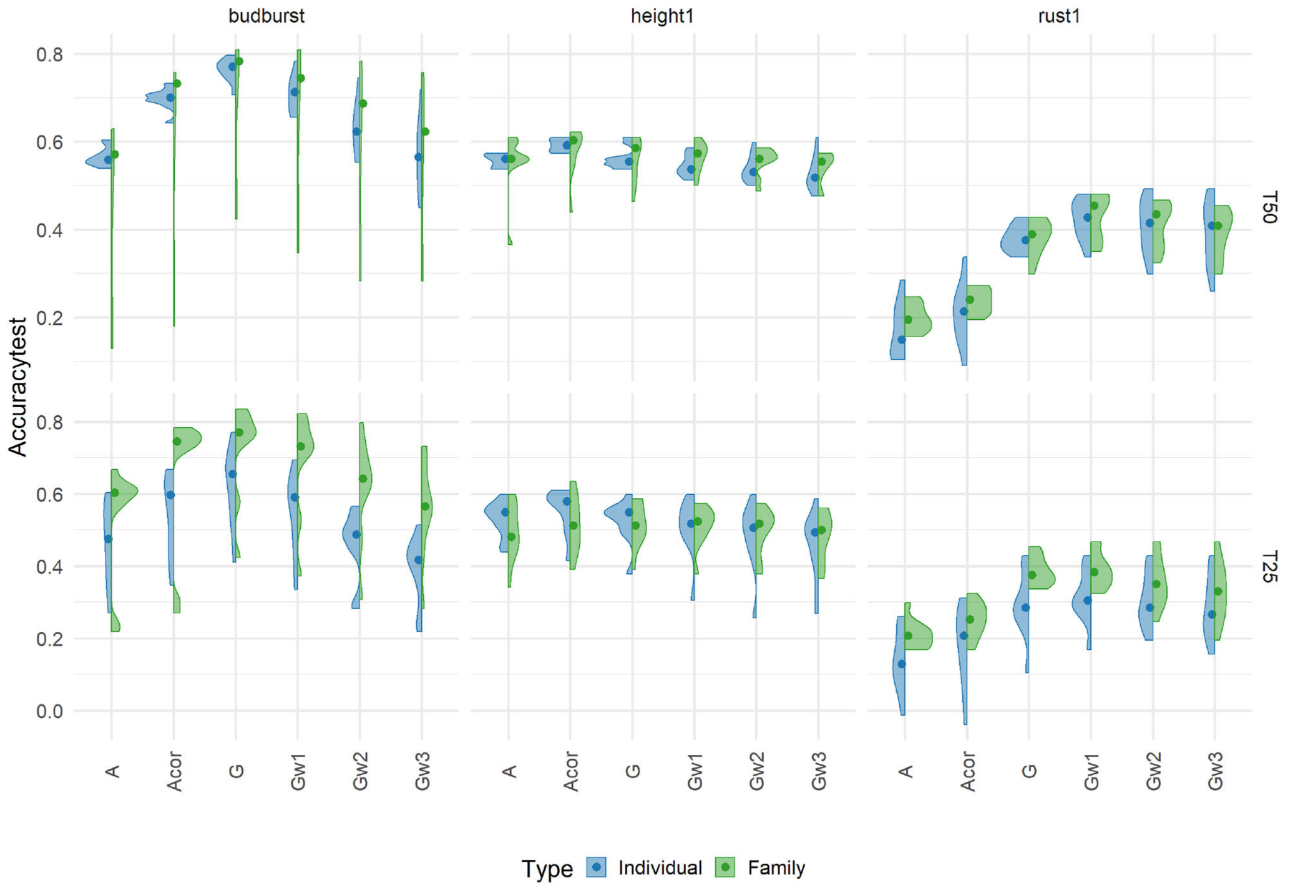


FIGURE 3 | Prediction accuracies using an additive model with 7K SNP for five traits, grouped by the proportion of individuals (Individual, in blue) or families (Family, in green), in training sets 50% (T50) and 25% (T25) on an independent Test Set representing the candidates for selection. Each violin plot represented the accuracy of ten repetitions for each scenario, and the dot represented the median of each distribution.

not equal to 1 (from 0.8 to 0.94: **Supplementary Figure 7**), and a *t*-test on paired data confirmed the difference to be of significance between the two sets of data. Resulting accuracies under this new evaluation scheme are presented in **Figure 6** (**Supplementary Figures 8**, **9** for the results in cross validation), involving the training scenario T50/individuals and the marker density set of 50K. The prediction accuracy was not significantly affected by the reduction in repetitions, across models and traits. This result was also observed for other training scenarios and for the remaining marker densities (not shown). Therefore, downgrading the phenotype with half the number of repetitions did not appear to affect pedigree-based predictions, which were almost equally competitive. This also suggests that evaluations under current conditions could have been simplified with either less field area or extended to extra candidates keeping the same field area.

3.7. Evaluation of Prediction Models With Complementary Criteria

Trends for slope of the linear regression between the adjusted clonal means used as phenotypes and the resulting GEBVs (or pedigree equivalents EBVs) across models showed that the pedigree-based approaches had the most robust behavior with values always around 1. Contrarily, G-based approaches often showed upwardly biased predictions (**Supplementary Figures 10**, **11**). This deviation was always more pronounced for G-BLUP than for weighted G-BLUP, with a decreasing trend in slope with increasing steps of weighting. Marker densities had the effect of increasing slopes, notably for G-BLUP and weighted G-BLUP schemes with fewer steps of weighting. With a less pronounced effect, the change in training scenarios from individuals to families and from T50 to T25 increased slopes. In general, G-BLUP schemes showed the largest deviation in slopes due to changes in training scenarios. Slopes larger than one correspond generally to biases in predictions that depend on the magnitude of the predicted variable, being larger the bias the larger the phenotype.

We compared two correlation coefficients: the classical Pearson correlation, on which predicting abilities are based, and a rank-based coefficient like Spearman. Such comparison was made across different tiers of the evaluated sample of candidates: from the 5% tier of best candidates to the totality of the TestSet, with the aim to explain the origin of biases. Results are shown in **Figure 7** for budburst (**Supplementary Figure 12** for the

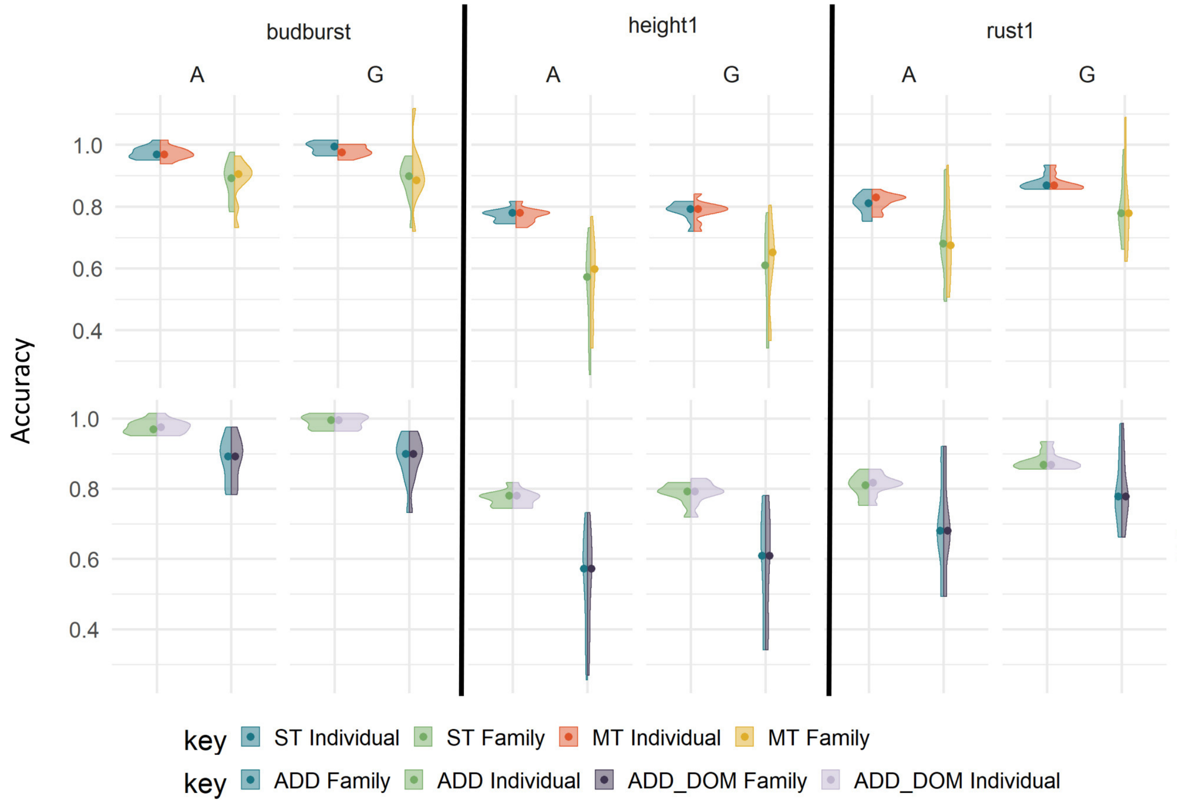


FIGURE 4 | Prediction accuracies using different evaluation models on the TestSet by cross-validation type "T50" with 7K SNP, and rust1, budburst and height1. The upper panels involve single-trait (ST) vs. multiple-trait (MT) additive models: with ST with individual sampling (blue), ST with family sampling (green), MT with individual sampling (orange), and MT with family sampling (yellow). The lower panels involve additive (ADD) vs. additive and dominance (ADD_DOM) single-trait models: with ADD and individual sampling (blue), ADD and family sampling (green), ADD_DOM and individual sampling (light purple), ADD_DOM and family sampling (dark purple).

cross-validation results). Differences between the two coefficients were substantial within the best 5% tier, where the Spearman correlation appeared to magnify the advantages of G-based models over that of pedigree-based counterparts. Such advantage became more pronounced for that particular elite tier with G-based models using higher marker densities. Differences were less pronounced for other less performing tiers, notably those closer to the mean. For the totality of the TestSet, Pearson resulted in slightly higher values than those of Spearman. Thus, the behavior of the two correlations were opposite whether we looked at the best tier or to the whole distribution, with Spearman revealing extra differences between evaluation methods for the tail of the distribution that is usually relevant for selection. Similar patterns were observed for rust1 (**Supplementary Figure 13** lower part). Height1 had a pattern slightly different, with an advantage of Spearman over Pearson for the G-based models relevant for the 50K SNP densities and for the 2 top tiers, and no advantage with the highest density 250K (**Supplementary Figure 13** upper part).

3.8. Genomic Model to Select Among Full-Sibs

Differences in Prediction ability at within-family level between genome-based and pedigree-based predictions are shown in **Figure 8**, in the shape of distributions over all available full-sib families and for three traits. Results show important variation across families, spanning from no advantage of genome-based methods with respect to the pedigree counterpart (zero differences and below), to advantages over 0.4 for the genome-based option for some of the families. The different methods of constructing the G matrix (G and weighted G) had little effect on the differences, while increasing the training set (T50 vs. T25) or sampling families instead of individuals augmented slightly the genome-based advantage in terms of median differences. These advantages were higher for budburst and rust1 than for height1. Overall, genome-based methods showed advantages over pedigree counterparts when ranking candidates at the within-family levels, for most of the families.

3.9. Ranking of Factors Impacting Prediction Accuracies

The Boruta algorithm was used to evaluate the different features explaining the variability of three performance parameters: accuracy, Spearman correlation and slope. Results in terms of Z-score for all features in the cross-validation are shown in **Figure 9**. Both correlation-based performance parameters, accuracy and Spearman, led to similar ranking of features, with Type (Individual vs. Family), trait, matrix (A and G matrices), and Perc (T50 vs. T25) being the factors explaining the most

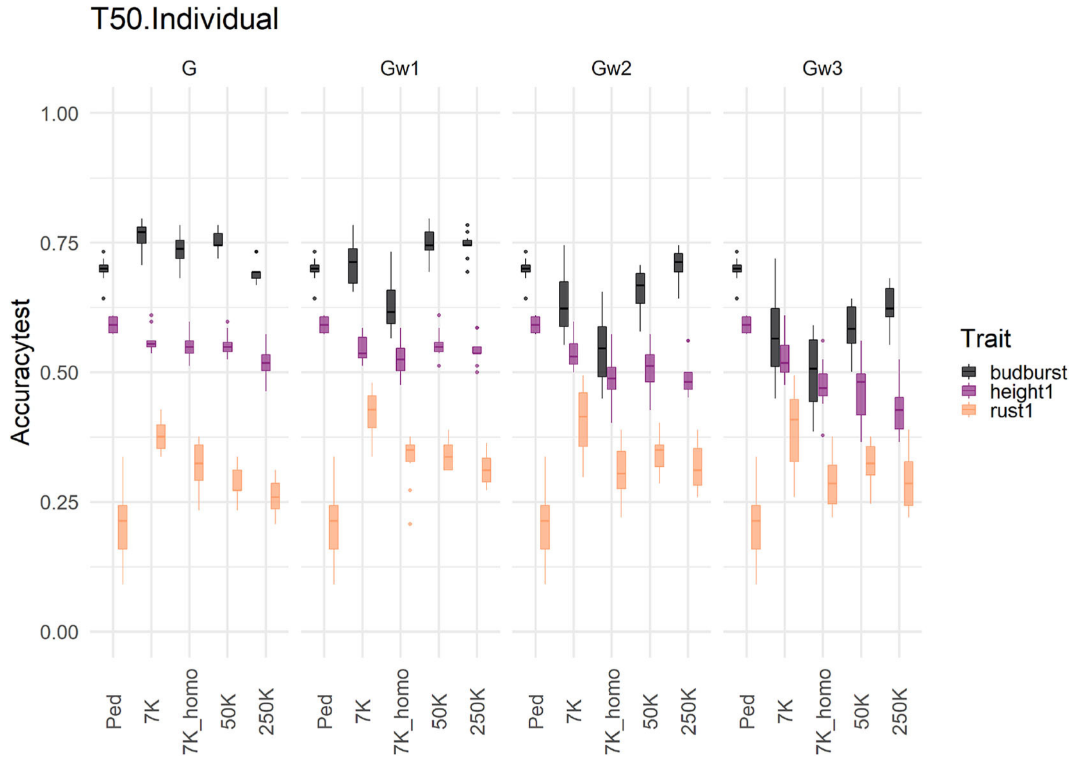


FIGURE 5 | Marker densification impact on predictive accuracy of a single trait additive model with T50 individual in the TestSet for four genomic relationships matrices (in columns) and three different traits : height1 (purple), budburst (black), and rust1 (orange). The range of accuracies obtained with the pedigree information was represented in each column by the tag Ped. The accuracies distribution is represented by a boxplot.

in performances. Thus, the A vs. G comparison, although important, was not the one at the top. For slope, however, the features related to modeling and integrating information were the most important ones, with those related to training and validation characteristics being negligible. A similar analysis was conducted on the results obtained with the TestSet (**Figure 10**). Results show patterns for accuracy and Spearman correlation similar to those of cross-validation, except for the fact that the impact of size and composition of validation was negligible in TestSet conditions. For slope, the effects of the different features were very small, again with features related to modeling and integrating information showing the most important roles. The main feature explaining variability of prediction within family is the trait variation (**Supplementary Figure 14**).

4. DISCUSSION

4.1. Genomics Does Not Improve Substantially Prediction Accuracy Over Pedigree in Standard Conditions

This study was conceived as a proof-of-concept of the genomic evaluation in the black poplar breeding program in order to evaluate feasibility and performance in a situation close to operational conditions for the species. Several main messages could be drawn from this study. Firstly, genome-based models captured higher heritabilities and higher additive variances than their pedigree equivalents, although this did not lead to a systematic advantage in terms of prediction accuracy for the former over the latter. Although G-BLUP obtained in general the best prediction accuracies, it was very closely followed by the evaluation based on a genomically corrected pedigree. Secondly, the benefit of densification of the marker panel for the prediction quality was not obvious, with results dependent on traits and treatment of the G matrix. Finally, the most clear advantages of genome-based methods and of marker densification were found in more challenging validation situations, when observing the ranking among the best 5% elite individuals or when importance was given to selection within families.

The genomic evaluation captured generally more genetic variance than pedigree evaluation, regardless of the trait. The number of markers fitted in the model generally increased the proportion of genetic variance explained by the model, but this occurred mostly under G-BLUP. When using a weighted GBLUP variant, the proportion of genetic variance explained by the model decreased with the cycles of weighting and selection of relevant markers. Without variable selection, plain G-BLUP, increasing the number of markers favored a better coverage of all genomic regions, including those close or inside relevant QTLs. Variable selection in weighted GBLUP could have eroded relevant variation, affecting the proportion of captured variation. This type of behavior could reflect an underlying infinitesimal-like

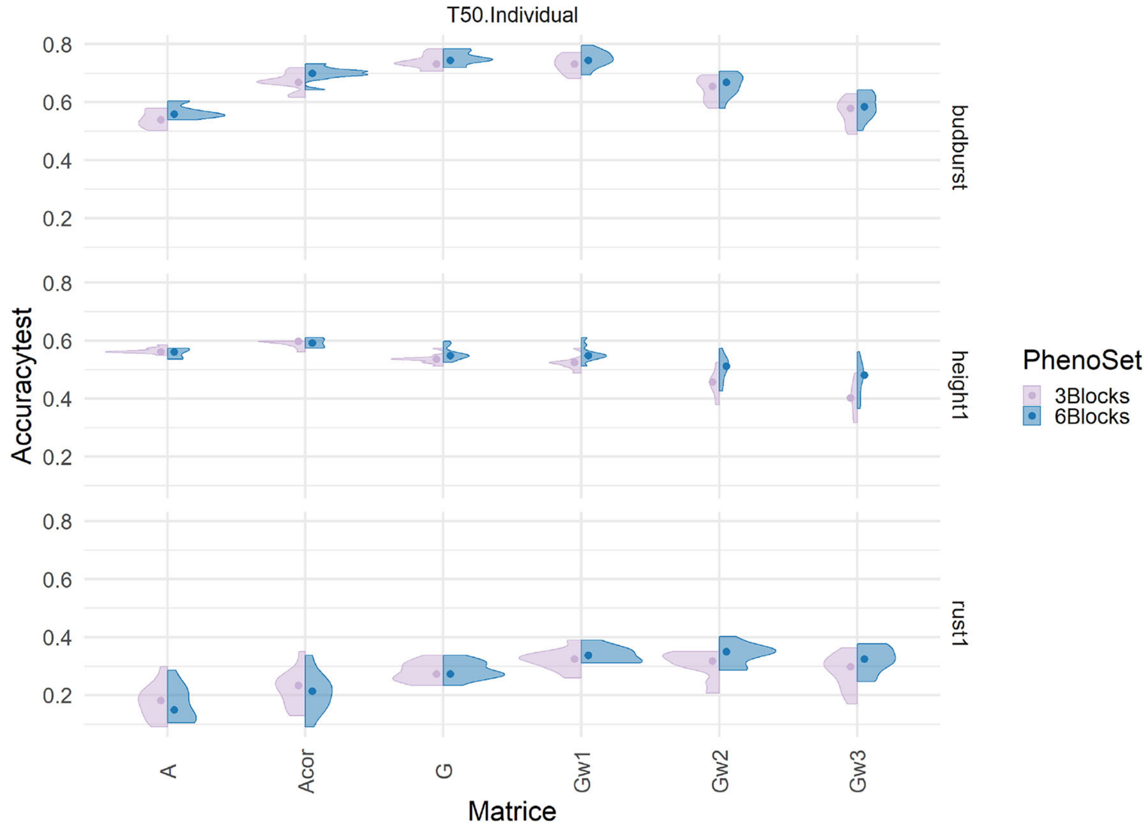


FIGURE 6 | Impact on predictive accuracy of two alternative ways of producing phenotypes, with 3 (pink) and with 6 (blue) replicates, with a single trait additive model in the Test set (T50 individual sampling strategy) by genomic relationships matrices (in columns) and three different traits (in rows): height1, budburst, and rust1. The accuracies distribution is represented by a violin plot and their median by the dot.

trait architecture of the traits studied rather than a few underlying QTLs with a substantial effect (Zhang et al., 2016).

Capturing more genetic variance with marker-based models did not result necessarily in a better prediction of the phenotype than using plain A models. Our prediction accuracy was already relatively high under pedigree evaluation, probably due to the fact of using a good evaluation design with enough repetitions and spatial adjustments at individual level. Markers did not help to improve this scenario or very little. Globally, when there was a difference between pedigree-based and genomic predictions, this occurred with G or Gw1 matrices. Using several weighting cycles (Gw2 and Gw3) did not show in any case better results. Comparable results with decreasing efficiency of several cycles of weighting were found in other recent studies (Teissier et al., 2018). Our results show little or no gain by increasing marker density, even when combining densification with a variable selection method, such as Gw. This lack of gain in accuracy may suggest that we have reached a plateau and that 7K markers are sufficient for this population. Some authors have already reported plateaus in performance when increasing the number of markers: in cocoa (Romero Navarro et al., 2017), wheat (Norman et al., 2018) and eucalyptus (Kainer et al., 2018). For eucalyptus, the plateau in correlation was still not reached at 500K, while for cocoa and wheat it was reached after thousands or tens of thousands of markers. Together with the fact that pedigree evaluations already obtained high levels of prediction accuracy, there is also the point that correcting pedigrees generally had a beneficial effect, making the resulting model truly competitive in some situations and with some traits compared to genome-based models. This is not new in forest assessments, given the fact that controlled crosses are cumbersome and prone to errors. In loblolly pine (Munoz et al., 2014) and in maritime pine (Bartholomé et al., 2016), pedigree errors led to decreases in predicting ability, and by completing or correcting the pedigree the predicting ability could be increased. In the maritime pine study (Bartholomé et al., 2016), the predicting ability was improved by the completion of the pedigree information in such a way that the genomic evaluation had little extra room for improvement in predicting ability. The error rate in our pedigree was 15%, involving in most cases wrong paternity attribution of complete or partial families, or individuals supposed to be different genetically.

In our study, model complexification using a dominance effect had no effect (positive or negative) on the quality of prediction. Our results are in line with previous studies. Several studies integrated dominance or epistatic effects in the GS. The results

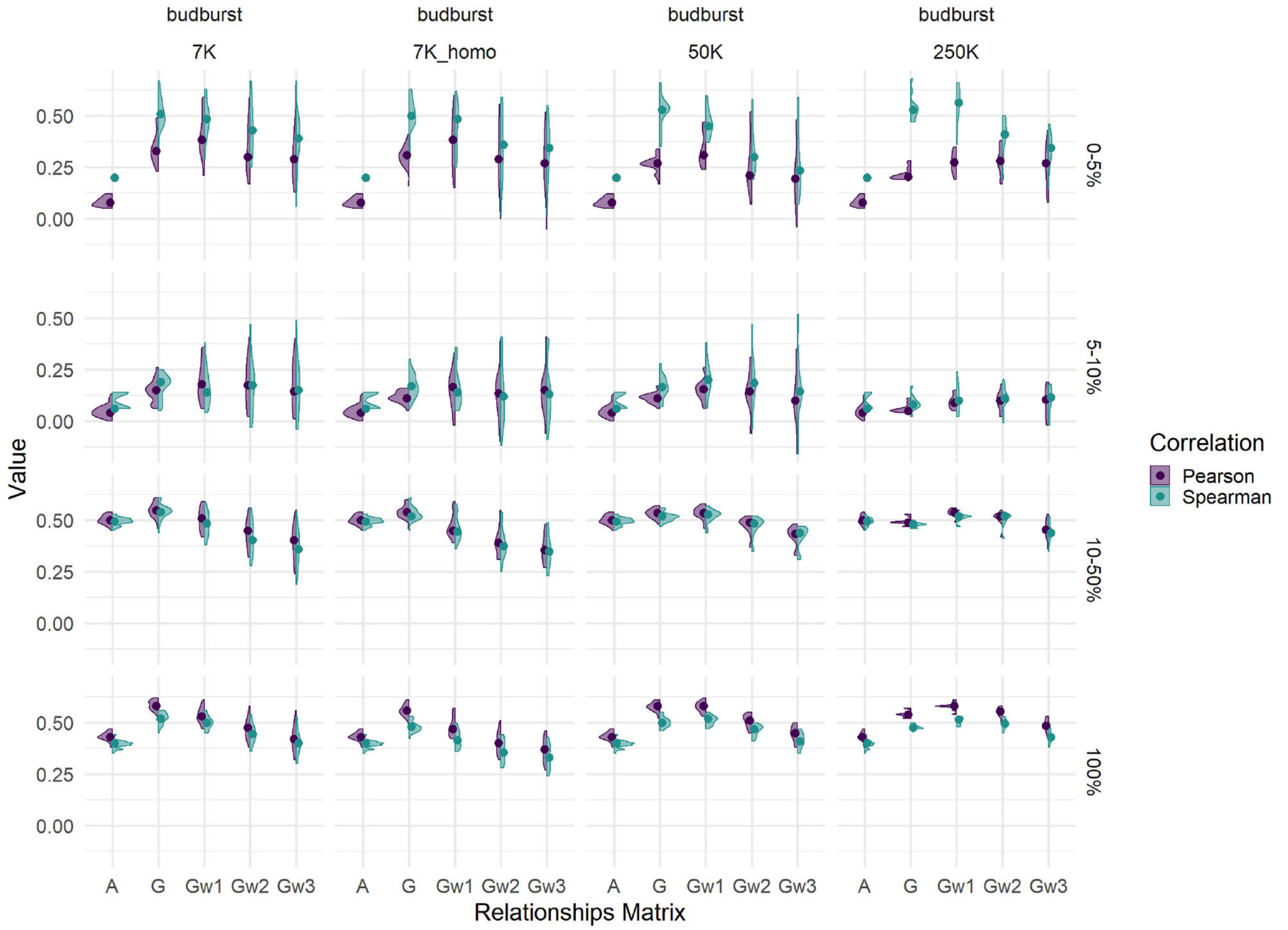


FIGURE 7 | Comparison of Spearman (green) and Pearson (purple) correlations between phenotypes and estimated breeding values for budburst in the Test set (T50 individual sampling strategy), for different relationship matrices (within panels abscissas) and SNP densities (across panel columns). Across panel rows represent the tier used for the calculation of correlations: 0–5% for the 5% best individuals; 5–10%, between the 5 and 10% best individuals; 10–50%, between the 10 and 50% best individuals, and 100% for the whole Test set.

on real datasets showed either no improvement in terms of accuracy (Heidaritabar et al., 2014; Gamal El-Dien et al., 2016; Jiang et al., 2017), even if a non-additive proportion of variance was observed for the traits, or a small improvement in prediction accuracy (Aliloo et al., 2016; Moghaddar and van der Werf, 2017; Tan et al., 2018). This so far limited success may be due to the fact that the populations under study were not big enough, nor with an optimal design to reveal the benefits of adding non-additive effects in genomic prediction. Despite a few strong genetic correlations in our population, the same observation can be drawn for the multi-trait approach, which did not bring a clear advantage to the quality of the predictions. One of the possible explanations could be found in the small difference in missing values between traits in our dataset. This has already been pinpointed as a cause of lack of performance by other authors working with a multi-trait approach (Jia and Jannink, 2012; Dos Santos et al., 2016; Lyra et al., 2017; Rambolarimanana et al., 2018). Multi-trait evaluation can help the prediction by compensating missing values in different traits and poor heritabilities (Calus and Veerkamp, 2011; Jia and Jannink, 2012; Marchal et al., 2016; Schulthess et al., 2016). It could also reduce prediction bias (Kadarmideen et al., 2003). An interesting and promising approach called “Trait-assisted genomic prediction” by Ben-Sadoun et al. (2020) allows to optimize the phenotyping cost by using a multiple-trait approach.

Apart from the general trends between pedigree vs. genomic models, results of prediction accuracy were fundamentally trait-dependent and mostly driven by the kind of training scenario being applied. This is clearly shown by the results of the Boruta algorithm, which found trait and training scenarios to be key features in explaining predicting accuracies. Similarly to other authors (Norman et al., 2018), we observed that prediction accuracy resulted in higher levels when the training and validation populations were closely related, as when the split between the two occurred at within family levels. On the contrary, prediction accuracy could be greatly affected when resulting from distant, independent validation sets. In our study, the cross-validation with individual sampling performed better than with family sampling, and this somehow limited the use of genomic evaluations to predict unobserved

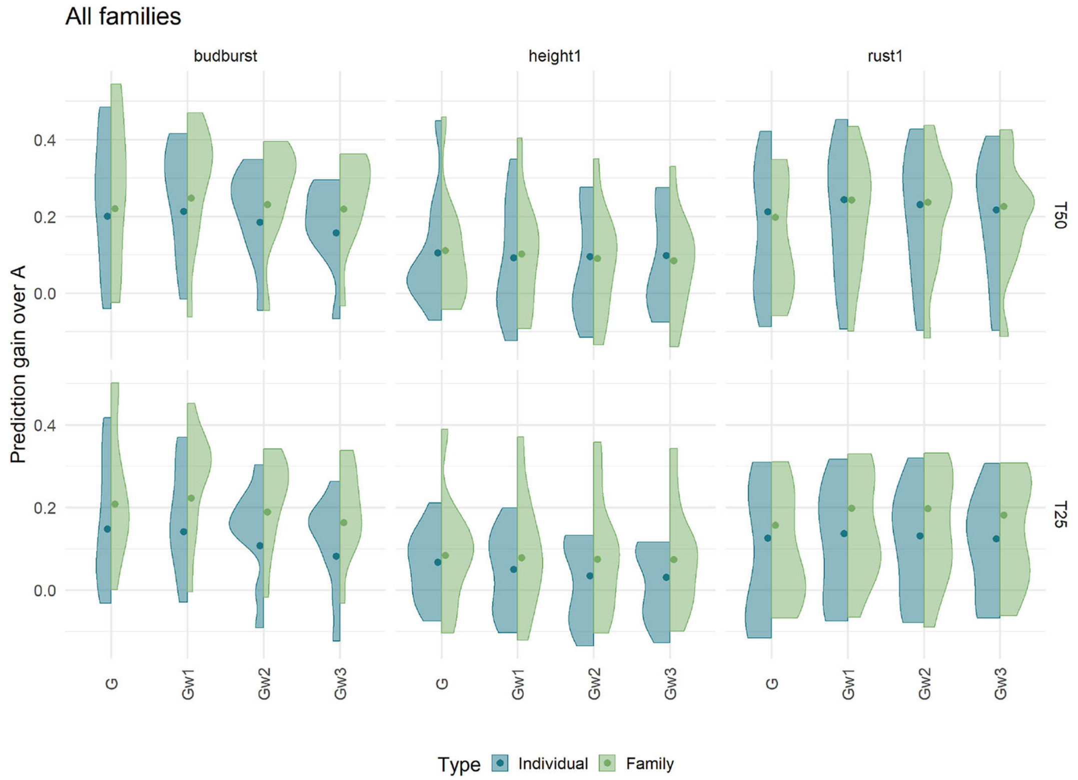


FIGURE 8 | Prediction gain compared to Pedigree based predicting ability within independent test families. Predicting abilities were obtained using an additive model with 7K SNP for three traits grouped by the proportion of individuals (Individual) or families (Family) in training sets 50% (T50) and 25% (T25). The color of violin plots correspond to the sampling strategy: in blue, the individual sampling strategy and in green the family sampling strategy. Each violin plot represented the accuracy of ten repetitions for each relationship matrix. The dot represented the weighted mean of the prediction gain, the mean was weighted by the number of offspring in each family.

crosses in our population with current approaches. The size of the training set used to develop prediction calibration is often cited as an important factor (Nakaya and Isobe, 2012). Curiously, the differences between our T50 and T25 schemes (50 and 25% of individuals to construct the calibration model, respectively) was not as large as one could expect and their performances overlapping to a large degree, making sometimes the differences between the two alternative training negligible. This is presumably very dependent on the properties of the populations being used for training.

4.2. Genomic Prediction Advantages Are Mostly Observed in Challenging Conditions

The choice of the training and validation sets is known to have a non-negligible impact on the prediction accuracy (Rincent et al., 2012). In that sense, our results showed that there was a substantial variation around each cross-validation realization, although often the ranking in performance between realizations was preserved across scenarios, notably for the individual sampling. In general, these cross-validation cases corresponded to operational situations where validation contributes with extra selection intensities, for instance, with new crosses from known parents or additional sibs across families to select from. One additional scenario of training that could be considered as especially challenging, corresponded to the validation set of newly obtained crosses from parents that were mostly underrepresented in the cross-validation sets. This could be seen as an operational demand to incorporate comparatively new material for selection. Our results showed that such challenges (represented by the validation in the test set) affected substantially the prediction accuracy across models, although G-BLUP and Gw1 were generally the most robust performers and pedigree-based evaluations the ones with the greatest loss overall. In the cross-validation scheme, the factorial design had a relatively large influence in demographic terms in the training set. Being a system that creates a well-interconnected network of families (Sørensen et al., 2005), the factorial design seemingly favored pedigree predictions to a level that made it competitive compared to genomic predictions in the cross-validation. However, the new testing set posed a challenging prediction problem to pedigree-based models, as the relatedness between training and validation was certainly weak to support quality predictions solely from a sparse A matrix. Despite that, the situation was not always a clear-cut difference between pedigree and genome-based evaluations, as shown by traits like height1.

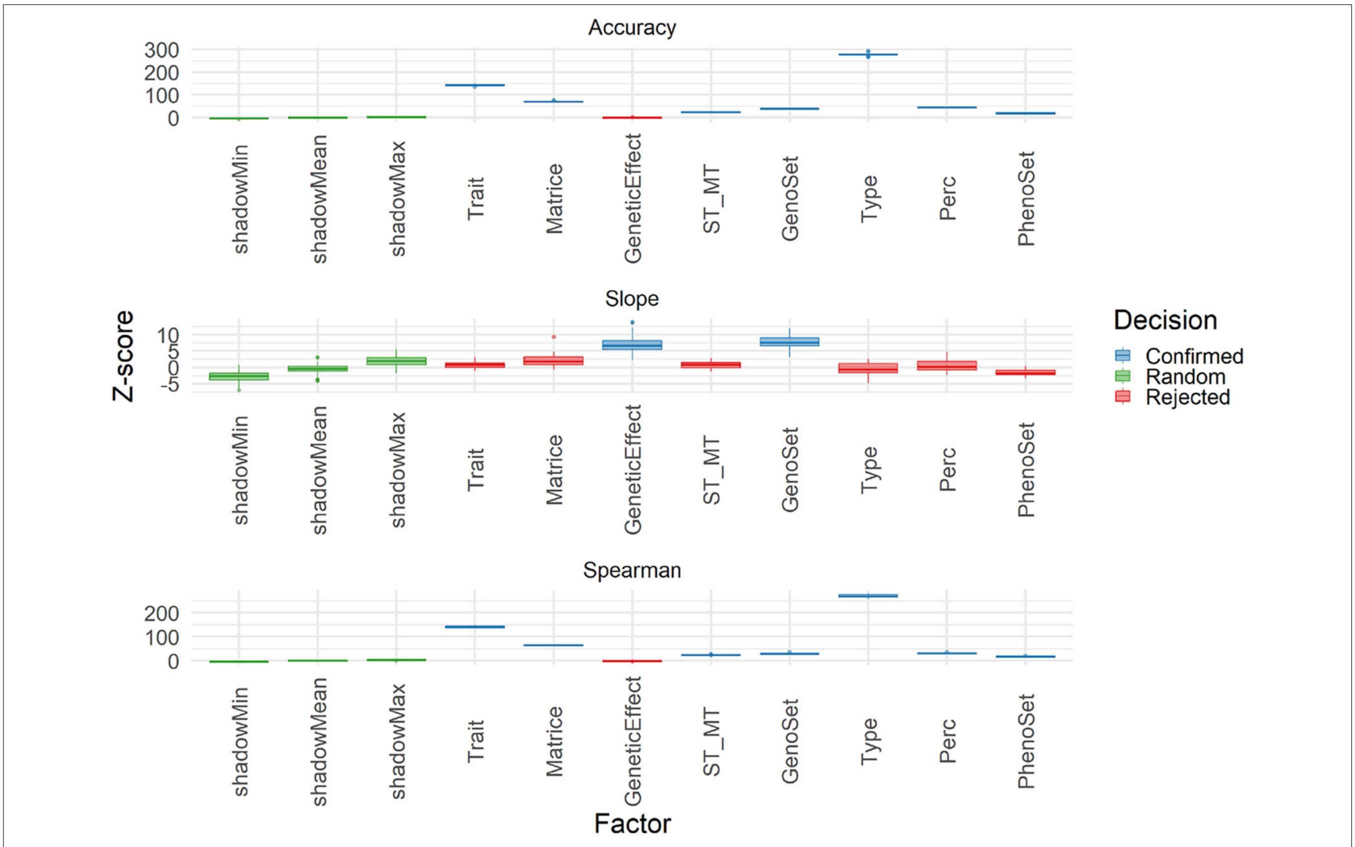


FIGURE 9 | Importance (Z-score) for each feature estimated with Boruta algorithm to explain Accuracy, slope, and Spearman correlation (Spearman) variability in the validation population. Boruta shadow features were ShadowMin, ShadowMean, and ShadowMax, as random references. The test factors were Trait (rust1, rust2, height1, height2, circ2, budburst, angbranch), Matrix (A, Acor, G, Gw1, Gw2, Gw3, D), GeneticEffect (Additive, Additive, and Dominance), ST_MT (Single-Trait, Multiple-Trait), GenoSet (none, 7K,7K_homo, 50K, 250K), Type (Individual, Family), and Perc (T50, T25). Algorithm decision for each factor, based on the significativity of the difference between factors and the shadow features are: shadow features (green), confirmed (blue), and rejected (red).

If the extent of relatedness thanks partly to the factorial design could have facilitated the competitiveness of pedigree-based predictions, the fact of using a high quality adjusted phenotype involving 6 repetitions was another element that could have a role in diminishing the differences between pedigree and genome-based performances in prediction terms. Actually, our results showed that downgrading the quality of clonal means used as phenotypes clearly had no differential effect between pedigree and genome-based predictions, with the latter retaining prediction quality at a level without replicate reduction. This evaluation simplification has also important operational implications for field evaluation, which need to be balanced with the genomic investments.

4.3. Genomic Prediction Enables the Ranking of Candidates to Selection

One of the main objectives of genetic evaluation is ultimately to rank individuals according to their breeding values, in order to use subsequently final selections as reproductors for the next generation. In that sense, identifying accurately the highest breeding values is a key element in genetic progress, and the use of predicting abilities based on a parametric correlation between predictions and true breeding values is one of the most common means of quality assessment (Daetwyler et al., 2013). This latter correlation shows a linear relationship with the genetic response (Falconer, 1981). For the poplar breeding program, however, the stress is given to the selection of genotypes for clonal dissemination at the production stage directly, rather than for gametic dispersion in seed orchards. This essential difference leads to the importance of ranking in selection decisions for poplars, as for any other domesticated species with clonal selection. When assessing the potential of genomic evaluations, it is essential to take into account the way predictions will be used for. Thus, we used alternative measures of prediction quality, like the slope of the regression of "true" breeding values on estimated breeding values. This slope represents a way to assess departures due to bias in predictions, generally caused by unequal representations of lineages in the training (Patry and Ducrocq, 2011), unbalanced data (Blair and Pollak, 1984), or the use of wrong variance estimation (Sorensen and Kennedy, 1984). Bias can lead eventually to wrong selection decisions when involving differently biased candidates. Our results suggest that G-BLUP was particularly affected by biases, with large departures toward greater slopes, i.e., best phenotypes gave proportionally

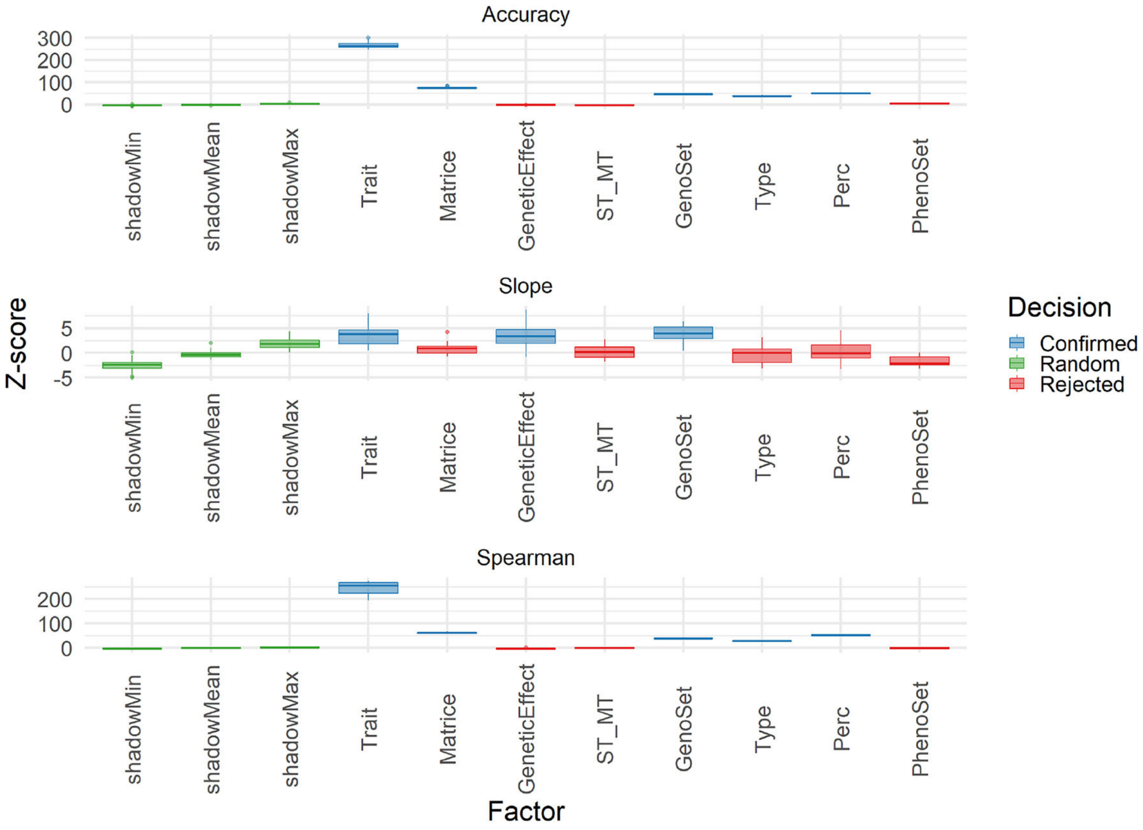


FIGURE 10 | Importance (Z-score) for each features estimated with Boruta algorithm to explain Accuracy, slope, and Spearman correlation (Spearman) variability in the TestSet population. Boruta shadow features were ShadowMin, ShadowMean, and ShadowMax. The test factors were Trait (rust1, rust2, height1, height2, circ2, budburst, angbranch), Matrix (A, Acor, G, Gw1, Gw2, Gw3, D), GeneticEffect (Additive, Additive and Dominance), ST_MT (Single-Trait, Multiple-Trait), GenoSet (none, 7K,7K_homo, 50K, 250K), Type (Individual, Family), and Perc (T50, T25). Algorithm decision for each factor, based on the significativity of the difference between factors and the shadow features are: in color: Green: shadow features (green), confirmed (blue), and rejected (red).

higher predictions than worst phenotypes. To a lesser extent, the best weighted G-BLUP (Gw1) also presented departures in slope. Comparatively, pedigree-based predictions were perfectly unbiased with slopes of one.

This result casted some doubts on the relevance of rankings derived from G-BLUP genomic predictions. We added an alternative measure of prediction quality, the Spearman correlation between predictions and true breeding values, which is a non-parametric estimate measuring the variation of the ranking. Moreover, this focus on ranking appeared as an appealing feature in the context of poplar breeding. Although less frequent in the literature than Pearson-based predicting abilities, a few authors used Spearman correlation to evaluate the prediction quality and to serve as criterion to select evaluation approaches (González-Recio et al., 2009; Mota et al., 2018). Some other authors suggest that individual ranking strategies could be more efficient (Blondel et al., 2015).

Our comparison of Spearman vs. Pearson correlations revealed that their differences in behavior were dependent on the selected tier in the distribution used for calculations, with Spearman magnifying the advantages of G-based models and high marker densities over pedigree for the best 5–10% tiers, the tail of the distribution that is usually relevant for selection. Pearson, on the other hand, attained its maximum correlation when considering the whole population. Such a difference in behavior could be of relevance when considering different levels of selection intensities, or weights given to each trait in a selection index. Usually, the interesting part of the distribution is the top percentiles, where Spearman could be a criterion of choice. However, in some cases the interest lies at intermediate values, like for budburst. The goal here is to have trees that do not budburst too early to avoid late frosts, nor too late to avoid shortening the growing season. For those central tiers, both correlations showed similar performances.

We have already pinpointed the fact that the population used for training, given the level of parental factorization in their mating, presented favorable conditions for pedigree-based evaluation. One condition where genome-based evaluation is expected to outperform a pedigree counterpart is when selecting at within-family levels. Our results showed that only genome-based evaluations were able to rank sibs with some degree of accuracy within family cohorts, where pedigrees do not bring any extra information. Although such advantage over the pedigree was not clear for all the families, a majority of them showed some

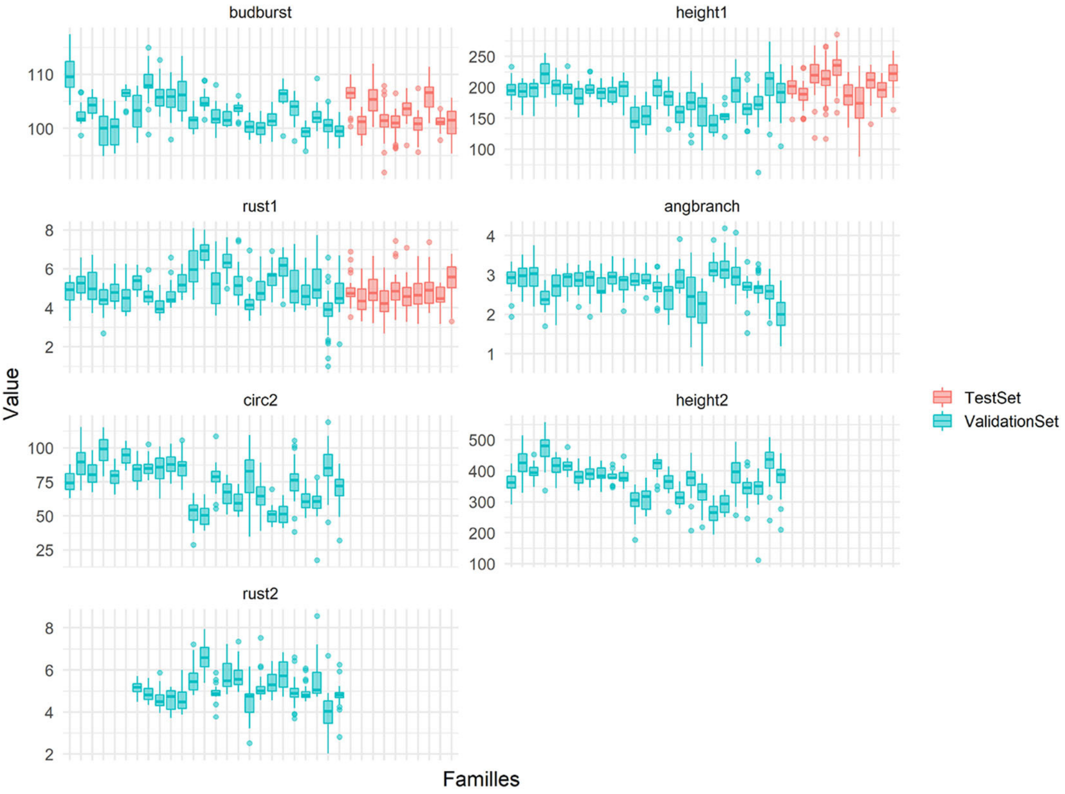


FIGURE 11 | Micro-environmentally adjusted phenotypic variability by full-sibs families (in x-axis) for the seven traits used in this study. In color the families includes in the TS/VS (in blue) or in the TestSet (in pink).

potential for gain over several traits. The fact that this ability did not translate into larger differences in our population could result from families of reduced size and/or from segregational variances too narrow to feed gain in a substantial way. While family sizes were not specially large for what is usual in breeding programs (on average 26 sibs per family), the variation at within-family level appeared indeed as notably reduced when compared with between-family differences (shown in **Figure 11**), and that for most of the traits in the analysis. This could be the result of a narrow parental variation in the training, but also from crossings between genetically similar parents, all characteristics of a reduced effective population size. Our initial estimates of effective population size (12) already pinpointed this narrow genetic diversity. A small effective size could explain to some extent the small difference that was found between our four training set scenarios, as well as the low impact of the densification in the number of markers. In that sense, it is clear that there is a need to expand this proof-of-concept approach with extra diversity.

4.4. Is There a Better Place in the Selection Scheme for Genomic Evaluation?

The present study took place at a particular step in the poplar breeding program, as illustrated in (**Figure 12**), specifically when evaluating selected candidates on juvenile traits in the nursery. The current selection scheme was the result of optimizing for many constraints derived from the phenotypic evaluation and operational factors over the years. It comprises several steps of selection conducted at the greenhouse, at the nursery, in the laboratory via *in vitro* tests and later in field trials, with each step implying different selection intensities and notably different selection accuracies. It is important to note that each selection step is done sequentially and conditionally onto the precedent (i.e., independent culling levels), instead of jointly and simultaneously, leading to inefficiencies with the risk of losing in the first steps important variation for subsequent steps. First steps of selection at the greenhouse and nursery are the less accurate, but the ones that screen most of the variation. Due to a limited field evaluation surface, a small number of individuals per family is kept for the next steps, reducing the phenotypic variance within each family. Conversely, later steps at the lab and in the fields are relatively accurate but screen through a subsample of original variation. Therefore, accuracy and genetic variation do not meet in a single same step for maximum efficiency in the current scheme.

Our test of genomic selection was performed with moderate to high heritability traits, well-evaluated in field trials, and on a relatively reduced set of individuals (with low effective population size) that were the result of two previous steps of selection conducted typically with a low precision and at a relatively high selection intensity (see **Figure 12**, with the red circle indicating where genomic evaluation was tested). These

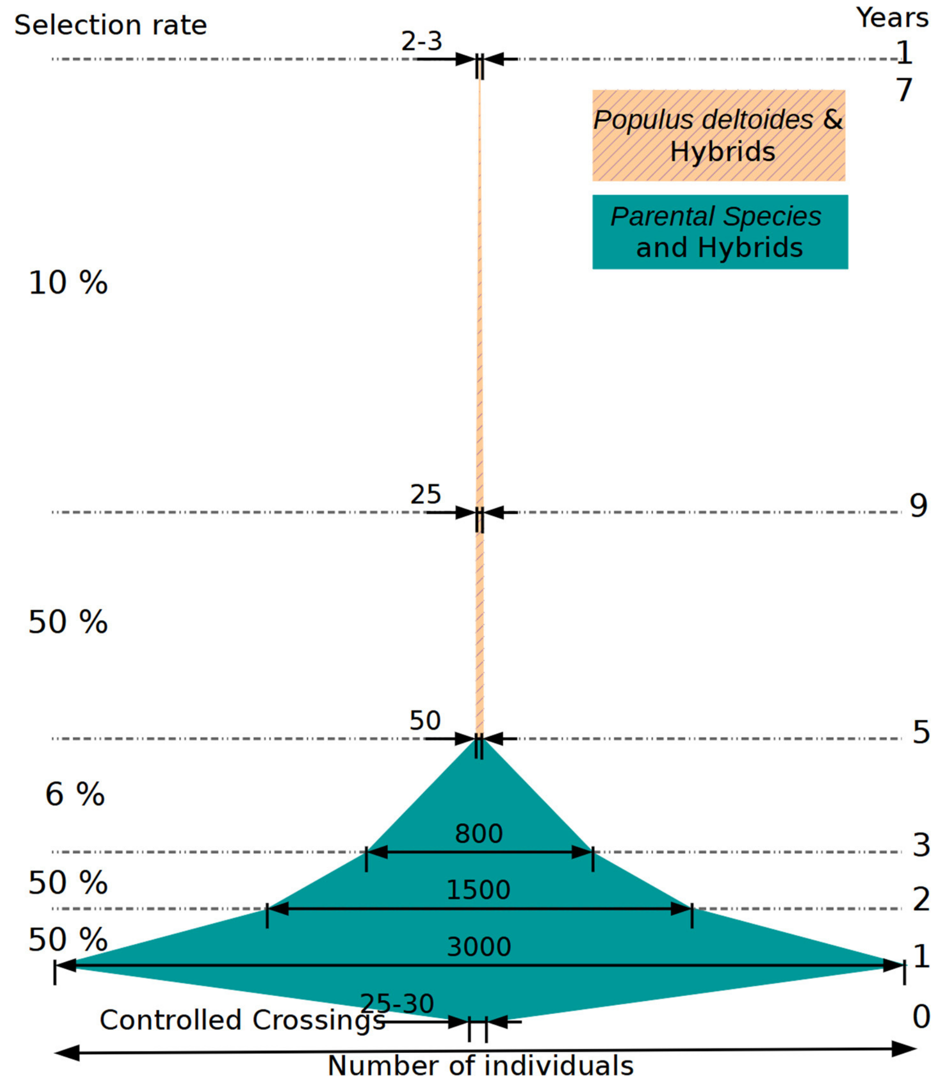


FIGURE 12 | Schematic representation of a breeding cycle in poplar, with the evolution in the number of individuals and the selection rate during the different steps of selection after crossings (year 0). Numbers correspond to one cycle of selection. Selection rate values correspond to a rate relative to the previous step. The place where the genome-based evaluation test was carried out is identified by a red circle.

conditions are often the ones encountered in late stages in breeding program cycles, when the implementation of genomic evaluation is typically devised, and where the precious genomic and phenotypic resources that are required are to be found. This is the case, probably, of other species undergoing domestication, with elites concentrating most of the evaluation resources, and founder bases only lightly evaluated. Theoretically, there is room for improvement in the way genomic evaluation is integrated in this kind of scheme, where extra precision is specially required: at the first stages of selection. Such a scenario would involve automatically larger effective population sizes than those used here. The only drawback of such an early implementation would certainly be the costs associated with a mass genotyping, involving thousands of candidates at the greenhouse. However, with current prices attaining record low levels every year, notably with custom SNP arrays shared between species (Silva-Junior et al., 2015; Gutierrez et al., 2017), such a possibility appears now within the reach of breeding program budgets. In the case of our study, sequencing had an average cost of 400*e* per individual, although with large variations due to techniques and depths, while genotyping experienced gradual reductions during data gathering from a starting 94*e* to late 46*e* per sample (not including chip design costs).

4.5. Recommendations for Future Studies in Genomic Evaluation in Poplar

One of the main limitations of the study was probably the use of a training population with a design that did not correspond necessarily to what is routinely done in poplar breeding. Indeed, the factorial mating design, although potentially interesting in terms of the parental variability, was more oriented for cognitive or mapping studies. This was partially overcome by the addition of extra families and crosses, well-connected to the breeding program. In that sense, a training population truly representative of the base population for the breeding program could have made more easily generalizable the results of the study. Resampling in the existing population is a good way to improve the training population and increase the prediction accuracy. For instance, to include 6–8 trees per family and evaluation site appears to be sufficient to guarantee an accurate estimation of genetic parameters for wood density and growth in an open pollinated test of black spruce (Perron et al., 2013). For some species (Cros et al., 2015; Tayeh et al., 2015), CDmeans has given good results in optimizing the training population (Rincent et al., 2012). Some preliminary work not shown in this study, however, suggested that there is no clear advantage for such an optimal procedure, and one of the reasons could be the lack of differentiation within the population to derive truly different training sets. The optimal procedure could also be tried with a denser SNP set, like the 50K. Another strategy to optimize the training step would be to integrate existing information in the pedigree and from genetic association studies in the way proposed by Cericola et al. (2017).

Further investigations are still necessary to improve the model prediction in terms of accuracy, but also to reduce systematic and overdispersion biases. The slope bias seemed to be positively correlated with the number of markers, while the use of variable selection models like wGBLUP was able to reduce the slope bias as density was allowed to increase. Density and marker distribution of the original 7K chip did not allow GS to get a clear advantage over the pedigree-based counterpart. Marker densities lower than 7K did not appear to be of interest here, given already the slight advantage at 7K. Marker selection could be optimized to select the best repartition. Our trial of an alternative SNP set with 7K being homogeneously distributed along the genome did not lead to gains in accuracy. The original 7K array was somehow enriched for markers in some genomic regions relevant for economically important traits (Faivre-Rampant et al., 2016). Alternatively, marker repartition could follow recombination rate maps obtained from a pedigreed population, enriching in SNPs around recombination hotspots. Such distributions could be combined with haplotypic approaches based on LD information. Some studies show that haplotypic approaches could increase the reliability of predictions because of the extra capture of linkage disequilibrium with respect to single SNPs (Hess et al., 2017).

Multi-trait and multi-environment evaluations are essential in plant and tree breeding programs, although performing single-step analyses in these circumstances could be methodologically and computationally challenging. In that sense, Montesinos-Lopez et al. (2018) have proposed efficient heuristic methods based on multi-trait deep learning (MTDL), which appear to be well-adapted when data is highly unbalanced, contain missing values data and there is a need for accommodating different design factors.

GS can contribute to accelerate genetic gain by increasing the individual selection accuracy at early stages, thus shortening the generation interval, and by increasing the selection intensity. We propose to implement GS sooner in the cycle, at the seedling stage, than what was assessed in this study. In the short term, a genomic selection scheme at the seedling stage, when there is a great number of individuals taking up the least space, would be of great benefit to the breeding program. Such an early scheme combined with a multi-trait approach with a selection index can increase the genetic gain in the short term for most traits simultaneously, even for those phenotyped at maturity like wood properties. For now, only the *P. nigra* parents could be selected with such early genome-based approach, and in order to identify the best black poplar parents at the same year as the controlled-crosses to produce both pure species descendants and hybrids with other species. Time-consuming and resource-intensive evaluations could then take place only on those genomically preselected parents, with the possibility to enlarge the panel of pre-selections. In the longer term, GS can be implemented in the other parental species, *P. deltoides*, and even at the hybrid progeny (Tan et al., 2017), depending on the breeding strategy for hybrids. In this case, in addition to the step at the nursery evaluation, new steps at the laboratory can focus on other targeted traits, like interaction genotype × rust strain and woolly aphid resistance for hybrids, increasing the accuracy of prediction for costly traits related to resistance. Such propositions could save eventually from 5 up to 9 years in the breeding program. One of the evaluations for which time gains are expected is that related to wood quality, with the interesting possibility of predicting potential uses at the individual level according to the wood properties.

However, there are limits to the rapid advancements of the cycle, and we can cite here two main ones: one is regulatory and the other is of biological nature. Even if accurate genomic evaluation is available at very early stages, the release of varieties under current regulations will require carrying out evaluations under production conditions in several environments, which usually takes 10 years. Biological constraints are related to sexual maturity. Indeed, if we want to use a selected individual from a parental species for hybridization, it is necessary to wait until sexual maturity at around 7 years of age. Another added problem when dealing with sex and early selection is the sex determination, which cannot be predicted accurately from markers (Müller et al., 2020). Sex prediction at early stages could indeed save resources among the selected candidates while waiting for sexual maturity for mating.

5. CONCLUSIONS AND PERSPECTIVES

Our proof-of-concept study shows that genomic evaluation advantages are context-dependent. Its performance could be comparable to the already well-optimized pedigree-based evaluation under certain standard conditions and with access to low to medium SNP density panels. Genomic evaluation appeared to be advantageous under less standard scenarios with a certain degree of challenge which have been pinpointed in our present work. Our study focused on a fairly advanced stage of the evaluation in the breeding program, where a substantial part of the variation has already been let aside by using pragmatic but less efficient early selections at the nursery (based on early growth, rooting ability . . .). We believe that genomic selection could be an interesting option at that early stage, where selection precision is typically poor and genetic variability abundant. Our study also showed that it is important to assess performances by looking at other alternative criteria, like those related to ranking, notably when these criteria respond to the operational context of the breeding program under scrutiny.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at https://data.inrae.fr/dataset.xhtml?persistentId=doi%3A10.15454%2F4FWANJ&version=DRAFT (doi: 10.15454/4FWANJ).

AUTHOR CONTRIBUTIONS

MP performed the analyses and drafted the manuscript. FM developed the scripts for spatial adjustment of phenotypes. VS contributed to the discussion on analytical models and data preparation, providing as well valuable scripts. CB provided the access to plant material and contributed to the view of the breeding program and ways of optimization as the scientist responsible for the *Populus nigra* breeding program. VJ and LS designed the study, discussed the analyses, assisted in drafting the manuscript and obtained funding. All authors significantly contributed to the present study, and read and approved the final manuscript.

FUNDING

This study was funded by the following sources: the INRA AIP Bioressource, EU NovelTree (FP7-211868), EU Evoltree (FP6-16322), and INRA SELGEN funding program (project BreedToLast) including sequencing and genotyping data. MP Ph.D. grant was jointly funded by INRA SELGEN funding program (BreedToLast) and by Region Centre—Val de Loire funding council. The Poplar GIS (MAAF) funded the acquisition of phenotypic data and the maintenance of the collections.

ACKNOWLEDGMENTS

The authors acknowledge the GIS peuplier, the UE GBFOR, and PNRGF (ONF) for access, maintenance, and sampling of plant material. The authors want to thank Vanina Guérin and Corinne Buret for their work on DNA extraction. The authors acknowledge Patricia Faivre-Rampant, Marie-Christine Le Paslier, and Aurélie Bérard from US EPGV for the genotyping and the sequencing. The authors acknowledge Philippe Barre for his detailed reading and his pertinent comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.581954/full#supplementary-material

REFERENCES

Akdemir, D., Sanchez, J. I., and Jannink, J. L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel. Evol.* 47, 1–10. doi: 10.1186/s12711-015-0116-6

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., and Hayes, B. J. (2016). Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. *Genet. Sel. Evol.* 48, 1–11. doi: 10.1186/s12711-016-0186-0

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., Fix, J., Van Tassell, C. P., et al. (2013). Methods of tagSNP selection and other variables affecting imputation accuracy in swine. *BMC Genet.* 14:8. doi: 10.1186/1471-2156-14-8

Ballesta, P., Bush, D., Silva, F. F., and Mora, F. (2020). Genomic predictions using low-density snp markers, pedigree and gwas information: a case study with the non-model species eucalyptus cladocalyx. *Plants* 9:99. doi: 10.3390/plants9010099

Bartholomé, J., Van Heerwaarden, J., 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

Ben-Sadoun, S., Rincent, R., Auzanneau, J., Oury, F., Rolland, B., Heumez, E., et al. (2020). Economical optimization of a breeding scheme by selective phenotyping of the calibration set in a multi-trait context: application to bread making quality. *Theor. Appl. Genet.* 133, 2197–2212. doi: 10.1007/s00122-020-03590-4

Bisognin, D. A. (2011). Breeding vegetatively propagated horticultural crops. *Crop Breed. Appl. Biotechnol.* 11, 35–43. doi: 10.1590/S1984-70332011000500006

Blair, H., and Pollak, E. (1984). Estimation of genetic trend in selected population with and without the use of control population. *J. Anim. Sci.* 58, 878–886. doi: 10.2527/jas1984.584878x

Blondel, M., Onogi, A., Iwata, H., and Ueda, N. (2015). A ranking approach to genomic selection. *PLoS ONE* 10:e128570. doi: 10.1371/journal.pone.0128570

Caballero, A. (2000). Interrelations between effective population size and other pedigree tools for the management of conserved populations interrelations between effective population size and other pedigree tools for the management of conserved populations. *Genet. Res.* 75, 26–27. doi: 10.1017/S0016672399004449

Calus, M., and Veerkamp, R. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* 43:26. doi: 10.1186/1297-9686-43-26

Cappa, E. P., and Cantet, R. J. (2007). Bayesian estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. *Can. J. For. Res.* 37, 2677–2688. doi: 10.1139/X07-116

Cappa, E. P., [de Lima,], B. M., [da Silva-Junior,], O. B., Garcia, C. C., Mansfield, S. D., and Grattapaglia, D. (2019). Improving genomic prediction of growth and wood traits in eucalyptus using phenotypes from non-genotyped trees by single-step gblup. *Plant Sci.* 284, 9–15. doi: 10.1016/j.plantsci.2019.03.017

Cappa, E. P., Muñoz, F., Sanchez, L., and Cantet, R. J. C. (2015). A novel individual-tree mixed model to account for competition and environmental heterogeneity: a Bayesian approach. *Tree Genet. Genomes* 11:120. doi: 10.1007/s11295-015-0917-3

Castellani, E., Freccero, V., Lapietra, G., and Castellani, E., and Freccero, V., and Lapietra, G. (1967). Proposta di una scala di differenziazione delle gemme fogliari del pioppo utile per gli interventi antiparas sitari. *Plant Biosyst.* 101, 355–360. doi: 10.1080/11263506709426301

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

Chamberland, V., Robichaud, F., Perron, M., Gélinas, N., Bousquet, J., and Beaulieu, J. (2020). Conventional versus genomic selection for white spruce improvement: a comparison of costs and benefits of plantations on quebec public lands. *Tree Genet. Genomes* 16, 1–16. doi: 10.1007/s11295-019-1409-7

Chen, Z.-Q., Baison, J., Pan, J., Karlsson, B., Andersson, B., Westin, J., et al. (2018). Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in norway spruce. *BMC Genomics* 19:946. doi: 10.1186/s12864-018-5256-y

Cros, D., Denis, M., Sánchez, L., Cochard, B., Flori, A., Durand-Gasselin, T., et al. (2015). Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). *Theor. Appl. Genet*. 128, 397–410. doi: 10.1007/s00122-014-2439-z

Cros, D., Mbo-Nkoulou, L., Bell, J. M., Oum, J., Masson, A., Soumahoro, M., et al. (2019). Within-family genomic selection in rubber tree (*Hevea brasiliensis*) increases genetic gain for rubber production. *Ind. Crops Products* 138:111464. doi: 10.1016/j.indcrop.2019.111464

Daetwyler, H. D., Calus, M. P., 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

de Almeida Filho, J. E., Guimarães, J. F. R., e Silva, F. F., de Resende, M. D. V., Muñoz, P., Kirst, M., et al. (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity (Edinb)*. 117:33. doi: 10.1038/hdy.2016.23

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

Dos Santos, J. P. R., De Castro Vasconcellos, R. C., Pires, L. P. M., Balestre, M., and Von Pinho, R. G. (2016). Inclusion of dominance effects in the multivariate GBLUP model. *PLoS ONE* 11:e152045. doi: 10.1371/journal.pone.0152045

Faivre-Rampant, P., Zaina, G., Jorge, V., Giacomello, S., Segura, V., Scalabrin, S., et al. (2016). New resources for genetic studies in *Populus nigra*: genome-wide SNP discovery and development of a 12K Infinium array. *Mol. Ecol. Resour*. 16, 1023–1036. doi: 10.1111/1755-0998.12513

Falconer, D. S. (1981). *Introduction to Quantitative Genetics, 2nd Edn*. Longman. p. 340.

Forni, S., Aguilar, I., and Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* 43:1. doi: 10.1186/1297-9686-43-1

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., and El-Kassaby, Y. A. (2015). Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics* 16:370. doi: 10.1186/s12864-015-1597-y

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., and El-Kassaby, Y. A. (2016). Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from nonadditive genetic effects. *G3 Genes Genomes Genet.* 6, 743–753. doi: 10.1534/g3.115.025957

Gianola, D. L., and Fernando, R. (2020). A multiple-trait Bayesian Lasso for genome-enabled analysis and prediction of complex traits. *Genetics*. 214, 305–331. doi: 10.1534/genetics.119.302934

Gilmour, A. R., Gogel, B. J., Cullis, B. R., and Thompson, R. (2009). *ASReml user guide release 3.0*. Hemel Hempstead: VSN International Ltd.

González-Recio, O., Gianola, D., Rosa, G. J., Weigel, K. A., and Kranis, A. (2009). Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. *Genet. Sel. Evol*. 41:3. doi: 10.1186/1297-9686-41-3

Grattapaglia, D., and Resende, M. D. V. (2011). Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7, 241–255. doi: 10.1007/s11295-010-0328-4

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet*. 15:30. doi: 10.1186/1471-2156-15-30

Gutierrez, A. P., Turner, F., Gharbi, K., Talbot, R., Lowe, N. R., Peñaloza, C., et al. (2017). Development of a medium density combined-species SNP array for pacific and european oysters (*Crassostrea gigas* and *Ostrea edulis*). *G3 Genes Genomes Genet.* 7, 2209–2218. doi: 10.1534/g3.117.041780

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

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

Heidaritabar, M., Vereijken, A., Muir, W. M., Meuwissen, T., Cheng, H., Megens, H. J., et al. (2014). Systematic differences in the response of genetic variation to pedigree and genome-based selection methods. *Heredity (Edinb)*. 113, 503–513. doi: 10.1038/hdy.2014.55

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

Hess, M., Druet, T., Hess, A., and Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet. Select. Evol.* 49:54. doi: 10.1186/s12711-017-0329-y

Howe, G. T., Saruul, P., Davis, J., and Chen, T. H. (2000). Quantitative genetics of bud phenology, frost damage, and winter survival in an F2family of hybrid poplars. *Theor. Appl. Genet*. 101, 632–642. doi: 10.1007/s001220051525

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

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

Jiang, J., Shen, B., O'Connell, J. R., VanRaden, P. M., Cole, J. B., and Ma, L. (2017). Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. *BMC Genomics* 18, 1–13. doi: 10.1186/s12864-017-3821-4

Kadarmideen, H. N., Thompson, R., Coffey, M. P., and Kossaibati, M. A. (2003). Genetic parameters and evaluations from single-and multiple-trait analysis of dairy cow fertility and milk production. *Livest. Prod. Sci*. 81, 183–195. doi: 10.1016/S0301-6226(02)00274-9

Kainer, D., Stone, E. A., Padovan, A., Foley, W. J., and Külheim, C. (2018). Accuracy of genomic prediction for foliar terpene traits in *Eucalyptus polybractea*. *G3 Genes Genomes Genet.* 8, 2573–2583. doi: 10.1534/g3.118.200443

Kursa, M. B., and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J. Stat. Softw*. 36, 1–13. doi: 10.18637/jss.v036.i11

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

Legionnet, A., Muranty, H., and Lefèvre, F. (1999). Genetic variation of the riparian pioneer tree species *Populus nigra*. II. Variation in susceptibility to the foliar rust Melampsora larici-populina. *Heredity (Edinb)*. 82, 318–327. doi: 10.1038/sj.hdy.6884880

Lenz, P. R., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M., and Bousquet, J. (2017). Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (*Picea mariana*). *BMC Genomics* 18:335. doi: 10.1186/s12864-017-3715-5

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. doi: 10.1111/eva.12823

Liaw, A., and Wiener, M. (2002). Classification and regression by random forest. *R News* 2, 18–22. Available online at: https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf

Lyra, D. H., de Freitas Mendonça, L., Galli, G., Alves, F. C., Granato, Í. S. C., and Fritsche-Neto, R. (2017). Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. *Mol. Breed.* 37:80. doi: 10.1007/s11032-017-0681-1

Marchal, A., Legarra, A., Tisné, S., Carasco-Lacombe, C., Manez, A., Suryana, E., et al. (2016). Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. *Mol. Breed.* 36, 1–13. doi: 10.1007/s11032-015-0423-1

Martini, J. W., Gao, N., Cardoso, D. F., Wimmer, V., Erbe, M., Cantet, R. J., et al. (2017). Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). *BMC Bioinformatics* 18:3. doi: 10.1186/s12859-016-1439-1

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.

Miller, A. J., and Gross, B. L. (2011). From forest to field: perennial fruit crop domestication. *Am. J. Bot*. 98, 1389–1414. doi: 10.3732/ajb.1000522

Moghaddar, N., and van der Werf, J. H. (2017). Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. *J. Anim. Breed. Genet*. 134, 453–462. doi: 10.1111/jbg.12287

Montesinos-Lopez, O. A., Montesinos-Lópes, 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 Genet*. 8, 3829–3840. doi: 10.1534/g3.118.200728

Mota, R. R., Silva, F. F., e., Guimarães, S. E. F., Hayes, B., Fortes, M. R. S., et al. (2018). Benchmarking Bayesian genome enabled-prediction models for age at first calving in Nellore cows. *Livest. Sci*. 211, 75–79. doi: 10.1016/j.livsci.2018.03.009

Müller, B. S., Neves, L. G., de Almeida Filho, J. E., Resende, M. F., Muñoz, P. R., dos Santos, P. E., et al. (2017). Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus. *BMC Genomics* 18:524. doi: 10.1186/s12864-017-3920-2

Müller, N. A., Kersten, B., Montalvão, A. P. L., Mähler, N., Bernhardsson, C., Bräutigam, K., et al. (2020). A single gene underlies the dynamic evolution of poplar sex determination. *Nat. Plants* 6, 630–637. doi: 10.1038/s41477-020-0672-9

Muñoz, F., and Sanchez, L. (2018). *breedR: Statistical Methods for Forest Genetic Resources Analysts*. Available online at: https://hal.archives-ouvertes.fr/hal-01269326

Munoz, P. R., Resende, M. D. M. F., Huber, D. A., Quesada, T., Resende, M. D. M. F., Neale, D. B., et al. (2014). Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy. *Crop Sci*. 54, 1115–1123. doi: 10.2135/cropsci2012.12.0673

Muñoz, P. R., Resende, M. F., Gezan, S. A., Resende, M. D. V., de los Campos, G., Kirst, M., et al. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* 198, 1759–1768. doi: 10.1534/genetics.114.171322

Muranty, H., Jorge, V., Bastien, C., Lepoittevin, C., Bouffier, L., and Sanchez, L. (2014). Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of mas in crops. *Tree Genet. Genomes* 10, 1491–1510. doi: 10.1007/s11295-014-0790-5

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

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, 2889–2899. doi: 10.1534/g3.118.200311

Patry, C., and Ducrocq, V. (2011). Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci*. 94, 1011–1020. doi: 10.3168/jds.2010-3804

Pegard, M., Rogier, O., Bérard, A., Faivre-Rampant, P., Le Paslier, M.-C., Bastien, C., et al. (2018). Sequence imputation from low density single nucleotide polymorphism panel in a black poplar breeding population. *bioRxiv*. doi: 10.1101/437426

Perron, M., DeBlois, J., and Desponts, M. (2013). Use of resampling to assess optimal subgroup composition for estimating genetic parameters from progeny trials. *Tree Genet. Genomes* 9, 129–143. doi: 10.1007/s11295-012-0540-5

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

Rambolarimanana, T., Ramamonjisoa, L., Verhaegen, D., and Tsy, J. L. P. (2018). Performance of multi-trait genomic selection for Eucalyptus robusta breeding program. *Tree Genet. Genomes* 14:71. doi: 10.1007/s11295-018-1286-5

Ratcliffe, B., El-Dien, O. G., Cappa, E. P., Porth, I., Klápťě, J., Chen, C., et al. (2017). Single-step BLUP with varying genotyping effort in open-pollinated *Picea glauca*. *G3 Genes Genomes Genet*. 7, 935–942. doi: 10.1534/g3.116.037895

Ratcliffe, B., El-Dien, O. G., Klápště, J., Porth, I., Chen, C., Jaquish, B., et al. (2015). A comparison of genomic selection models across time in interior spruce (*Picea engelmannii glauca*) using unordered SNP imputation methods. *Heredity (Edinb)*. 115, 547–555. doi: 10.1038/hdy.2015.57

Resende, M. D., Resende Jr, M. F., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., et al. (2012a). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol*. 194, 116–128. doi: 10.1111/j.1469-8137.2011.04038.x

Resende, M. F. R. D. V., Munoz, P., Resende, M. F. R. D. V., Garrick, D. J., 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

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

Romero Navarro, J. A., Phillips-Mora, W., Arciniegas-Leal, A., Mata-Quirós, A., Haiminen, N., Mustiga, G., et al. (2017). Application of genome wide association and genomic prediction for improvement of cacao productivity and resistance to black and frosty pod diseases. *Front. Plant Sci*. 8:1905. doi: 10.3389/fpls.2017.01905

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. *Theor. Appl. Genet*. 129, 273–287. doi: 10.1007/s00122-015-2626-6

Silva-Junior, O. B., Faria, D. A., and Grattapaglia, D. (2015). A flexible multi-species genome-wide 60k SNP chip developed from pooled resequencing of 240 eucalyptus tree genomes across 12 species. *New Phytol*. 206, 1527–1540. doi: 10.1111/nph.13322

Sørensen, A. C., Berg, P., and Woolliams, J. A. (2005). The advantage of factorial mating under selection is uncovered by deterministically predicted rates of inbreeding. *Genet. Select. Evol*. 37:57. doi: 10.1186/1297-9686-37-1-57

Sorensen, D., and Kennedy, B. (1984). Estimation of genetic variances from unselected and selected populations. *J. Anim. Sci*. 59, 1213–1223. doi: 10.2527/jas1984.5951213x

Souza, L. M., Francisco, F. R., Gonçalves, P. S., Scaloppi Junior, E. J., Le Guen, V., Fritsche-Neto, R., et al. (2019). Genomic selection in rubber tree breeding: a comparison of models and methods for managing G × E interactions. *Front. Plant Sci*. 10:1353. doi: 10.3389/fpls.2019.01353

Su, G., Madsen, P., Nielsen, U. S., Mäntysaari, E. A., Aamand, G. P., Christensen, O. F., et al. (2012). Genomic prediction for Nordic Red Cattle using one-step and selection index blending. *J. Dairy Sci*. 95, 909–917. doi: 10.3168/jds.2011-4804

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., and Ingvarsson, P. K. (2017). Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC Plant Biol*. 17:110. doi: 10.1186/s12870-017-1059-6

Tan, B., Grattapaglia, D., Wu, H. X., and Ingvarsson, P. K. (2018). Genomic relationships reveal significant dominance effects for growth in hybrid Eucalyptus. *Plant Sci*. 267, 84–93. doi: 10.1016/j.plantsci.2017.11.011

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*. 6:941. doi: 10.3389/fpls.2015.00941

Teissier, M., Larroque, H., and Robert-Granié, C. (2018). Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: a quantitative trait influenced by a major gene. *Genet. Sel. Evol*. 50, 1–12. doi: 10.1186/s12711-018-0400-3

Toro, M. A., and Varona, L. (2010). A note on mate allocation for dominance handling in genomic selection. *Genet. Sel. Evol*. 42, 1–9. doi: 10.1186/1297-9686-42-33

Ukrainetz, N. K., and Mansfield, S. D. (2020). Assessing the sensitivities of genomic selection for growth and wood quality traits in lodgepole pine using bayesian models. *Tree Genet. Genomes* 16:14. doi: 10.1007/s11295-019-1404-z

VanRaden, P. M. (2007). Genomic measures of relationship and inbreeding. *Interbull Bull*. 25, 111–114.

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

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195, 1223–1230. doi: 10.1534/genetics.113.155176

Wang, H., Misztal, I., Aguilar, I., Legarra, A., and Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res. (Camb)*. 94, 73–83. doi: 10.1017/S0016672312000274

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

Wientjes, Y. C., Veerkamp, R. F., and Calus, M. P. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193, 621–631. doi: 10.1534/genetics.112.146290

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H. J., Wang, Y., and Schön, C. C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* 195, 573–587. doi: 10.1534/genetics.113.150078

Wolak, M. E. (2012). nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. *Methods Ecol. Evol.* 3, 792–796. doi: 10.1111/j.2041-210X.2012.00213.x

Zhang, X., Lourenco, D., Aguilar, I., Legarra, A., and Misztal, I. (2016). Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. *Front. Genet.* 7:151. doi: 10.3389/fgene.2016.00151

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 Plant Science

ORIGINAL RESEARCH
published: 09 November 2020
doi: 10.3389/fpls.2020.583277

Accurate Prediction of a Quantitative Trait Using the Genes Controlling the Trait for Gene-Based Breeding in Cotton

Yun-Hua Liu[1†], Yang Xu[2†], Meiping Zhang[1†], Yanru Cui[2], Sing-Hoi Sze[3], C. Wayne Smith[1], Shizhong Xu[1] and Hong-Bin Zhang[2]**

[1] Department of Soil and Crop Sciences, Texas A&M University, College Station, TX, United States, [2] Botany and Plant Sciences, University of California, Riverside, Riverside, CA, United States, [3] Department of Computer Science and Engineering and Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX, United States

OPEN ACCESS

Edited by:
Sarah J Hearne,
International Maize and Wheat
Improvement Center (Mexico), Mexico

Reviewed by:
Paulino Pérez-Rodríguez,
Colegio de Postgraduados
(COLPOS), Mexico
Yongzhong Xing,
Huazhong Agricultural University,
China

***Correspondence:**
Shizhong Xu
shizhong.xu@ucr.edu
Hong-Bin Zhang
hbz7049@tamu.edu

†These authors have contributed
equally to this work

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: 14 July 2020
Accepted: 15 October 2020
Published: 09 November 2020

Citation:
Liu Y-H, Xu Y, Zhang M, Cui Y,
Sze S-H, Smith CW, Xu S and
Zhang H-B (2020) Accurate
Prediction of a Quantitative Trait Using
the Genes Controlling the Trait
for Gene-Based Breeding in Cotton.
Front. Plant Sci. 11:583277.
doi: 10.3389/fpls.2020.583277

Accurate phenotype prediction of quantitative traits is paramount to enhanced plant research and breeding. Here, we report the accurate prediction of cotton fiber length, a typical quantitative trait, using 474 cotton (*Gossypium* ssp.) fiber length (*GFL*) genes and nine prediction models. When the SNPs/InDels contained in 226 of the *GFL* genes or the expressions of all 474 *GFL* genes was used for fiber length prediction, a prediction accuracy of $r = 0.83$ was obtained, approaching the maximally possible prediction accuracy of a quantitative trait. This has improved by 116%, the prediction accuracies of the fiber length thus far achieved for genomic selection using genome-wide random DNA markers. Moreover, analysis of the *GFL* genes identified 125 of the *GFL* genes that are key to accurate prediction of fiber length, with which a prediction accuracy similar to that of all 474 *GFL* genes was obtained. The fiber lengths of the plants predicted with expressions of the 125 key *GFL* genes were significantly correlated with those predicted with the SNPs/InDels of the above 226 SNP/InDel-containing *GFL* genes ($r = 0.892$, $P = 0.000$). The prediction accuracies of fiber length using both genic datasets were highly consistent across environments or generations. Finally, we found that a training population consisting of 100–120 plants was sufficient to train a model for accurate prediction of a quantitative trait using the genes controlling the trait. Therefore, the genes controlling a quantitative trait are capable of accurately predicting its phenotype, thereby dramatically improving the ability, accuracy, and efficiency of phenotype prediction and promoting gene-based breeding in cotton and other species.

Keywords: quantitative trait, phenotype prediction, fiber length, fiber length gene, genic SNP, gene expression, *Gossypium*

INTRODUCTION

Many traits of agricultural and medical importance, such as crop yield, livestock productivity and human diseases, are known as quantitative traits that are each controlled by numerous genes. Therefore, it has been one of the principle aims and interests of current molecular and genomic research to accurately predict the phenotypes of quantitative traits for progeny selection using omic data, thereby enhancing the ability, accuracy, and efficiency of breeding in crop plants (Crossa et al., 2010, 2013; De Los Campos et al., 2010b; Heffner et al., 2011a,b; González-Camacho et al., 2012; Gouy et al., 2013; Desta and Ortiz, 2014; Xu et al., 2014, 2016; Beyene et al., 2015; Dan et al., 2016) and livestock (Meuwissen et al., 2001; Daetwyler et al., 2012; Morota et al., 2014), and medicine in humans (Khan et al., 2001; Lee et al., 2008; De Los Campos et al., 2010a; Speed and Balding, 2014; Weissbrod et al., 2016). This has been known as genomic selection (GS) in crop plant and livestock breeding (Meuwissen et al., 2001; Desta and Ortiz, 2014) and as genomic medicine in humans (De Los Campos et al., 2010a). A so-called training population, usually a subpopulation of individuals randomly selected from a targeted breeding population, is both phenotyped and genotyped, and used to train and validate a statistical prediction model. The utility and efficiency of the trained model for phenotype prediction of the objective trait are often estimated by prediction accuracy presented by Pearson's correlation coefficient between observed and predicted phenotypes. The remaining individuals of the targeted population are genotyped only and their genetic values or phenotypes of the objective trait are then estimated using the trained and validated prediction model. The predicted phenotypes of the trait for the individuals of the targeted population are finally used to make decision for progeny selection in crop plant and livestock breeding, and for medicine practice in humans (De Los Campos et al., 2010a).

Because of their polygenic controls and sensitivity to varying environments, accurate prediction of quantitative traits is very challenging. Initially, genome-wide DNA markers were used to predict the phenotypes of quantitative traits (Meuwissen et al., 2001; Lee et al., 2008; Crossa et al., 2010, 2013; De Los Campos et al., 2010b; Heffner et al., 2011a,b; Daetwyler et al., 2012; González-Camacho et al., 2012; Gouy et al., 2013; Morota et al., 2014; Speed and Balding, 2014; Xu et al., 2014; Beyene et al., 2015; Weissbrod et al., 2016). Then, genome-wide gene expressions (Takagi et al., 2014; Xu et al., 2016) and genome-wide metabolites (Dan et al., 2016; Xu et al., 2016) have been used to improve the prediction accuracy of the trait phenotype. Attempts have been also made to improve the prediction accuracy of quantitative traits by increasing training population size, from hundreds to thousands of lines, and/or increasing the omic dataset size, from hundreds to millions of features (Lee et al., 2008; González-Camacho et al., 2012; Speed and Balding, 2014; Xu et al., 2016). Furthermore, approximately 20 statistical multiple regression models, including parametric and non-parametric, have been tested for the phenotype prediction of quantitative traits using the omic features (Desta and Ortiz, 2014; Speed and Balding, 2014; Weissbrod et al., 2016). These efforts have improved the prediction accuracy of quantitative traits, but the prediction accuracy still remains relatively low for the quantitative traits thus far investigated. The lower prediction accuracy and increased cost for phenotype prediction, due to the increased numbers of DNA markers and/or training population size, have substantially influenced applications of GS in practical breeding in crop plants and livestock. Most importantly, plant or livestock breeding usually consists of three parts: parent selection, cross design, and progeny selection. GS is effective for progeny selection, but it is ineffective for parent selection and cross design, while both are crucial to success of plant or livestock breeding.

Therefore, Zhang et al. (2020a), for the first time worldwide, proposed a novel molecular breeding technology, designated gene-based breeding (GBB), and demonstrated its utility and efficiency for enhanced breeding for maize grain yield. GBB is designed to develop new varieties by design by making full use of the genes controlling the objective trait(s), especially the number of their favorable alleles (NFAs), their SNPs/InDels as DNA markers and/or their expression abundances as omic features, through the entire breeding process, including parent selection, cross design, and progeny selection. Zhang et al. (2020a) showed that the prediction accuracy of maize grain yield using either of these three datasets of the grain yield genes for GBB was over 60% more accurate and several-fold more cost-efficient than those with genome-wide random SNPs. When the phenotypes of grain yield predicted with two or all of three datasets of the genes were jointly used for progeny selection, the top 10% plants selected using the predicted grain yields were completely consistent with those selected based on the grain yields of the plants determined by replicated field trials. Therefore, their results showed that GBB is promising to substantially continue crop improvement. Nevertheless, additional research is needed to test the utility and efficiency of GBB for different traits in different species and to optimize it for enhanced breeding of different crops and livestock.

In the present study, we explored the ability, utility, and efficiency of the genes significantly contributing to quantitative traits for prediction of their phenotypes using fiber length as the objective trait in cotton. Cotton, including *Gossypium hirsutum* L. (Upland cotton) and *Gossypium barbadense* L. (Sea Island cotton), is the world's leading textile fiber crop and an important oilseed crop. Fiber length is a typical quantitative trait and also one of the economically most important fiber quality traits for the textile industry and cotton fiber produce. We previously cloned 474 *GFL* (*Gossypium* fiber length) genes significantly contributing to fiber length (upper half mean length, UHML) and estimated their effects on fiber length (Liu, 2014). In this study, we investigated the phenotype prediction ability and efficiency of cotton fiber length for gene-based breeding using these *GFL* genes. We also discussed the applicability of the concepts and methods obtained in the present study to development of GBB for enhanced breeding in other crops and livestock of agricultural importance.

MATERIALS AND METHODS

Plant Materials and Fiber Length Phenotyping

One hundred ninety-eight recombinant inbred lines (RILs) at F_7, F_8, and F_9 generations derived by the single-seed descent method from a cross of TAM 94L-25 (*G. hirsutum*) x NMSI 1331 (*G. barbadense*) were used for this study. These RILs and their parents were grown at the Texas A&M AgriLife Research Farm near College Station, TX, United States, in 2009 (F_7), 2010 (F_8), and 2011 (F_9) to phenotype their fiber lengths. The 2010 and 2011 field trials were performed in a randomized complete block

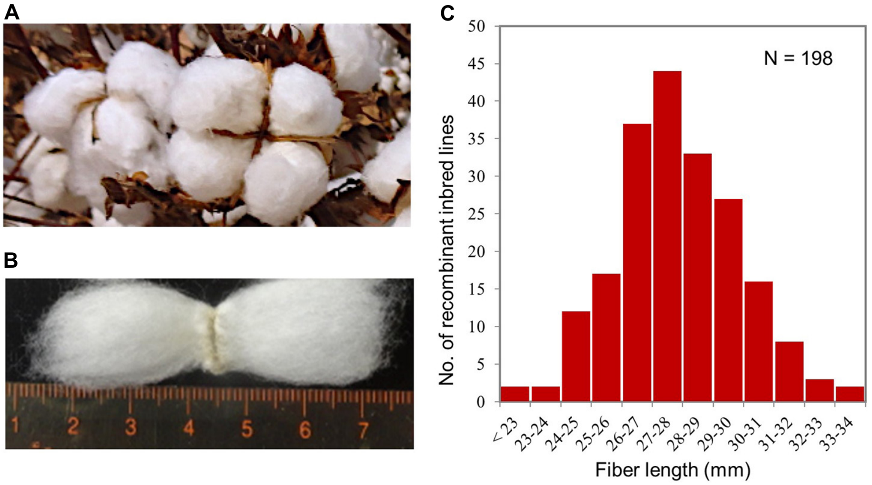


FIGURE 1 | Field trial of the RIL population for fiber length phenotyping. **(A)** Matured fiber bolls used for fiber length phenotyping. **(B)** Fiber lengths. **(C)** Variation of fiber length in UHML, measured by high-volume instrumentation, in the RIL population showing that fiber length is a typical quantitative trait. The fiber length data were collected from the 2011 field trial and the mean fiber lengths of three replicates.

design, with three replicates, while the 2009 trial only included a single five-plant plot per line, with no replication, because it was used for seed production for the 2010 and 2011 trials. The field practices followed those used for standard cotton breeding trials in our cotton breeding program. When the fiber bolls completely ripened (**Figure 1A**), they were hand-harvested from entire plots and ginned. A sample of the fibers from each line was used to measure its fiber length (**Figure 1B**), presented as upper half mean length (UHML), using High-Volume Instrumentation (HVI) at Fiber and Biopolymer Research Institute, Texas Tech University, Lubbock, TX, United States.

The mean fiber length of each line was calculated from those of the three replicates for each of the 2010 and 2011 trials (**Figure 1C**). The fiber length of the 2009 trial was from single five-plant entry. The broad sense heritability (H^2) of fiber length was estimated separately for the 2010 and 2011 trials by subtracting the mean fiber length variance of the two parents among their entries ($n = 33$ for each parent) [$\sigma^2{}_e = (\sigma^2_{p1} + \sigma^2{}_{p2})/2$] from the fiber length variance of the 198 RILs ($\sigma^2{}_p$) and then dividing by the fiber length variance of the 198 RILs ($\sigma^2{}_p$).

Genes

GFL Genes

The 474 *GFL* genes were previously cloned by our laboratory and coded from 001 through 474 (Liu, 2014) were used for this study (**Supplementary Table S1A**; NCBI GenBank accession numbers: MW082098-MW082571). These 474 *GFL* genes included 17 of the 18 published fiber length genes (**Supplementary Tables S2**, **S3**; Zhang et al., 2020b). Liu (2014) showed that each of these *GFL* genes had an effect on fiber length varying from 2.6% to 7.9%, with 88.6% of them significantly decreasing and 11.4% significantly increasing fiber length, when activated or up-regulated (**Supplementary Table S1A**). Network analysis showed that for 19 of these 474 *GFL* genes, variation of their edge numbers in the *GFL* network was significantly associated with fiber length (**Supplementary Table S1B**) (Liu, 2014; for more related information, see Zhang et al., 2020b).

Published Fiber Length Genes

A literature search was conducted as of December 2014 and found that a total of 18 fiber length genes were cloned from cotton using different gene cloning methods, including gene expression repression (RNAi or antisense) and gene overexpression (**Supplementary Table S2**; Zhang et al., 2020b). These 18 published fiber length genes were used as the positive control to test the ability of the *GFL* genes to predict the phenotype of fiber length in this study.

Randomly Selected Cotton Unknown Non-474 *GFL* Genes

A cotton database consisting of 79,708 transcripts of developing fibers sampled on the 10th day of post-anthesis (10-dpa fibers) (Zhang et al., 2019) were used for sampling the randomly selected cotton unknown non-474 *GFL* genes used as the negative control in this study.

Gene Transcript Expression Profiling and Gene Transcript Expression Dataset Construction

The sequences of the TAM 94L-25 transcripts expressed in 10-dpa fibers (Zhang et al., 2019), including those of the 474 *GFL* genes, were used as the reference to determine the expression profiles of the targeted transcripts of the *GFL* genes in the 10-dpa developing fibers of each line. Because a plant gene may be alternatively spliced into multiple transcripts, with each transcript likely being translated into different proteins having different biological functions (Syed et al., 2012; Zhang et al., 2019), the expression abundances of only the transcripts of the *GFL* genes that are responsible for fiber length (Zhang et al., 2020b) were quantified as predictors for phenotype prediction of fiber length in this study. The targeted transcript expression abundance of each *GFL* gene in a line was quantified with the RNA-seq 100-nucleotide clean reads using the RSEM software (Li and Dewey, 2011) bundled with the Trinity software (Grabherr et al., 2011; Haas et al., 2013) and presented as Transcripts Per Million mapped reads (TPM) (**Supplementary Table S4**).

GFL SNP/InDel Genotyping and SNP/InDel Dataset Construction

We previously sequenced all the genes expressed in 10-dpa developing fibers of the cotton population from the 2011 trial (Liu, 2014; Zhang et al., 2019). In this study, we first identified the single nucleotide polymorphisms (SNPs) and/or nucleotide insertions/deletions (InDels) of all the expressed genes using the RNA-seq 100 nucleotide clean reads and SAMtools (Li et al., 2009; Li, 2011). The cotton acc. TM-1 genome (Zhang et al., 2015) was used as the reference. Only the SNPs or InDels identified at the same position in the two parents, TAM 94L-25 and NMSI 1331, and two or more lines were used for further analysis. Since the transcript assemblies of the expressed genes had an average length of 778 bp (Liu, 2014), the probability that the two parents and two RILs had an SNP or InDel at the same position by chance, such as sequencing, base calling, and/or transcript assembly errors, would be close to zero [$P = (1/778)^4 = 2.7\text{E-}12$]. This filtration excluded almost all SNPs or InDels, if not all, resulted from sequencing, base calling, and/or transcript assembly errors from this study.

Then, we extracted the SNPs and/or InDels (hereafter, SNPs/InDels) contained in the *GFL* genes. To identify the SNPs/InDels of the *GFL* genes that significantly influenced fiber length, we conducted association analysis between the *GFL* genic SNPs/InDels and fiber length using the single marker analysis method for QTL mapping (Liu, 1997). Given that cotton has a genome size of 2,450 Mb/1C, the probability of the *GFL* genic SNPs/InDels linked to a gene controlling fiber length within an interval of 10 Mb, if they were the SNPs/InDels contained in the *GFL* genes, would be extremely low [$(10/2{,}450)^2 = 1.67\text{E-}05$]. Therefore, the association of a *GFL* genic SNP/InDel with fiber length indicated that the SNP/InDel of the *GFL* gene highly likely had a significant effect on fiber length. Therefore, only the SNPs/InDels contained in the *GFL* genes significantly influenced fiber length ($P \leq 0.05$) were selected and used as DNA markers for this study. These genes were defined in this article as the SNP/InDel-containing *GFL* genes. Furthermore, the *GFL* genic SNPs were verified by allele-specific PCR using the genomic DNAs of four cotton genotypes, including the two parents of the cotton population, as templates (Gaudet et al., 2007).

For the construction of the *GFL* genic genotype dataset, their SNPs or InDels were scored as bi-allelic DNA markers, as those genome-wide SNPs used for prediction of phenotype for genomic selection. The homozygote for one allele was scored as "0," the homozygote for the other allele scored as "2," and their heterozygote scored as "1." Because cotton is a frequently outcrossing species and the RIL population used in this study was developed in the field condition, with no bagged selfing pollination, heterozygotes for some plants were expected, even though the RILs at F_7–F_9 generation were used for this study.

Fiber Length Prediction

Prediction of fiber length using the *GFL* genes was carried out with two genic datasets compiled above separately: (i) the SNPs or InDels contained in the SNP/InDel-containing *GFL* genes as DNA markers and (ii) the targeted transcript expressions of the *GFL* genes in 10-dpa developing fibers. Nine prediction models, including five parametric and four non-parametric models (Desta and Ortiz, 2014; Zhang et al., 2020b), that have been widely used for GS were used to predict fiber length using the *GFL* genes. The five parametric models were genomic best linear unbiased prediction (GBLUP) (VanRaden, 2008), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), partial least square (PLS) (Geladi and Kowalski, 1986), BayesA (González-Recio and Forni, 2011), and BayesB (González-Recio and Forni, 2011). The four non-parametric models were support vector machine using the radial basis function kernel (SVMRBF) (Maenhout et al., 2007), support vector machine using the polynomial kernel function (SVMPOLY) (Maenhout et al., 2007), random forest (RF) (Svetnik et al., 2003), and reproducing kernel Hilbert space regression (RKHS) (De Los Campos et al., 2010a). We tested these nine prediction models because some of them may not be well suited for these two datasets, while others may be well fitted for the prediction of fiber length using the datasets.

GBLUP was implemented in an R program (Xu et al., 2014); LASSO was implemented in the GlmNet/R program (Friedman et al., 2010); BayesA, BayesB, and RKHS were implemented in the BGLR package (Pérez and De Los Campos, 2014); SVMRBF and SVMPOLY were implemented in the kernlab R program (Karatzoglou et al., 2004); PLS was implemented using the pls R package (Mevik and Wehrens, 2007); and RF was implemented in an R program (Liaw and Wiener, 2018). Among the nine prediction models, several require tuning parameters, which were selected based on the 10-fold cross validation used for the prediction (see below). Parameter values that maximize the predictability (squared correlation between predicted and observed trait values) were chosen as the optimal values. The shrinkage parameter of LASSO was chosen in this way. For the PLS prediction, the number of components extracted was considered as a tuning parameter and was obtained *via* 10-fold cross validation also. For BayesA, BayesB, and RKHS, the number of iterations, burnIn and thin were set to 10000, 1000 and 10, respectively. For RKHS, a multi-kernel approach was used, as proposed by De Los Campos et al. (2010b), and the bandwidth parameter was set to {0.5, 2, 10}.

A 10-fold cross-validation scheme widely used for GS was used for the prediction of fiber length using the *GFL* genes. The 10-fold cross validation scheme was described in our previous study (Zhang et al., 2020a), with each subset consisting of 19 or 20 RILs and 100 replications.

Statistical Analysis

The statistical analyses, including the two-way ANOVA, Tukey's HSD (honest significant difference), and parametric correlation tests, were performed using an R program and Microsoft

Excel 2013. For the ANOVA and correlation tests, *P*-value was presented at a two-tailed significance, and for the Tukey's HSD test, a confidence interval (CI) of 95% was applied.

RESULTS

Variation of Cotton Fiber Length, and Transcript Expression Variation and SNPs/InDels of the *GFL* Genes

Phenotype analysis confirmed that the fiber length trait (**Figures 1A,B**) under this study exhibited a normal distribution (**Figure 1C**), the variation of a typical quantitative trait, for the field trials through all three years (2009, 2010, and 2011) and all three generations (F_7, F_8, and F_9) among the 198 RILs of the population studied. The fiber lengths of the population from the 2009, 2010, and 2011 trials varied from 23.0 to 34.6, 23.1 mm to 35.8 mm, and from 23.1 mm to 34.8 mm, respectively. **Figure 1C** shows the variation of fiber length determined through the 2011 field trial. The Pearson's correlation coefficients (r) of the fiber length phenotypes between the three replicates of the 2010 and 2011 trials were 0.80–0.85 (N = 164, P = 0.000) and 0.76 (N = 198, P = 0.000), respectively. The Pearson's correlation coefficients (r) of the fiber length phenotypes between the 2009, 2010, and 2011 trials were 0.67–0.91 (N = 164 or 198, P = 0.000), even though the weather of the trial location in 2011 was unusual hot and drought, which was quite different from those normal weathers in 2010 and 2009. The broad sense heritability of the fiber length was H^2 = 0.90 and 0.83 for 2010 and 2011, respectively, which were similar to those previously reported (Ulloa, 2006; Khan et al., 2010). We were unable to calculate the H^2 for 2009 because there was no replication for the parents for the 2009 trial to estimate the environmental variance (σ_e^2).

SNP/InDel analysis revealed that 400 of the 474 *GFL* genes contained one or more SNPs/InDels and 74 had no SNPs/InDels for the population. The 400 *GFL* genes had a total of 10,766 SNPs/InDels, with an average of 26.9 SNPs/InDels per gene. Gene mutation effect analysis showed that 740 (6.9%) of the SNPs/InDels contained in 226 of the 400 *GFL* genes, with an average of 3.2 SNPs/InDels per gene, significantly increased or decreased fiber length ($P \leq 0.05$) of the RILs (**Supplementary Tables S1C**, **S6**) by 2.1% to 22.6%. The multiple SNPs/InDels per *GFL* gene suggested that there are multiple alleles for a *GFL* gene, if each of its SNPs/InDels was considered to be biallelic. The number of SNPs that significantly influenced fiber length was expected, because a vast majority of the SNPs contained in protein-coding genes are known to be synonymous, not leading to protein sequence change and likely having no biological effects (Graur and Li, 2000). Furthermore, we randomly selected 20 SNPs from the 740 *GFL* SNPs/InDels, with one SNP from a *GFL* gene, and analyzed them by allele-specific PCR using the genomic DNAs of four cotton genotypes as templates, including the two parents of the population used in this study. The result confirmed the existence of all 20 SNPs in the four genotypes, with the sizes of the PCR products as expected (**Supplementary Figure S1**), thus confirming the *GFL* genic SNPs identified. Therefore, these 226 *GFL* genes were hereafter defined as SNP/InDel-containing *GFL* genes and further used as DNA markers for phenotype prediction of fiber length.

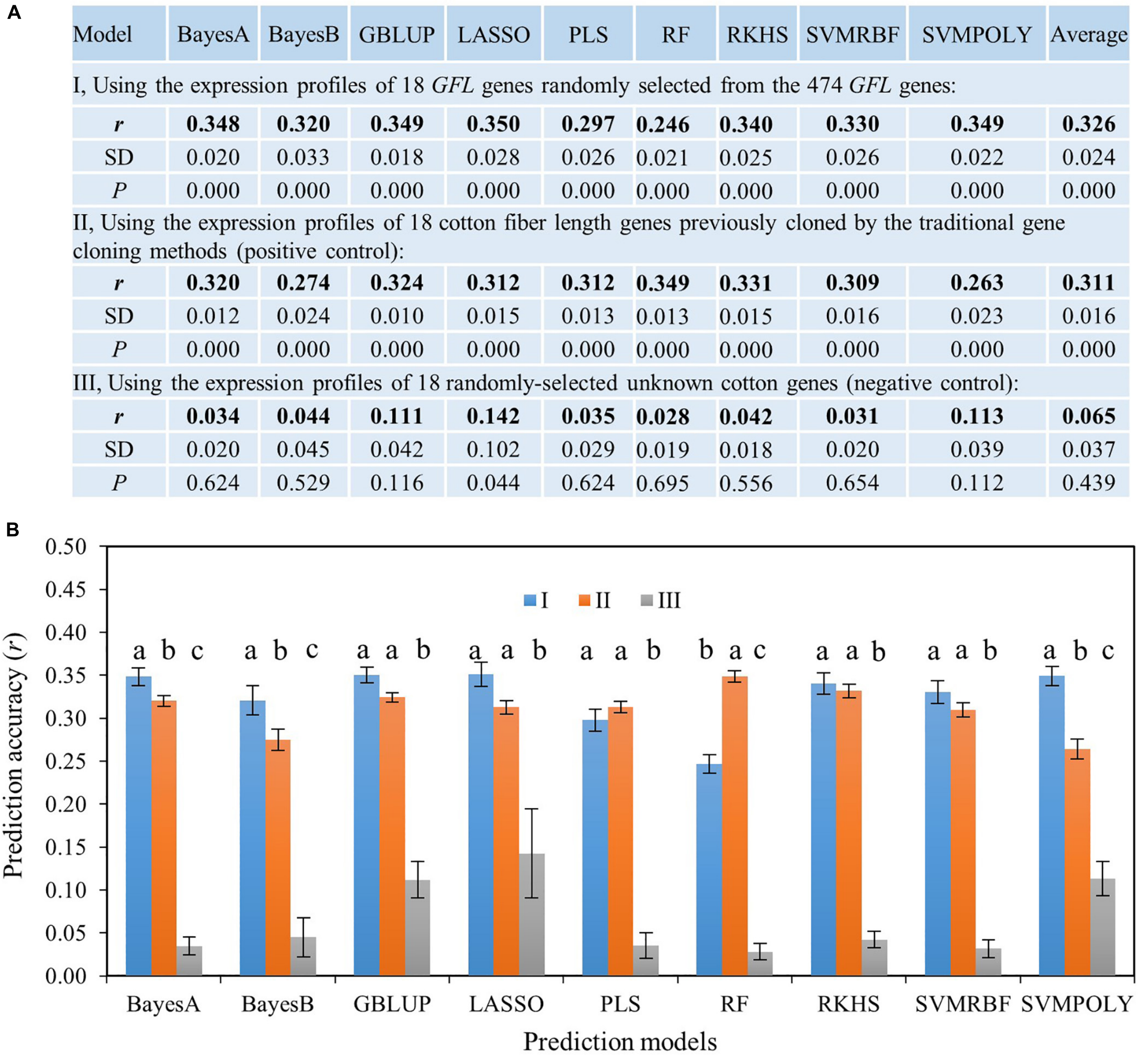
The 474 *GFL* genes all expressed in 10-dpa developing fibers of the population, but their expressions varied by thousands fold, from 0.75 TPM to 23,601 TPM (**Supplementary Table S4**). The expression of each *GFL* gene also varied dramatically among the RILs of the population, with a coefficient of variance (CV%) of 18.5%–202.5%. The expressions of all 474 *GFL* genes exhibited quantitative variations, with approximately 60% showing normal distributions and approximately 40% having distributions biased to lower expressions. Correlation analysis showed that the expressions of all 474 *GFL* genes in 10-dpa developing fibers were significantly correlated with the variation of the fiber length in the population ($P \leq 0.05$), which was consistent with the expression correlation of previously published fiber length genes (**Supplementary Tables S2**, **S3**) with the variation of fiber length (Zhang et al., 2020b). Therefore, both SNP/InDel and expression analyses further confirmed that the 474 *GFL* genes controlling fiber length.

Predicting the Phenotype of Fiber Length Using the *GFL* Genes

We tested the utility and efficiency of the *GFL* genes for phenotype prediction of fiber length for enhanced cotton fiber length breeding through GBB, especially progeny selection in this study, using expression abundances and SNP/InDel genotypes of the *GFL* genes. We first trained and validated the nine prediction models using the fiber length data collected from the 2011 trial, because the RILs of the population from the 2011 trial were also genotyped using the expressions and SNPs/InDels of the *GFL* genes. Then, we tested the utility and efficiency of the trained prediction model selected above for phenotype prediction of fiber length for the 2009 (F_7) and 2010 (F_8) trials using the genotypic data from the 2011 trial.

Predicting the Phenotype of Fiber Length Using the Expressions of the *GFL* Genes

We first tested the ability of the *GFL* genes for predicting the phenotype of fiber length, in which the published fiber length genes previously cloned by different researchers using different gene cloning methods (**Supplementary Tables S2**, **S3**) were used as the positive control. Since only 18 published genes controlling cotton fiber length were previously cloned as of December 2014, the ability of the *GFL* genes to predict the phenotype of fiber length was first evaluated using only 18 *GFL* genes randomly selected from these 474 *GFL* genes. These 18 published fiber length genes were used as the positive control, and 18 randomly selected unknown cotton genes were used as the negative control. Nine prediction models widely used for prediction of quantitative traits for GS and the expressions of the 18 *GFL* genes (**Supplementary Table S4**), 18 previously published fiber length genes (**Supplementary Table S2**) and 18 randomly selected unknown genes were used to predict fiber length, respectively. Results showed that only the randomly selected *GFL* genes and the published fiber length genes could predict the phenotype of fiber length, with a prediction accuracy of r = 0.246–0.350



A

Model	BayesA	BayesB	GBLUP	LASSO	PLS	RF	RKHS	SVMRBF	SVMPOLY	Average
I, Using the expression profiles of 18 *GFL* genes randomly selected from the 474 *GFL* genes:										
r	**0.348**	**0.320**	**0.349**	**0.350**	**0.297**	**0.246**	**0.340**	**0.330**	**0.349**	**0.326**
SD	0.020	0.033	0.018	0.028	0.026	0.021	0.025	0.026	0.022	0.024
P	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
II, Using the expression profiles of 18 cotton fiber length genes previously cloned by the traditional gene cloning methods (positive control):										
r	**0.320**	**0.274**	**0.324**	**0.312**	**0.312**	**0.349**	**0.331**	**0.309**	**0.263**	**0.311**
SD	0.012	0.024	0.010	0.015	0.013	0.013	0.015	0.016	0.023	0.016
P	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
III, Using the expression profiles of 18 randomly-selected unknown cotton genes (negative control):										
r	**0.034**	**0.044**	**0.111**	**0.142**	**0.035**	**0.028**	**0.042**	**0.031**	**0.113**	**0.065**
SD	0.020	0.045	0.042	0.102	0.029	0.019	0.018	0.020	0.039	0.037
P	0.624	0.529	0.116	0.044	0.624	0.695	0.556	0.654	0.112	0.439

FIGURE 2 | Ability of the *GFL* genes to predict the phenotype of fiber length using nine prediction models. **(A)** Ability of the *GFL* genes to predict the phenotype of fiber length using 18 *GFL* genes randomly selected from the 474 *GFL* genes. *r*, prediction accuracy presented by Pearson's correlation coefficient between predicted and observed fiber lengths; SD, standard deviation for 100 replications. **(B)** Statistics of prediction accuracies between these three sets of genes described in **(A)** for fiber length using the Tukey's HSD. I, 18 randomly-selected *GFL* genes. II, 18 published cotton fiber length genes (**Supplementary Tables S2**, **S3**); III, 18 randomly selected unknown cotton non-474 *GFL* genes. Different letters, significant at a confidence interval (CI) ≥ 95%; error bar, standard deviation for 100 replications. GBLUP, genomic best linear unbiased prediction; LASSO, least absolute shrinkage and selection operator; PLS, partial least square; SVMRBF, support vector machine using the radial basis function kernel; SVMPOLY, support vector machine using the polynomial kernel function; RF, random forest; RKHS, reproducing kernel Hilbert space regression (RKHS).

($P = 0.000$). The randomly selected unknown cotton genes could not predict the fiber length ($r = 0.028–0.142$, $P > 0.05$ for all nine prediction models, except for LASSO that had $P = 0.044$) (**Figure 2A**). Tukey's HSD test showed that the *GFL* genes had a similar prediction ability of fiber length to the published fiber length genes for five of the nine prediction models tested (confidence interval, CI < 95%), a higher prediction ability of fiber length than the published fiber length genes for three of the models, BayesA, BayesB, and SVMPOLY (CI ≥ 95%), and a lower prediction ability of fiber length than the published fiber length genes for only one of the nine models, RF (CI ≥ 95%). Both the *GFL* genes and the published fiber length genes had significantly higher prediction abilities than the randomly selected unknown genes for all nine prediction models (**Figure 2B**). These results indicated that the *GFL* genes had similar or better abilities to predict the fiber length than the published fiber length genes, thus verifying the contributions of the *GFL* genes to fiber length and their utility and efficiency to predict the phenotype of the objective trait.

Then, we further confirmed the ability of the *GFL* genes to predict the fiber length using a series of numbers of the randomly selected *GFL* genes sampled by bootstrap sampling, from 6 to all 474 (**Figure 3** and **Supplementary Table S5**). The experiment had ten bootstrap selections for each number of genes. As expected, all sets of the randomly selected *GFL* genes tested, no matter how many *GFL* genes there were in the selection, from 6 to 474, and which of the prediction models was used, were able to predict the fiber length ($P = 0.010$ for 6 *GFL* genes and $P = 0.000$ for all selections of genes with a number of *GFL* genes greater than 6). Again, none of the randomly selected unknown gene selections, regardless of how many there were in the selection, from 6 to 474, and which of the nine prediction models was used, could predict the fiber length ($P = 0.091–0.505$) (**Figure 3A** and **Supplementary Table S5**). Furthermore, as the number of the *GFL* genes used for the prediction increased, the prediction accuracy of fiber length increased (**Figures 3A,B**). When 200 or more of the *GFL* genes were used, the prediction accuracy plateaued (**Figure 3C**). Comparative analysis showed that the prediction models, PLS, BayesA, and RKHS, best predicted the phenotype of fiber length among the nine prediction models tested, with a prediction accuracy of $r = 0.830$, 0.817, and 0.814, respectively, when all 474 *GFL* genes were used (**Figure 3B** and **Supplementary Table S5**). In contrast, the prediction accuracies of the randomly-selected unknown gene sets remained non-significant, low, and consistent, for all of the randomly-selected cotton unknown gene selections, from 6 to 474 (**Figure 3A** and **Supplementary Table S5**). These results further confirmed the ability, utility, and efficiency of the *GFL* genes for accurate prediction of fiber length.

Prediction of Fiber Length Using the SNPs/InDels of the *GFL* Genes as DNA Markers

Moreover, we further tested the ability, utility, and efficiency of the *GFL* genes in predicting the phenotype of fiber length using the 226 SNP/InDel-containing *GFL* genes (**Supplementary Table S1C**). The SNPs or InDels contained in the 226 SNP/InDel-containing *GFL* genes were only used as DNA markers (**Supplementary Tables S6**, **S7**), as those DNA markers used for GS, with no effect of the *GFL* genes on fiber length considered, for the prediction. We first compared the prediction accuracy of fiber length using all 740 SNPs/InDels contained in the 226 *GFL* genes (**Supplementary Table S6**) and a selection of the 740 genic SNPs/InDels, with only one SNP/InDel that had the largest effect on fiber length per *GFL* gene (**Supplementary Table S7**). As expected, the 740 *GFL* SNPs/InDels better predicted the phenotype of fiber length, with a prediction accuracy varying from $r = 0.650$ ($P = 0.000$) for the RF model to $r = 0.832$ ($P = 0.000$) for the SVMRBF model, than the selection of the 226 *GFL* SNPs/InDels, with a prediction accuracy varying from $r = 0.671$ ($P = 0.000$) for the SVMPOLY model to $r = 0.779$ ($P = 0.000$) for the BaysA, BayesB, GBLUP, or RKHS

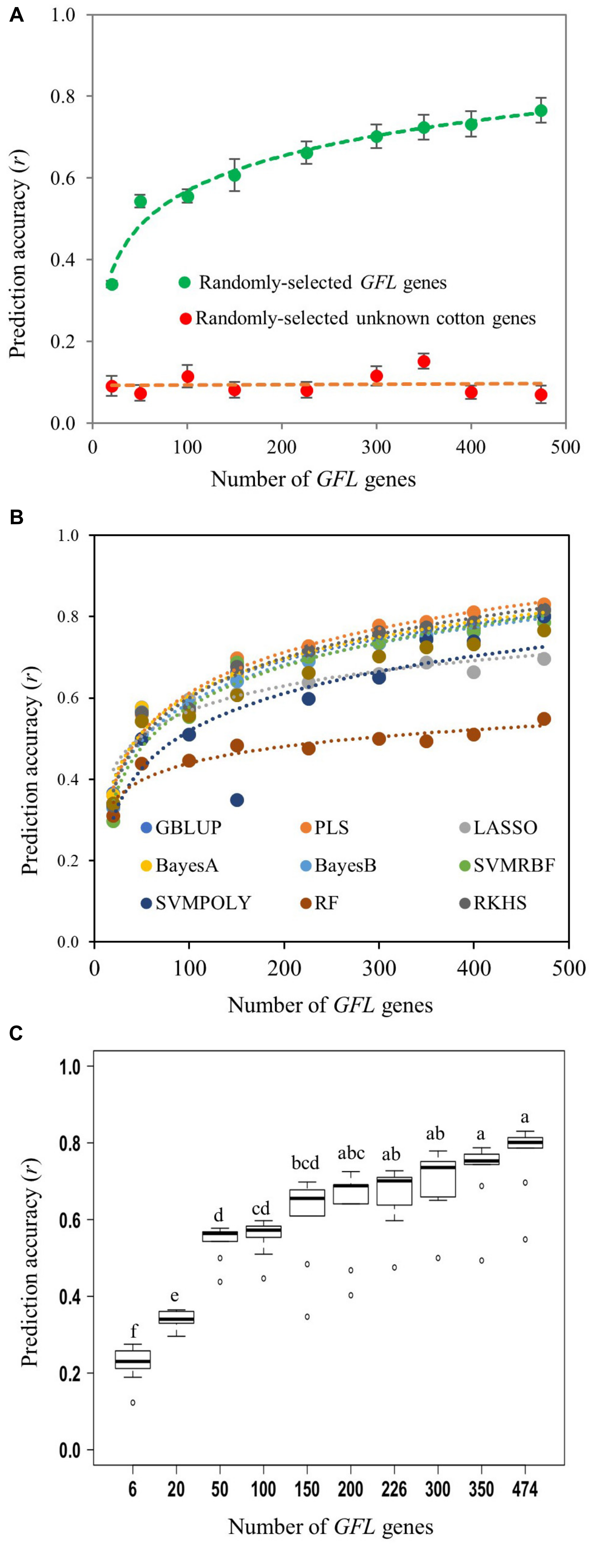


FIGURE 3 | Continued

FIGURE 3 | Prediction of fiber length with different numbers of randomly selected *GFL* genes and nine prediction models using expression profiles. **(A)** Mean prediction accuracy of fiber length with the *GFL* genes using the nine prediction models. A series of numbers of the 474 *GFL* genes ranging from 6 to 474 were tested using the same numbers of randomly selected unknown cotton non-474 *GFL* genes as the negative control (**Supplementary Table S5**). For prediction models, see **Figure 2**. **(B)** Prediction accuracy of fiber length with the *GFL* genes using different prediction models. **(C)** Statistics of the mean prediction accuracies between different numbers of the *GFL* genes predicted by the nine prediction models using the Tukey's HSD. Different letters, significant at CI ≥ 95%; same letter, not significant at CI ≥ 95%.

model, in seven of the nine prediction models. The 740 *GFL* SNPs/InDels had a similar to or lower prediction accuracy than the selection of the 226 *GFL* SNPs/InDels for the LASSO and RF models (**Figure 4A**).

However, if the selection of 226 SNPs/InDels was used for the prediction, although the prediction accuracy would be slightly lower, the cost of genotyping for the prediction would be reduced by 2.3-fold. Therefore, we further tested the prediction accuracies of different numbers of the SNPs/InDels selected from the 226 *GFL* SNPs/InDels for the phenotype of fiber length. Overall, the RKHS model showed the best prediction results of fiber length among the nine models (**Figure 4B**), and as more of the 226 *GFL* SNPs/InDels were used, a more accurate prediction of fiber length was obtained (**Figure 4C**). The fiber lengths of the cotton lines were predicted at an accuracy of $r = 0.783$ ($P = 0.000$), when all the 266 *GFL* SNPs/InDels were used with the RKHS model.

In comparison, the prediction accuracies of fiber length using all 740 SNPs/InDels contained in 226 *GFL* genes were essentially the same high as the prediction accuracies of fiber length using the expressions of all 474 *GFL* genes, thus demonstrating the ability, utility and efficiency of the *GFL* genes in phenotype prediction of fiber length for progeny selection.

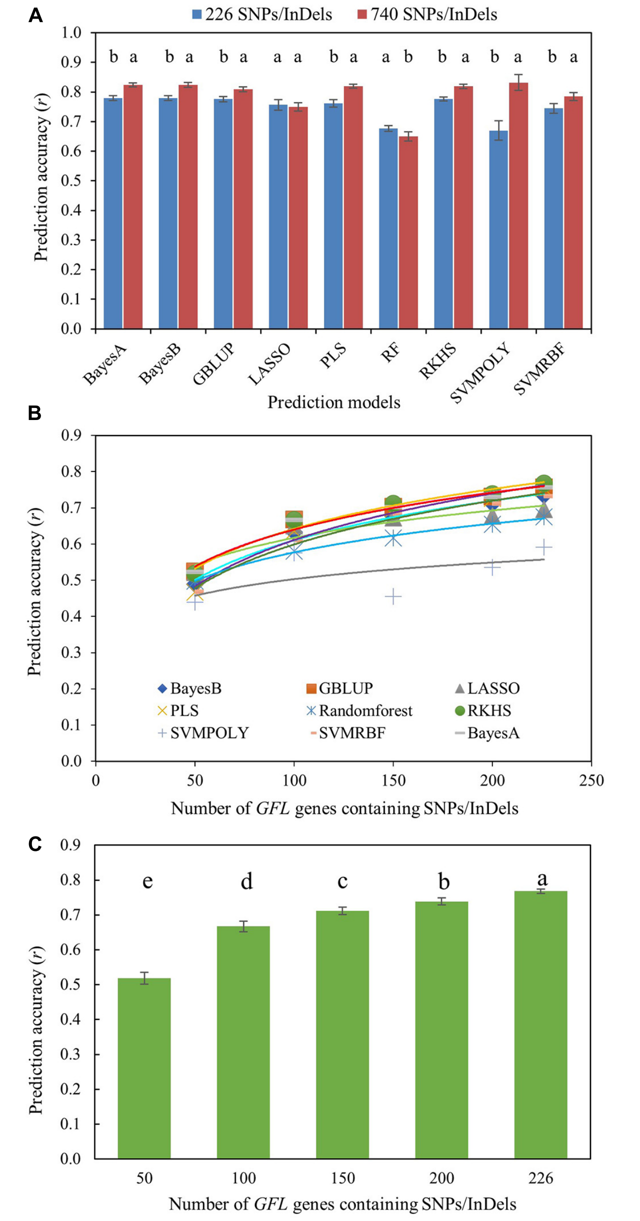


FIGURE 4 | Prediction of fiber length with the *GFL* SNPs/InDels as DNA markers using nine prediction models. **(A)** Prediction accuracy of fiber length using the genotypes of all 740 *GFL* SNPs/InDels (**Supplementary Table S6**) versus a selection of 226 *GFL* SNPs/InDels that had the largest effects on fiber length, with only one SNP/InDel per gene (**Supplementary Table S7**). Different letters, significant at CI ≥ 95%; same letter, not significant at CI ≥ 95%; error bar, standard deviation for 100 replications. **(B)** Prediction accuracy of fiber length with the selection of the 226 *GFL* SNPs/InDels (**Supplementary Table S7**) using different prediction models. **(C)** Prediction of fiber length with different numbers of the 226 *GFL* SNPs/InDels (**Supplementary Table S7**) using the RKHS model. Different letters, significant at CI ≥ 95%; error bar, standard deviation.

Identification of the Key *GFL* Genes to Phenotype Prediction of Fiber Length for Progeny Selection

The above experiments indicated that the *GFL* genes were able to accurately predict the fiber length with either *GFL* expression abundances in 10-dpa developing fibers or *GFL* genic SNPs/InDels as DNA markers. The question was whether the *GFL* genes equally contributed to the phenotype prediction of fiber length. If not, whether a subset of the *GFL* genes, defined herein the key *GFL* genes, selected from the 474 *GFL* genes could predict the phenotype of fiber length as accurate as all 474 *GFL* genes for progeny selection. Therefore, we tested the ability and efficiency of the *GFL* genes according to their roles in the *GFL* network (Liu, 2014; **Supplementary Table S1B**), the effects of their SNP/InDel mutations on fiber length (**Supplementary Table S1C**), or their effects on fiber length (Liu, 2014; **Supplementary Table S1A**). The *GFL* genes randomly selected from the 474 *GFL* genes were used as the control. The expression abundances of the selected *GFL* genes were used for the prediction. Results showed that both the roles of the *GFL* genes in the *GFL* network (**Supplementary Figure S2A**) and their effects on fiber length (**Supplementary Figure S2C**) increased the ability of the genes to predict fiber length, but

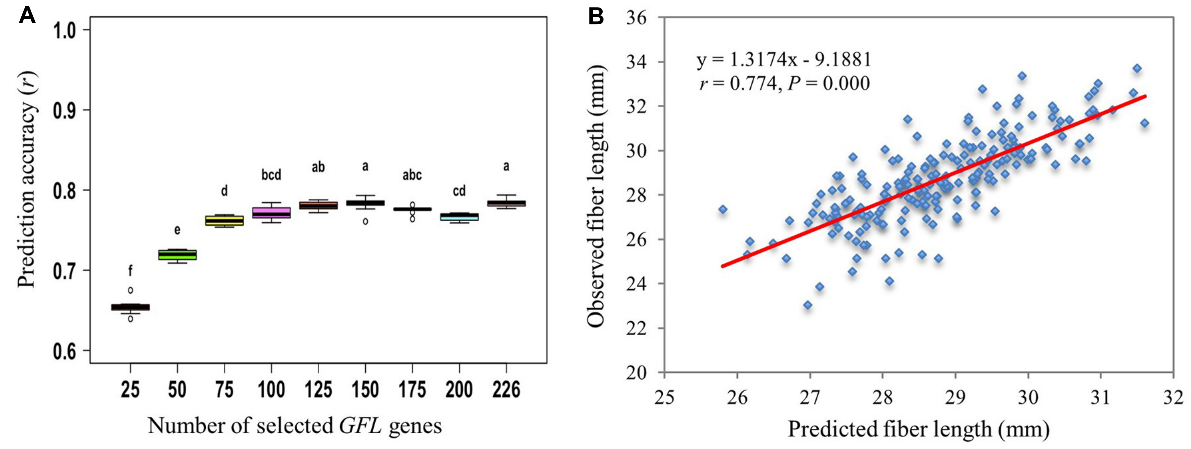


FIGURE 5 | Prediction of fiber length using the 226 *GFL* genes selected according to their effects on fiber length (Subset X, **Supplementary Figure S2C**). **(A)** Prediction of fiber length using different numbers of the 226 selected *GFL* genes and the SVMRBF model. Different letters, significant at CI ≥ 95%; same letter, not significant at CI ≥ 95%; error bar, standard deviation for 100 replications. **(B)** Prediction of fiber length with the 125 *GFL* genes selected from the 226 *GFL* genes (**Supplementary Table S8**) using the SVMRBF model.

the effects of SNP/InDel mutations of the *GFL* genes on fiber length (**Supplementary Figure S2B**) decreased the ability of the genes to predict fiber length (CI ≥ 95%). Since the effects of the *GFL* genes on fiber length had a larger increase than their roles in the *GFL* network for phenotype prediction of fiber length, the subset of the 226 *GFL* genes consisting of all 54 positively effective *GFL* genes, 59 smallest negatively effective *GFL* genes, and 113 largest negatively effective *GFL* genes (Subset X, **Supplementary Figure S2C**) was selected for further analysis (**Supplementary Table S1A**).

Furthermore, we predicted the phenotype of fiber length using different numbers of *GFL* genes randomly selected from the subset of 226 *GFL* genes above (Subset X, **Supplementary Figure S2C**). When 125 or more of the *GFL* gene subset were used, the prediction accuracy of fiber length plateaued for eight of the nine prediction models and the SVMRBF model best predicted the phenotype of fiber length using these numbers of the selected *GFL* genes (**Figure 5A** and **Supplementary Figure S3**). Therefore, a subset of 125 *GFL* genes were identified from the 226 selected *GFL* genes for phenotype prediction of fiber length using expression profiles in 10-dpa developing fibers (**Supplementary Table S8**). These 125 *GFL* genes were herein defined the key *GFL* genes to phenotype prediction of fiber length for progeny selection. When the 125 key *GFL* genes were used, the prediction accuracy of fiber length approached r = 0.774 (P = 0.000) (**Figure 5B**), suggesting that they were well suited for accurate prediction of fiber length and therefore, could be used for progeny selection in a breeding program. Comparative analysis showed that the prediction results of these 125 key *GFL* genes were significantly correlated with those predicted with all 474 *GFL* genes (r = 0.888, P = 0.000; **Supplementary Figure S4**). The fiber lengths predicted with the expression of the 125 key *GFL* genes were also significantly correlated with those predicted using the 226 SNPs/InDels contained in the 226 *GFL* genes (r = 0.892, P = 0.000).

Prediction of Fiber Length Using the *GFL* Genes Across Years or Generations

To further explore the ability, utility, and efficiency of the *GFL* genes for fiber length prediction, we examined the prediction accuracy of fiber length for the RILs across years or environments (generations) using the two datasets of the selected *GFL* genes genotyped from the 2011 (F_9) trial only and the fiber lengths phenotyped in 2009 (F_7), 2010 (F_8), and 2011 (F_9), respectively. The result showed that the *GFL* genes genotyped in the 2011 (F_9) trial could also predict the fiber length of the RILs grown in 2010 (F_8) at a prediction accuracy similar to that achieved from the 2011 trial that was used for genotyping the genes using either of the two genic datasets, 125 key *GFL* expressions or 226 *GFL* SNPs/InDels as DNA markers. However, the prediction accuracy of fiber length for the RILs grown in 2009 (F_7) was slightly lower than those achieved for the RILs in 2010 and 2011 (**Table 1**). Since the 2009 trial had no replication (those of 2010 and 2011 had three replications) and the prediction accuracy was determined by Pearson's correlation coefficient between the predicted and observed phenotypes, the reduced prediction accuracy for 2009 could be more likely attributed to the fiber length phenotyping accuracy rather than the gene x environment interactions. These results confirmed that the prediction accuracy of fiber length for different environments or years and suggested that the prediction accuracy of fiber length using the *GFL* genes was largely consistent across environments or years at the late generations of progeny for plant breeding.

The Proper Training Population Size for Accurate Prediction of Fiber Length Using the *GFL* Genes

Furthermore, we determined what was the appropriate training population size to train a prediction model for fiber length prediction using the *GFL* genes by using their expression

TABLE 1 | Prediction accuracies of fiber length for different generations or years using the two datasets of the selected *GFL* genes for GBB collected in 2011, individually: **(A)** The RKHS model was used for the prediction and **(B)** The SVMRBF model was used for the prediction.

Year	Generation	(A) 226 *GFL* SNPs/InDels as markers		(B) Expression of 125 selected *GFL* genes	
		r	*P*-value	*r*	*P*-value
2011	F_9	0.7830	0.00E + 00	0.7872	0.00E + 00
2010	F_8	0.8334	0.00E + 00	0.7761	0.00E + 00
2009	F_7	0.6719	0.00E + 00	0.6515	0.00E + 00

The observed fiber lengths measured in 2010 or 2011 were the means of three replicates, while the observed fiber length measured in 2009 was from only one five-plant plot with no replicate, largely explaining the lower prediction accuracy of fiber length in 2009.

abundances (**Figure 6A**) and their SNPs/InDels as DNA markers (**Figure 6B**), individually. This is because the training population size is regarded to prediction accuracy and also to the cost for prediction model training. The populations consisting of a series of numbers of lines, from 40 to 198, were used to predict the fiber length using the selected optimal prediction models. Although the variation of the prediction accuracy increased as the training population size decreased, the prediction accuracy of the *GFL* genes for fiber length plateaued, when 100 lines were used, with the expressions of the 125 key *GFL* genes (**Figure 6A**). For prediction of fiber length using the 226 SNPs/InDels of the 226 SNP/InDel-containing *GFL* genes as DNA markers, the prediction accuracy of fiber length plateaued, when 120 lines were used (**Figure 6B**). Therefore, a training population size of 100–120 lines seemed proper to train a prediction model for accurate prediction of fiber length for progeny selection using either genotypes or expressions of the *GFL* genes.

DISCUSSION

One of the most important aims of molecular and genomic research is to develop molecular technologies that can enhance breeding in crop plants and livestock, and enhance medicine in humans. This study has demonstrated that the phenotype of a quantitative trait can be accurately predicted using the genes controlling the trait. The prediction accuracy of the cotton fiber length, which is used as the objective trait in this study, has approached its plateaued accuracy, with an accuracy of $r = 0.83$ ($P = 0.000$) using either the SNPs/InDels of 226 of the 474 *GFL* genes or the expressions of the 474 *GFL* genes. This prediction accuracy is as accurate as the prediction accuracy of maize grain yield ($r = 0.85$, $P = 0.000$), which is one of the most complex quantitative traits, using the maize grain yield (*ZmINGY*) genes (Zhang et al., 2020a). Moreover, the cotton fiber lengths predicted using these two genic datasets of the *GFL* genes are significantly correlated ($r = 0.892$, $P = 0.000$), further verifying the prediction accuracy of fiber length. The prediction accuracy of fiber length achieved using its contributing genes are 4%–315%, with an average of 95%, higher than those of $r = 0.20$–0.80 achieved for different quantitative traits using genome-wide DNA markers, genome-wide gene expressions, or genome-wide metabolites consisting of thousands to tens of thousands of omic features (Meuwissen et al., 2001; Lee et al., 2008; Crossa et al., 2010, 2013; De Los Campos et al., 2010b; Heffner et al., 2011a,b; Daetwyler et al., 2012; González-Camacho et al., 2012; Gouy et al., 2013; Morota et al., 2014; Speed and Balding, 2014; Xu et al., 2014, 2016; Beyene et al., 2015; Dan et al., 2016; Weissbrod et al., 2016; Islam et al., 2020). If the same species (cotton), same trait (fiber length, UHML), same prediction models (BayesB, GBLUP and RKHS), and same cross-validation scheme are considered for the comparison, the prediction accuracy of the cotton fiber length using the 740 SNPs/InDels of the 226 *GFL* genes as DNA markers were $r = 0.80$, 0.80, and 0.82 ($P = 0.000$) for GBLUP, BayesB, and RKHS, respectively, in this study (**Figure 4A**). These prediction accuracies are 116% higher than those of the fiber length predicted using 6,292 genome-wide SNPs (Islam et al., 2020). Furthermore, the prediction accuracy of cotton fiber length using the *GFL* genes is highly consistent across years (environments), even though the weathers between the years were quite different, with 2011 having unusual weather. This result is consistent with that of Zhang et al. (2020a) who showed that the genes controlling maize grain yields consistently predicted the maize grain yield across diverse climates and across different eco-agricultural systems. Finally, 100–120 plants are sufficient to properly train a model for accurate prediction of fiber length using the *GFL* genes, thus significantly reducing the cost for training and validating a model for phenotype prediction of a quantitative trait (Islam et al., 2020). These results, therefore, indicate that the genes controlling a quantitative trait are capable of and desirable for accurate prediction of the phenotype of a quantitative trait for progeny selection.

Zhang et al. (2020a) first proposed gene-based breeding (GBB), based on the ability, utility, and efficiency of the maize grain yield genes for accurate prediction of maize grain yield. GBB is an innovative plant breeding method that makes full use of the genes controlling the objective trait(s) through the entire process of plant breeding, including parent selection, cross design, and progeny selection. Three genic datasets of the genes are used for GBB individually or jointly: (i) the number of their favorable alleles (NFAs), (ii) their SNPs/InDels as DNA markers, and (iii) their expression abundances and networks. The results of this study that used two of the genic datasets for GBB provide a strong support for development and application of GBB for enhanced and accelerated plant breeding. Because the datasets of genes controlling the objective trait(s) are used for the entire breeding process, GBB allows not only accurately selecting for the progeny that are the most high-yielding, high-quality and highly resistant to biotic and abiotic stresses, but also accurately selecting the most desirable breeding materials or

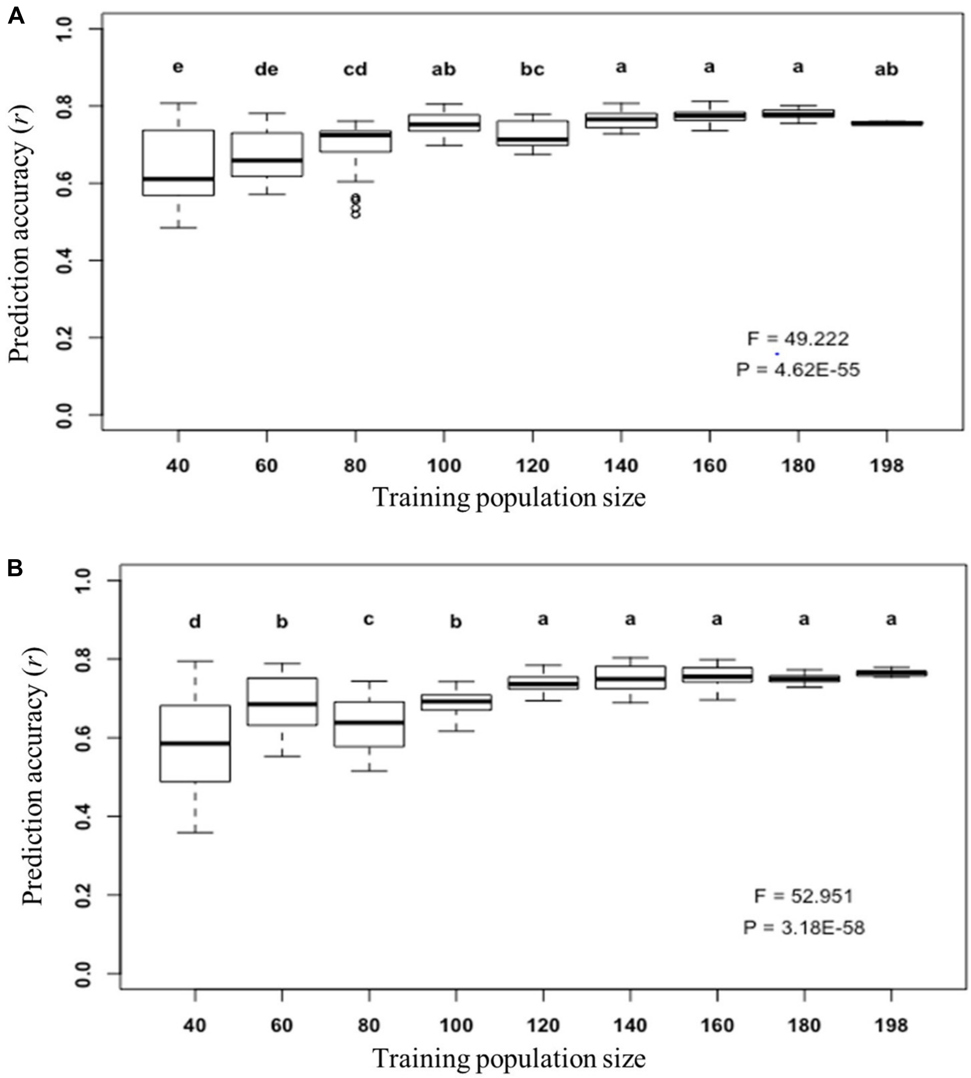


FIGURE 6 | Prediction of fiber length with the selected *GFL* genes for GBB using different training population sizes. **(A)** Prediction of fiber length using the transcript expression abundances of the 125 selected *GFL* genes and the SVMRBF model (**Figures 5A,B**). **(B)** Prediction of fiber length using the 226 selected *GFL* SNPs/InDels as DNA markers and the RKHS model (**Figure 4C**). The prediction was carried out for 100 replications. Each number of lines was sampled for 10 times by bootstrap sampling, with each number sample being tested with 10 replications. Different letters, significant at CI ≥ 95%; same letter, not significant at CI ≥ 95%; error bar, standard deviation.

parents to approach the breeding objectives and wisely designing crosses that maximally combine the favorable alleles and heterotic genotypes of the genes controlling the objective trait(s) from the breeding materials into progeny. Therefore, GBB sheds great light on substantial and continued crop improvement, thus promising to help feed the world.

The findings of this study are achieved using cotton fiber length as the objective trait; nevertheless, the concepts and methods developed in this study are applicable to accurate prediction of other quantitative traits in crop plants, livestock, and humans, to development of GBB for enhanced crop and livestock improvement, and to development of gene-based medicine for enhanced human disease prevention, diagnosis and medicine. This conclusion is supported not only by the results of this study, but also by Zhang et al. (2020a) who accurately predicted the phenotype of grain yield in maize within and across diverse environments (locations). However, concerns may exist for practical use of the trait contributing genes in phenotype prediction of quantitative traits. The first concern may be genome-wide high-throughput cloning of the genes controlling an objective quantitative trait. We previously invented an innovative technology and developed an associated pipeline for genome-wide high-throughput cloning of the genes controlling quantitative traits and used it to have successfully cloned the 1,501 *ZmINGY* genes used by Zhang et al. (2020a) and the 474 *GFL* genes used for this study. Both the accurate prediction of cotton fiber length using the *GFL* genes (this study) and the accurate prediction of maize grain yield using the *ZmINGY* genes

(Zhang et al., 2020a) consistently indicated that our novel gene cloning technology enables to genome-wide, high-throughput, and reliably clone the genes controlling quantitative traits. Because its gene cloning throughput, efficiency, and reliability are independent of the genome size, complexity, ploidy level, and availability of genomic knowledge and resources of a species, our gene cloning technology is applicable to genome-wide high-throughput cloning of genes controlling a quantitative trait in any species, including plants, animals, humans, and microbes. This technology and associated pipeline will be published and made available to the public soon.

The second concern may be variation of gene expression across environments. First, gene expression is the determinant of phenotype of a trait that results from interaction of numerous factors, including gene effects (additive and dominant), gene mutation, gene x gene interaction (epistasis), gene x genetic background or non-gene element interaction, epigenetic factors, and G x E interaction; therefore, it is a desirable type of omics for omics-based prediction of phenotypes. This study and Zhang et al. (2020a, b) revealed that the variation of a quantitative trait, such as cotton fiber length, maize grain yield, and ginseng ginsenoside content (Zhang et al., 2020b), is contributed by not only gene mutation, such as SNPs/InDels, but also by variation of gene expression. Therefore, the expression abundances of genes controlling the objective quantitative trait accurately predicted the phenotype of the fiber length in this study and the phenotype of the maize grain yield by Zhang et al. (2020a). Moreover, Zhang et al. (2019) conducted an extensive study on the variation of gene expression across environments and showed that that gene transcript expressions were highly consistent and highly reproducible across plants growing within a field trial replicate, between field trial replicates, and sampled from different years/locations (r = 0.90–0.98, P = 0.000). In addition, we recently showed that the phenotypic performance of offspring could be also accurately predicted using the expression abundances of genes related to the objective trait (grain yield) in parents in maize across very diverse climates, across eco-agricultural systems, and across populations (MZ, Y-HL, Y Wang, CF Scheuring, X Qi, J Pekar, SC Murray, W Xu, S-HS, H-BZ, submitted). These results together consistently indicate that the expression abundances of the genes contributing to the objective trait could predict the phenotype of the trait across environments, including different years, different climates, and different eco-agricultural systems, and across populations.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article or Supplementary Material. The sequences of the 474 *GFL* genes can be found in NCBI GenBank under accession numbers: MW082098-MW082571.

AUTHOR CONTRIBUTIONS

H-BZ conceived, designed, and supervised the entire project. SX supervised the prediction of fiber length with the *GFL* genes using the nine prediction models. Y-HL performed the experiments and data analysis. YX and YC performed the fiber length prediction with the *GFL* genes using the prediction models. MZ helped with the data analysis and prepared the manuscript. CWS developed the RIL population, helped conduct the field trials, and phenotyped the fiber length. S-HS genotyped the SNPs or InDels of the *GFL* genes in the cotton population and parents. All authors contributed to the article and approved the submitted version.

FUNDING

This project was supported in part by the Agriculture and Food Research Initiative Competitive grant no. 2013-67013-21109 of the USDA National Institute of Food and Agriculture, the Texas A&M AgriLife Research Cotton Improvement Program (Grant Nos: 06-124475-85360 and 06-504755-85360) and the National Science Foundation Collaborative Research grant (DBI-1458515). The open access publishing fees for this article have been partially covered by the Texas A&M University Open Access to Knowledge Fund (OAKFund), supported by the University Libraries.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.583277/full#supplementary-material

Supplementary Figure 1 | Examples of validation of cotton *GFL* SNPs by allele-specific PCR.

Supplementary Figure 2 | Selection of key *GFL* genes for GBB.

Supplementary Figure 3 | Prediction of fiber length using different numbers of the effect-selected *GFL* genes with nine prediction models.

Supplementary Figure 4 | Correlation of predicted fiber lengths between the 125 selected *GFL* genes and all 474 *GFL* genes.

Supplementary Table 1 | Selection of the key *GFL* genes for GBB, according to their effects on fiber length (A),their roles in the *GFL* network (B), or the effects of their SNP/InDel mutations on fiber length (C).

Supplementary Table 2 | Published cotton fiber length genes cloned by the traditional gene cloning methods and used as the positive control in this study.

Supplementary Table 3 | The transcript sequences of the published cotton fiber length genes used as the positive control in this study.

Supplementary Table 4 | Expression profile variation of the 474 *GFL* genes, presented in TPM (transcripts per million), in 10-dpa developing fibers of the cotton RIL population.

Supplementary Table 5 | Prediction accuracy of fiber length with different numbers of randomly-selected *GFL* genes and randomly-selected unknown non-474 *GFL* cotton genes using nine prediction models.

Supplementary Table 6 | Genotypes of all 740 SNPs/InDels contained in 226 *GFL* genes for prediction of fiber length.

Supplementary Table 7 | Genotypes of 226 SNPs/InDels contained in 226 *GFL* genes, with only one SNP or InDel per gene, for prediction of fiber length.

Supplementary Table 8 | The 125 key *GFL* genes selected for GBB.

REFERENCES

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

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 (Bethesda)* 3, 1903–1926. doi: 10.1534/g3.113.008227

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

Daetwyler, H. D., Swan, A. A., Werf, J. H. J., and van der Hayes, B. J. (2012). Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genet. Sel. Evol. GSE* 44:33.

Dan, Z., Hu, J., Zhou, W., Yao, G., Zhu, R., Zhu, Y., et al. (2016). Metabolic prediction of important agronomic traits in hybrid rice (*Oryza sativa* L.). *Sci. Rep.* 6:21732.

De Los Campos, G., Gianola, D., and Allison, D. B. (2010a). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Genet. Rev.* 11, 880–886. doi: 10.1038/nrg2898

De Los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., and Crossa, J. (2010b). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* 92, 295–308. doi: 10.1017/s0016672310000285

Desta, Z. A., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* 19, 592–601. doi: 10.1016/j.tplants.2014.05.006

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22.

Gaudet, M., Fara, A.-G., Sabatti, M., Kuzminsky, E., and Mugnozza, G. S. (2007). Single-reaction for SNP genotyping on agarose gel by allele-specific PCR in black poplar (*Populus nigra* L.). *Plant Mol. Biol. Rep.* 25, 1–9. doi: 10.1007/s11105-007-0003-6

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Anal. Chim Acta* 185, 1–17. doi: 10.1016/0003-2670(86)80028-9

González-Camacho, J. M., De Los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125, 759–771. doi: 10.1007/s00122-012-1868-9

González-Recio, O., and Forni, S. (2011). Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genet. Sel. Evol.* 43:7.

Gouy, M., Rousselle, Y., Bastianelli, D., Lecomte, P., Bonnal, L., Efile, J.-C., et al. (2013). Experimental assessment of the accuracy of genomic selection in sugarcane. *Theor. Appl. Genet.* 126, 2575–2586. doi: 10.1007/s00122-013-2156-z

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Graur, D., and Li, W.-H. (2000). *Fundamentals of Molecular Evolution*, 2nd Edn. Sunderland, MA: Sinauer Associates, Inc.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Heffner, E. L., Jannink, J.-L., Iwata, H., Souza, E., and Sorrells, M. E. (2011a). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop. Sci.* 51, 2597–2606. doi: 10.2135/cropsci2011.05.0253

Heffner, E. L., Jannink, J.-L., and Sorrells, M. E. (2011b). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* 4, 65–75. doi: 10.3835/plantgenome.2010.12.0029

Islam, M. S., Fang, D. D., Jenkins, J. N., Guo, J., McCarty, J. C., and Jones, D. C. (2020). Evaluation of genomic selection methods for predicting fiber quality traits in Upland cotton. *Mol. Genet. Genomics* 295, 67–79. doi: 10.1007/s00438-019-01599-z

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – An S4 package for kernel methods in R. *J. Stat. Softw.* 11, 1–20.

Khan, A. A., Azhar, F. M., Khan, I. A., Raiz, A. H., and Athar, M. (2010). Genetics basis of variation for lint color, yield, and quality in cotton (*Gossypium hirsutum* L.). *Plant Biosyst.* 143, S17–S24.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7, 673–679. doi: 10.1038/89044

Lee, S. H., Werf, J. H. J., van der Hayes, B. J., Goddard, M. E., and Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* 4:e1000231. doi: 10.1371/journal.pgen.1000231

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liaw, A., and Wiener, M. (2018). *Breiman and Cutler's Random Forests for Classification and Regression. CRAN*. Available from: https://www.stat.berkeley.edu/~breiman/RandomForests/

Liu, B. H. (1997). *Statistical Genomics: Linkage, Mapping and QTL Analysis*. Boca Raton, FL: CRC Press.

Liu, Y.-H. (2014). *Molecular Basis Of Quantitative Genetics Revealed By Cloning And Analysis Of 474 Genes Controlling Fiber Length In Cotton*. Ph.D. Dissertation, Texas A&M University, College Station, TX.

Maenhout, S., De Baets, B., Haesaert, G., and van Bockstaele, E. (2007). Support vector machine regression for the prediction of maize hybrid performance. *Theor. Appl. Genet.* 115, 1003–1013. doi: 10.1007/s00122-007-0627-9

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.

Mevik, B.-H., and Wehrens, R. (2007). The pls Package: principal component and partial least squares regression in R. *J. Stat. Softw.* 18, 1–24.

Morota, G., Abdollahi-Arpanahi, R., Kranis, A., and Gianola, D. (2014). Genome-enabled prediction of quantitative traits in chickens using genomic annotation. *BMC Genomics* 15:109. doi: 10.1186/1471-2164-15-109

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

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

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., and Sheridan, R. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958. doi: 10.1021/ci034160g

Syed, N. H., Kalyna, M., Marquez, Y., Barta, A., and Brown, J. W. S. (2012). Alternative splicing in plants – coming of age. *Trends Plant Sci.* 17, 616–623. doi: 10.1016/j.tplants.2012.06.001

Takagi, Y., Matsuda, H., Taniguchi, Y., and Iwaisaki, H. (2014). Predicting the phenotypic values of physiological traits using SNP genotype and gene expression data in mice. *PLoS One* 9:e115532. doi: 10.1371/journal.pone.0115532

Tibshirani, R. (1996). Regression Shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Ulloa, M. (2006). Heritability and correlations of agronomic and fiber traits in an okra-leaf upland cotton population. *Crop. Sci.* 46, 1505–1514. doi: 10.2135/cropsci2005.08-0271

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

Weissbrod, O., Geiger, D., and Rosset, S. (2016). Multikernel linear mixed models for complex phenotype prediction. *Genome Res.* 26, 969–979. doi: 10.1101/gr.201996.115

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

Xu, S., Zhu, D., and Zhang, Q. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12456–12461. doi: 10.1073/pnas.1413750111

Zhang, M. P., Cui, Y., Liu, Y.-H., Xu, W., Sze, S.-H., Murray, S. C., et al. (2020a). Accurate prediction of maize grain yield using its contributing genes for gene-based breeding. *Genomics* 112, 225–236. doi: 10.1016/j.ygeno.2019.02.001

Zhang, M. P., Liu, Y.-H., Chang, C.-S., Zhi, H., Wang, S., Xu, W., et al. (2019). Quantification of gene expression while taking into account RNA alternative splicing. *Genomics* 111, 1517–1528. doi: 10.1016/j.ygeno.2018.10.009

Zhang, M. P., Liu, Y.-H., Xu, W., Smith, C. W., Murray, S. C., and Zhang, H.-B. (2020b). Analysis of the genes controlling three quantitative traits in three diverse plant species reveals the molecular basis of quantitative traits. *Sci. Rep.* 10:10074.

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537.

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 Plant Science

OPINION
published: 16 April 2021
doi: 10.3389/fpls.2021.651480

The Modern Plant Breeding Triangle: Optimizing the Use of Genomics, Phenomics, and Enviromics Data

Jose Crossa[1,2], *Roberto Fritsche-Neto*[3], *Osval A. Montesinos-Lopez*[4], *Germano Costa-Neto*[3], *Susanne Dreisigacker*[1], *Abelardo Montesinos-Lopez*[5] *and Alison R. Bentley*[1]*

[1] International Maize and Wheat Improvement Center (CIMMYT), Carretera México-Veracruz, de Mexico, Mexico, [2] Colegio de Postgraduados, Montecillo, Edo. de Mexico, Mexico, [3] Department of Genetics, "Luiz de Queiroz" Agriculture College, University of São Paulo, São Paulo, Brazil, [4] Facultad de Telemática, Universidad de Colima, Colima, Mexico, [5] Departamento de Matemáticas, Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, Guadalajara, Mexico

Keywords: genomics, phenomics, enviromics, high throughput phenotype, multi-trait, multi-environment

OPEN ACCESS

Edited by:
Brian Gardunia,
Bayer Crop Science, United States

Reviewed by:
Flavio Breseghello,
Brazilian Agricultural Research Corporation (EMBRAPA), Brazil
Hans D. Daetwyler,
La Trobe University, Australia
Cathy Colette Jubin,
University of Göttingen, Germany

****Correspondence:***
Alison R. Bentley
A.Bentley@cgiar.org

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *09 January 2021*
Accepted: *11 February 2021*
Published: *16 April 2021*

Citation:
Crossa J, Fritsche-Neto R, Montesinos-Lopez OA, Costa-Neto G, Dreisigacker S, Montesinos-Lopez A and Bentley AR (2021) The Modern Plant Breeding Triangle: Optimizing the Use of Genomics, Phenomics, and Enviromics Data. Front. Plant Sci. 12:651480. doi: 10.3389/fpls.2021.651480

INTRODUCTION

Continued increases in genetic gain demonstrate the success of established public and private plant breeding programs. Nevertheless, in the last two decades, a growing body of modern technologies has been developed and now awaits efficient integration into traditional breeding pipelines. This integration offers attractive benefits, yet comes with the challenges of making modifications in established and operational systems, a recent example of which is rice breeding (Collard et al., 2019). Newly available technologies, genomics rapid cycling (Crossa et al., 2017), high throughput phenotyping (HTP, phenomics) (Montesinos-López et al., 2017) and historical descriptions of environmental relatedness (enviromics) (Costa-Neto et al., 2020a,b; Resende et al., 2020; Rogers et al., 2021) are crucial to improving conventional breeding schemes and increasing genetic gain. Integrating these new technologies into routine breeding pipelines will support the delivery of cultivars with robust yields in the face of the expected unfavorable future environmental conditions caused by climate change and the consequently increased occurrence of biotic and abiotic stresses. Here, we briefly describe the use of these technologies and their implementation to provide cost-effective and time-saving approaches to plant breeding. We also give an overview of the interconnections between these techniques. Finally, we envision future perspectives to implement a more interconnected breeding approach that takes advantage of the so-called modern plant breeding triangle: integrating genomics, phenomics, and enviromics.

Why Genomics for Improving Breeding?

One of the most popular uses of genomics in breeding is the prediction of breeding values. Genomic selection (GS) reduces cycle time, increases the accuracy of estimated breeding values and improves selection accuracy. For instance, in maize, the effectiveness of GS has been proven for the case of bi-parental populations (Massman et al., 2013; Beyene et al., 2015; Vivek et al., 2017), as well as in multi-parental populations (Zhang et al., 2017). Its use has also been documented in species with long generation times such as trees (Grattapaglia et al., 2018) and dairy cattle breeding, where the reduction of the breeding cycle has increased the response to selection in comparison with the progeny testing system (García-Ruiz et al., 2016).

Genomic selection has been implemented in many crops, including wheat, chickpea, cassava and rice (Roorkiwal et al., 2016; Crossa et al., 2017; Wolfe et al., 2017; Huang et al., 2019), and the number of programs that are moving from "conventional" to GS is growing. Results in wheat show that genomic predictions used early in the breeding cycle led to a substantial increase in performance in later generations (Bonnett et al., 2021 this issue).

Defining Foundational Core Parents for Genomic Selection-Assisted Breeding

In genomic selection, the optimization of the training set composition is an important topic because training and testing sets should be genetically related in such a way that the genetic diversity present in the testing set could be covered and captured by the diversity in the training set. Breeding programs must start forming initial foundational core parents (training populations) that represent the genetic diversity found in the current progeny and conform to the testing population(s) to the greatest extent possible (Hickey et al., 2012). These foundational parents should be extensively phenotyped in different target populations of environments and genotyped with high-density marker systems. These training sets of foundation parents will be able to produce a model with a high accuracy for current highly selected progenies (Zhang et al., 2017).

Why Detailed Phenomics and the Use of Multi-Trait Analysis to Improve Breeding?

The most important limitation to determining accurate phenotypes has been the time and cost required to measure traits in the field. Field phenomics aims to study all plant phenotypes under a range of environmental conditions. Modern phenomics methods are able to use hyperspectral/multispectral cameras to provide hundreds of reflectance data points at discrete narrow bands in many environments and at many stages of crop development. Phenotyping technology can now be used to quickly and accurately obtain data on agronomic traits based on advancements in plant phenotyping technologies (Atkinson et al., 2018). Therefore, the main goal of a high-throughput phenotype (HTP) is to reduce the cost of data per plot and to increase the prediction accuracy early in the crop-growing season with the use of highly heritable secondary phenotypes, closely related to the selection phenotypes. The cost of processing HTP data can be minimized by using open-source software, such as FieldImageR (Matias et al., 2020).

There is evidence that multi-trait analyses improve prediction accuracies when the genetic and residual correlations are considered in the modeling process. New genomic models that take the multiple traits and the multiple environments into consideration, along with trait × environment, trait × genotype, and trait × genotype × environment interactions, offer a huge potential for the exploitation of correlations between different variables and for the differentiation between effects. Integrating current GBLUP multi-trait models with models that consider the environmental information with the two- and three-way interaction terms provides a powerful, unified, whole genome prediction model.

The Bayesian multi-trait and multi-environment model (BMTME) (Montesinos-López et al., 2016, 2019a) allows for general covariance matrices for traits and environments that capture the correlations among traits and environments better. This unified model could be implemented to select genotypes with traits measured in one environment and to predict in other, untested environments. It could also be applied to predict traits that are costly or difficult to measure in all environments.

It is crucial to obtain large and inter-operable phenomics datasets from field phenotyping. This should be used to characterize the foundational core parents in the different environments and incorporate them into the visual data collected in the different environments. These data, along with pedigree and genomic information, can be used to fit Bayesian linear mixed models to compute BLUPs of the genetic values of the material in the training set. Breeding programs should collect multi-trait data on the multi-environment used for foundational core parents and exploit possible correlations among traits that will eventually increase prediction accuracy. The genomics and phenomics of the multi-trait foundation core parents are essential for use alongside enviromics data.

Why Enviromics to Improve Multi-Environment Trials for Genomics-Assisted Plant Breeding?

The phenotypic variation observed across diverse environments is a product of genetic and environmental variation. Thus, enviromics acts as a central bottleneck for the application of modern genomics-assisted prediction tools, especially for use across multiple environments. Novel approaches have integrated field trial data with DNA sequences using different sources of enviromics, such as linear and nonlinear reaction-norm models (e.g., Jarquín et al., 2014; Morais-Júnior et al., 2018; Millet et al., 2019; Monteverde et al., 2019; Costa-Neto et al., 2020a), crop growth model (CGM) outputs (Heslot et al., 2014; Rincent et al., 2017, 2019), CGM integrated with GS (Cooper et al., 2016; Messina et al., 2018; Robert et al., 2020) and historical weather records to predict cultivars in years to come (de los Campos et al., 2020).

For example, the strategy proposed by de los Campos et al. (2020) assesses genomic × environment (G × E) patterns learned from field trials and predicts the expected performance of a cultivar in an environment but also evaluates the expected distribution of a cultivar performance over other possible weather conditions, while accounting for uncertainty in model parameters. This is a new method for the analysis of multi-environment trials and can speed up the assessment of grain yield adaptability and stability.

Another recent example is the approach that can increase the resolution in multi-environment prediction for stability by taking advantage of large-scale enviromics with different kernel methods (Costa-Neto et al., 2020a). The environmental relatedness among field trials can be shaped using linear covariances (as proposed by Jarquín et al., 2014) and non-linear methods (Gaussian kernel, deep learning, and deep kernel) (Cuevas et al., 2016, 2017, 2018, 2019; Montesinos-López et al., 2018a,b, 2019b,c). The use of non-linear kernels has led to higher accuracy gains in the prediction of novel genotypes under known conditions, but mostly in the prediction of novel environment conditions (untested environments). This approach was expanded to take account of several environmental structures across different crop development stages (Costa-Neto et al., 2020b). For the latter, the authors observed an increased ability to explain G × E in terms of genotype-specific reaction norms for key environmental factors or key development stages. This increased ability to explain G × E was important to achieve higher accuracy gains in comparison with models without enviromic information.

In a recent research article, Rogers et al. (2021) emphasized the importance of incorporating high throughput environmental data into genomic prediction models in order to carry out predictions in new environments characterized with the same environmental characteristics. The author concluded that, among other factors, G × E interactions and environmental covariates should be incorporated into prediction models to improve prediction accuracy.

Interconnection in Modern Plant Breeding

Progress toward the modernization of the statistical and quantitative genetic models for the analysis of plant breeding in multi-environment trials has become clearer as the availability of genomics, phenomics, and environments information has increased (see, among others, Vargas et al., 1998; Crossa et al., 2010; Burgueño et al., 2012; Heslot et al., 2014; Jarquín et al., 2014; Montesinos-López et al., 2017; Millet et al., 2019; Costa-Neto et al., 2020a; de los Campos et al., 2020; Robert et al., 2020). Thus, we see that all the elements described above offer a clear potential for the acceleration of genetic gains in plant breeding. However, an efficient data-based integration is required to achieve greater opportunity, particularly in terms of increasing prediction accuracy. Some of the major links between genomics, phenomics, and enviromics are outlined below, and their potential impacts are summarized in **Figure 1**.

Linking Genomics and Phenomics

Linking massive data sets from genomics and phenomics has complexities that require statistical models to deal with a very large number of correlated predictors. Montesinos-López et al. (2017) proposed linking genomics and phenomics with Bayesian functional regression models that consider all available reflectance bands (250 bands or wavelength), genomic or pedigree information, the main effects of lines and environments, as well as the effects of interaction. They observed that the models with wavelength × environment interaction terms were the most accurate for the prediction of performance in three different environments and at various crop development time points. The functional regression models are parsimonious and computationally efficient because the mathematical basis functions allow the selection of only 21 beta coefficients (rather than using all 250). Recently, Lopez-Cruz et al. (2020) proposed a method to predict the genetic merit of cultivars from high-dimensional HTP data by integrating high-dimensional regressions into the standard selection index methodology.

Linking Multi-Trait and Multi-Environment Data

Multi-trait multi-environment data (MTME) take advantage of large-scale correlations among different traits evaluated across

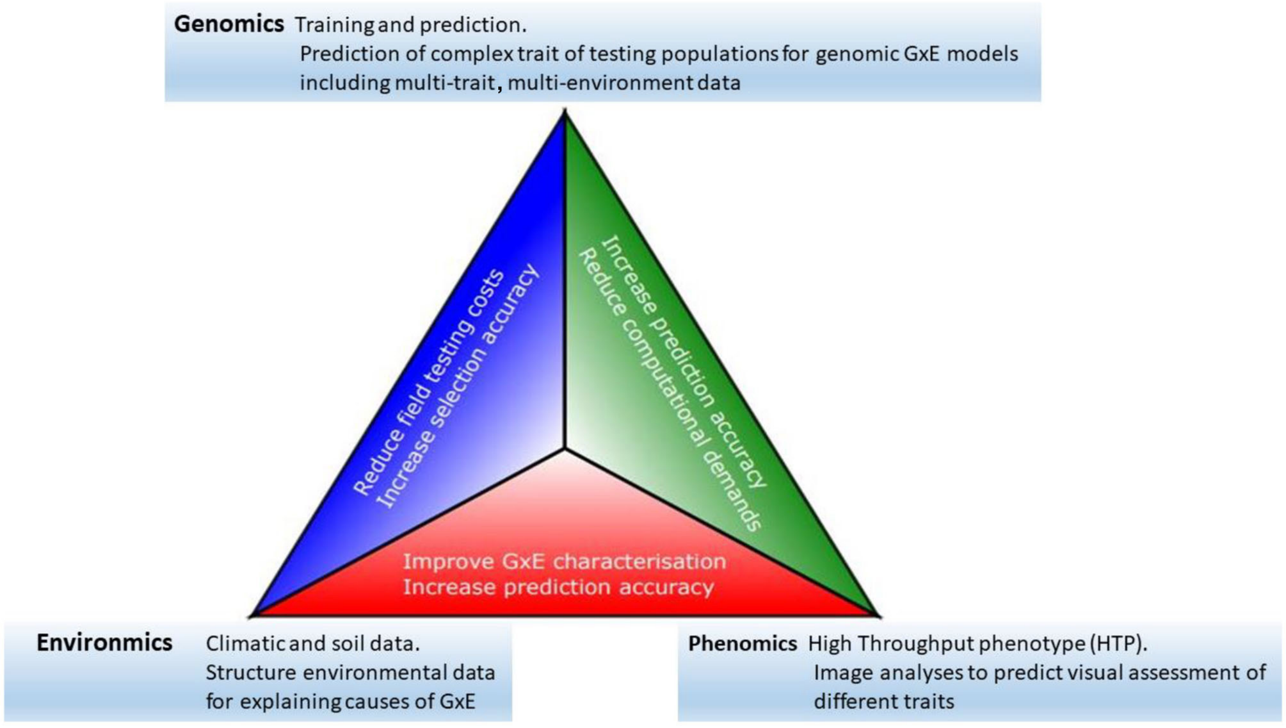

Genomics Training and prediction.
Prediction of complex trait of testing populations for genomic GxE models including multi-trait, multi-environment data
Reduce field testing costs
Increase selection accuracy
Increase prediction accuracy
Reduce computational demands
Improve GxE characterisation
Increase prediction accuracy
Environmics Climatic and soil data.
Structure environmental data for explaining causes of GxE
Phenomics High Throughput phenotype (HTP).
Image analyses to predict visual assessment of different traits

FIGURE 1 | The modern plant-breeding triangle incorporates genomics, phenomics, and enviromics. Connections between each of these elements can be beneficial for the acceleration of genetic gains.

diverse environments to train accurate GS models. Because of this, the use of GS in MTME data is a promising approach to reduce field phenotyping efforts. For example, Ibba et al. (2020) evaluated the prediction performance of 13 quality traits in wheat using two multi-trait models and five data sets based on field evaluations over two consecutive years. In the second year (testing), lines were predicted using the quality information obtained in the first year (training). For most of the quality traits, they found moderate to high prediction accuracies, suggesting that the use of GS at earlier stages could be recommendable. Overall, the results indicate that the Bayesian MTME model helps capture the correlation among traits and the correlation among years, thus increasing prediction accuracy. Finally, we envision perspectives of modeling MTME-based reaction norms involving other omics, such as phenomics and enviromics. The latter can enhance the MTME analysis in terms of creating more biological models of crop growth, development, and yield components (e.g., Robert et al., 2020).

Interplay Genomics and Enviromics

Since the 1960s, several researchers have suggested the use of environmental information to explain the differences in cultivars due to G × E interactions (e.g., Perkins and Jinks, 1968; Freeman and Perkins, 1971; Wood, 1976; Vargas et al., 1998; Crossa et al., 1999). The use of genomics with enviromics is the basis for the prediction of cultivars across diverse growing conditions (e.g., Jarquín et al., 2014; Messina et al., 2018; Millet et al., 2019), which is useful for the prediction of global warming.

However, the efforts to implement environmental covariates into genomic selection models usually focus on a few environmental covariates such as temperature, precipitation, and sun radiation defined over specific developmental stages of the crop. With the use of large-scale envirotyping data, it is possible to design a global-scale envirotyping network of field trials to train GS models and perform "enviromic assembly" to predict a wider number of growing conditions from historical climate and soil data (R package EnvRtype, Costa-Neto et al., 2020b). In addition, research is underway for the study of model Enviromic + Genomic prediction (E-GP) to link genotype-phenotype variations, as well as to explain phenotypic variations across environments. As a predictive breeding tool, E-GP can contribute to the study of G × E structures, in which, as an exploratory tool, E-GP can contribute to the optimization of experimental networks of field trials and lead to more efficient training sets for GS (e.g., Rincent et al., 2017). In addition, for the early stages of selection, genomics and enviromics can be used to design optimized phenotyping trials and predict the breeding values of the selection candidate (Morais-Júnior et al., 2018) or single cross-hybrid prediction (Costa-Neto et al., 2020a).

Through enviromic assembly, it is possible to establish relatedness among field trials and thus use only the most representative set of experiments for training GS models. Another perspective of E-GP is the use of large-scale environmental data in training models involving genotype-specific reaction norms (e.g., Ly et al., 2018; Millet et al., 2019) and phenotypic landscapes implemented by genomics with crop growth models (CGM) (e.g., Messina et al., 2018; Bustos-Korts et al., 2019; Robert et al., 2020). The possible use of image-based responses related to main environmental stresses, such as heat and drought-stress, can also boost the implementation of genomic-assisted platforms for predictive purposes and are capable of better representing the plant-environment interplay.

Future Perspectives

In order to meet the well-documented challenges of food and nutrition security, there is a pressing need to use new technologies to accelerate the progress of plant breeding. These methods can be incorporated into conventional phenotypic breeding programs or help redesign established phenotypic breeding pipelines to enable a gradual shift toward a more data-driven perspective. The benefits of phenomics and enviromics together in benchmark genomic pipelines offer the potential to deliver larger increases in accuracy and efficiency of breeding pipelines when we select better-adapted genotypes in a cost-effective manner, as well as in a reduced timeframe. Genomics, phenomics, multi-trait, and enviromics analyses are interconnected, and their use can be optimized based on resources and program structure. Together, they offer a pathway for conventional phenotypic breeding to envision a diverse set of opportunities to accelerate genetic gains.

AUTHOR CONTRIBUTIONS

JC prepared the first drafts of the opinion, RF-N, OM-L, GC-N, SD, and AM-L read and corrected the first version. AB produced several reviews of the documents and worked with JC to finalize the definitive version. All authors contributed to the article and approved the submitted version.

FUNDING

Open Access fees are 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)] that generated the CIMMYT data analyzed in this study. We are also thankful for the financial support provided by the Foundations for Research Levy on Agricultural Products (F.F.J.) and the Agricultural Agreement Research Fund (J.A.) in Norway through NFR grant 267806 as well as the CIMMYT CRP (maize and wheat).

REFERENCES

Atkinson, J. A., Jackson, R. J., Bentley, A. R., Ober, E., and Wells, D. M. (2018). Field phenotyping for the future. *Annu. Plant Rev. Online.* doi: 10.1002/9781119312994.apr0651

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

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/cropsci 2011.06.0299

Bustos-Korts, D., Malosetti, M., Chenu, K., Chapman, S., Boer, M. P., Zheng, B., et al. (2019). From QTLs to adaptation landscapes: using genotype-to-phenotype models to characterize G×E over time. *Front. Plant Sci.* 10, 1–23. doi: 10.3389/fpls.2019.01540

Collard, B. C. Y., Gregorio, G. B., Thomson, M. J., Islam, M. R., Vergara, G. V., Laborte, A. G., et al. (2019). Transforming rice breeding: re-designing the irrigated breeding pipeline at the International Rice Research Institute (IRRI). *Crop Breed Genet. Genome* 1:e190008. doi: 10.20900/cbgg20190008

Cooper, M., Technow, F., Messina, C., Gho, C., and Radu Totir, L. (2016). Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial. *Crop Sci.* 56, 2141–2156. doi: 10.2135/cropsci2015.08.0512

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2020a). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity (Edinb).* 126, 92–106. doi: 10.1038/s41437-020-00353-1

Costa-Neto, G., Galli, G., Fanelli, H., Crossa, J., and Fritsche-Neto, R. (2020b). EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. *bioRxiv*[preprint]. doi: 10.1101/2020.10.14.339705

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O. A., 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

Crossa, J., Vargas, M., and Joshi, A. K. (2010). Linear, bilinear, and linear-bilinear fixed and mixed models for analyzing genotype x environment interaction in plant breeding and agronomy. *Can. J. Plant Sci.* 90, 561–574. doi: 10.4141/CJPS10003

Crossa, J., Vargas, M., Van Eeuwijk, F. A., Jiang, C., Edmeades, G. O., and Hoisington, D. (1999). Interpreting genotype x environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theor. Appl. Genet.* 99, 611–625. doi: 10.1007/s001220051276

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P., and de los Campos, G. (2017). Bayesian genomic prediction with genotype x environment kernel models. *G3: Genes|Genomes|Genetics* 7, 41–53. doi: 10.1534/g3.116.035584

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., de los Campos, G., et al. (2016). Genomic prediction of genotype x environment interaction kernel regression models. *Plant Genome* 9, 1–20. doi: 10.3835/plantgenome2016.03.0024

Cuevas, J., Granato, I., Fritsche-Neto, R., Montesinos-Lopez, O. A., Burgueño, J., Bandeira e Sousa, M., et al. (2018). Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. *G3: Genes|Genomes|Genetics* 8, 1347–1365. doi: 10.1534/g3.117.300454

Cuevas, J., Montesinos-López, O. A., Juliana, P., Guzmán, C., Pérez-Rodríguez, P., González-Bucio, J., et al. (2019). Deep kernel for genomic and near infrared prediction in multi-environments breeding trials. *G3: Genes|Genomes|Genetics* 9, 2913–2924. doi: 10.1534/g3.119.400493

de los Campos, G., Pérez-Rodríguez, P., Bogard, M., Gouache, D., and Crossa, J. (2020). A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. *Nat. Commun.* 11:4876. doi: 10.1038/s41467-020-18480-y

Freeman, G. H., and Perkins, J. M. (1971). Environmental and genotype-environmental components of variability: Viii Relations between genotypes grown in different environments and measures of these environments. *Heredity (Edinb).* 27, 15–23. doi: 10.1038/hdy.1971.67

García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-lópez, F. J., and Van Tassell, C. P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc. Natl. Acad. Sci. U.S.A.* 113, E3395–E4004. doi: 10.1073/pnas.1519061113

Grattapaglia, D., Silva-Junior, O. B., Resende, R. T., Cappa, E. P., Müller, B. S. F., Tan, B., et al. (2018) Quantitative genetics and genomics converge to accelerate forest tree breeding. *Front. Plant Sci.* 871, 1–10. doi: 10.3389/fpls.2018.01693

Heslot, N., Akdemir, D., Sorrels, 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

Hickey, J. M., Crossa, J., Babu, R., and de los Campos, G. (2012). Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Sci.* 52, 654–663. doi: 10.2135/cropsci 2011.07.0358

Huang, M., Balimponya, E. G., Mgonja, E. M., McHale, L. K., Luzi-Kihupi, A., Guo-Liang Wang, G.-L., et al. (2019). Use of genomic selection in breeding rice (*Oryza sativa* L.) for resistance to rice blast (*Magnaporthe oryzae*). *Mol. Breed.* 39:114. doi: 10.1007/s11032-019-1023-2

Ibba, M. I., Crossa, J., Montesinos-López, O. A., Montesinos-López, A., Juliana, P., Guzman, C., et al. (2020). Genome-based prediction of multiple wheat quality traits in multiple years. *Plant Genome* 1:14. doi: 10.1002/tpg2.20034

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucour, 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

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. doi: 10.1038/s41598-020-65011-2

Ly, D., Huet, S., Gauffreteau, A., Rincent, R., Touzy, G., Mini, A., et al. (2018). Whole-genome prediction of reaction norms to environmental stress in bread wheat (*Triticum aestivum* L.) by genomic random regression. *F. Crop. Res.* 216, 32–41. doi: 10.1016/j.fcr.2017.08.020

Massman, J. M., Jung, H.-J. G., and Bernardo, R. (2013). Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Sci.* 53, 58–66. doi: 10.2135/cropsci2012.02.0112

Matias, F. I., Caraza-Harter, M. V., and Endelman, J. B. (2020). FIELDimageR: an R package to analyze orthomosaic images from agricultural field trials. *Plant Phenome J.* 3, 1–6. doi: 10.1002/ppj2.20005

Messina, C. D., Technow, F., Tang, T., Totir, R., Gho, C., and Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: integrating crop growth models (CGM) with whole genome prediction (WGP). *Eur. J. Agron.* 100, 151–162. doi: 10.1016/j.eja.2018.01.007

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

Montesinos-López, A., Montesinos-López, O. A., Cuevas, J., Mata-López, W. A., Burgueño, J., Mondal, S., et al. (2017). Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper-spectral image data. *Plant Methods* 13:62. doi: 10.1186/s13007-017-0212-4

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018a). Multi-environment genomic prediction of plant traits using deep learners with a dense architecture. *G3: Genes|Genomes|Genetics* 8, 3813–3828. doi: 10.1534/g3.118.200740

Montesinos-López, M. A., Montesinos-López, A., Luna-Vázquez, J. F., Toledo, F. H., Paulino Pérez-Rodríguez, P., Lillemo, M., et al. (2019a). An R package for bayesian analysis of multi-environment and multi-trait multi-environment data for genome-based prediction. *G3. Genes|Genomes|Genetics* 9, 1355–1369. doi: 10.1534/g3.119.400126

Montesinos-López, O. A., Martin-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C., Montesinos-López, A., et al. (2019b). 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|Genetics* 9, 601–618. doi: 10.1534/g3.118.200998

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., et al. (2019c). New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3: Genes|Genomes|* 9, 1545–1556. doi: 10.1534/g3.119.300585

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

Montesinos-López, O. A., Montesinos-López, A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018b). Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant. *G3: Genes|Genomes|Genetics* 8, 3829–3840. doi: 10.1534/g3.118.200728

Monteverde, E., Gutierrez, L., Blanco, P., Pérez de Vida, F., Rosas,J. E., Bonnecarrère, V., et al. (2019). Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (*Oryza sativa* L.) Grown in Subtropical Areas. *G3 Genes|Genomes|Genetics* 9, 1519–1531. doi: 10.1534/g3.119.400064

Morais-Júnior, O. P., Duarte, J. B., Breseghello, F., Coelho, A. S. G., and Magalhães, A. M., Jr. (2018). Single-step reaction norm models for genomic prediction in multienvironment recurrent selection trials. *Crop Sci.* 58, 592–607. doi: 10.2135/cropsci2017.06.0366

Perkins, J. M., and Jinks, J. L. (1968). Environmental and genotype-environmental components of variability. 3. Multiple lines and crosses. *Heredity (Edinb).* 23, 339–356. doi: 10.1038/hdy.1968.48

Resende, R. T., Piepho, H. P., Rosa, G. J. M., Silva-Junior, O. B., Silva, F. F. E., de Resende, M. D. V., et al. (2020). Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theor. Appl. Genet.* 134, 95–112. doi: 10.1007/s00122-020-03684-z

Rincent, R., Kuhn, E., Monad, H., Oury, F. X., Rousset, M., Allard, V., et al. (2017). Optimization of multi - environment trials for genomic selection based on crop models. *Theor. Appl. Genet.* 130, 1735–1752. doi: 10.1007/s00122-017-2922-4

Rincent, R., Malosetti, M., Ababaei, B., Touzy, G., Mini, A., Bogard, M., et al. (2019). Using crop growth model stress covariates and AMMI decomposition to better predict genotype-by-environment interactions. *Theor. Appl. Genet.* 132, 3399–3411. doi: 10.1007/s00122-019-03432-y

Robert, P., Le Gouis, J., and Rincent, R. (2020). Combining crop growth modeling with trait-assisted prediction improved the prediction of genotype by environment interactions. *Front. Plant Sci.* 11, 1–11. doi: 10.3389/fpls.2020.00827

Rogers, A. R., Dunne, J. C., Romay, M. C., Bohn, M., Buckler, E. S., Ciampitti, I. A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize. *G3:Genes|Genomes|Genetics* 1:jkaa050.doi: 10.1093/g3journal/jkaa050

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., et al. (2016). Genome-enabled prediction models for yield related traits in Chickpea. *Front. Plant Sci.* 7:1666. doi: 10.3389/fpls.2016.01666

Vargas, M., Crossa, J., Sayre, K., Reynolds, M., Ramirez, M. E., and Talbot, M. (1998). Interpreting genotype X environment interaction using partial least squares regression. *Crop Sci.* 38, 679–689. doi: 10.2135/cropsci1998.0011183X003800030010x

Vivek, B. S., Krishna, G. K., Vengadessan, V., Babu, R., Zaidi, P. H., Kha, L. Q., et al. (2017). Use of genomic estimated breeding values results in rapid genetic gains for drought tolerance in maize (*Zea mays* L.). *Plant Genome* 10, 1–8, doi: 10.3835/plantgenome2016.07.0070

Wolfe, M. D., Del Carpio, D. P., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., Kayondo, I. S., et al. (2017). Prospects for Genomic Selection in Cassava Breeding. *Plant Genome* 10:15. doi: 10.3835/plantgenome2017.03.0015

Wood, J. T. (1976). The use of environmental variables in the interpretation of genotype-environment interaction. *Heredity (Edinb).* 37, 1–7. doi: 10.1038/hdy.1976.61

Zhang, X., Pérez-Rodríguez, P., Burgueño, J., Olsen M Buckler, E., Atlin, G., Prasanna, B. M., et al. (2017) Rapid cycling genomic selection in a multiparent tropical maize population. *G3:Gene/Genome/Genet* 7, 2315–2326. doi: 10.1534/g3.117.043141

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 Plant Science

REVIEW
published: 30 April 2021
doi: 10.3389/fpls.2021.665195

Genomic Selection in Tropical Forage Grasses: Current Status and Future Applications

Rosangela M. Simeão[1†], Marcos D. V. Resende[2†], Rodrigo S. Alves[3†], Marco Pessoa-Filho[4†], Ana Luisa S. Azevedo[5†], Chris S. Jones[6†], Jorge F. Pereira[5†] and Juarez C. Machado[5*†]

[1] Embrapa Gado de Corte, Campo Grande, Brazil, [2] Embrapa Café, Universidade Federal de Viçosa, Viçosa, Brazil, [3] Instituto Nacional de Ciência e Tecnologia do Café, Universidade Federal de Viçosa, Viçosa, Brazil, [4] Embrapa Cerrados, Brasília, Brazil, [5] Embrapa Gado de Leite, Juiz de Fora, Brazil, [6] International Livestock Research Institute, Nairobi, Kenya

OPEN ACCESS

Edited by:
*Fernando H. Toledo,
International Maize and Wheat Improvement Center, Mexico*

Reviewed by:
*Russell Jessup,
Texas A & M University, United States
Filipe Matias,
University of Wisconsin-Madison, United States
Jose Crossa,
International Maize and Wheat Improvement Center, Mexico*

***Correspondence:**
*Juarez C. Machado
juarez.machado@embrapa.br*

†ORCID:
*Rosangela M. Simeão
orcid.org/0000-0001-8351-846X
Marcos D. V. Resende
orcid.org/0000-0002-3087-3588
Rodrigo S. Alves
orcid.org/0000-0002-3038-6210
Marco Pessoa-Filho
orcid.org/0000-0002-7333-4429
Ana Luisa S. Azevedo
orcid.org/0000-0003-1939-0339
Chris S. Jones
orcid.org/0000-0001-9096-9728
Jorge F. Pereira
orcid.org/0000-0001-9340-065X
Juarez C. Machado
orcid.org/0000-0003-2551-2080*

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

***Received:** 07 February 2021*
***Accepted:** 06 April 2021*
***Published:** 30 April 2021*

Citation:
Simeão RM, Resende MDV, Alves RS, Pessoa-Filho M, Azevedo ALS, Jones CS, Pereira JF and Machado JC (2021) Genomic Selection in Tropical Forage Grasses: Current Status and Future Applications. Front. Plant Sci. 12:665195. doi: 10.3389/fpls.2021.665195

The world population is expected to be larger and wealthier over the next few decades and will require more animal products, such as milk and beef. Tropical regions have great potential to meet this growing global demand, where pasturelands play a major role in supporting increased animal production. Better forage is required in consonance with improved sustainability as the planted area should not increase and larger areas cultivated with one or a few forage species should be avoided. Although, conventional tropical forage breeding has successfully released well-adapted and high-yielding cultivars over the last few decades, genetic gains from these programs have been low in view of the growing food demand worldwide. To guarantee their future impact on livestock production, breeding programs should leverage genotyping, phenotyping, and envirotyping strategies to increase genetic gains. Genomic selection (GS) and genome-wide association studies play a primary role in this process, with the advantage of increasing genetic gain due to greater selection accuracy, reduced cycle time, and increased number of individuals that can be evaluated. This strategy provides solutions to bottlenecks faced by conventional breeding methods, including long breeding cycles and difficulties to evaluate complex traits. Initial results from implementing GS in tropical forage grasses (TFGs) are promising with notable improvements over phenotypic selection alone. However, the practical impact of GS in TFG breeding programs remains unclear. The development of appropriately sized training populations is essential for the evaluation and validation of selection markers based on estimated breeding values. Large panels of single-nucleotide polymorphism markers in different tropical forage species are required for multiple application targets at a reduced cost. In this context, this review highlights the current challenges, achievements, availability, and development of genomic resources and statistical methods for the implementation of GS in TFGs. Additionally, the prediction accuracies from recent experiments and the potential to harness diversity from genebanks are discussed. Although, GS in TFGs is still incipient, the advances in genomic tools and statistical models will speed up its implementation in the foreseeable future. All TFG breeding programs should be prepared for these changes.

Keywords: apomixis, brachiaria, elephant grass, forage breeding, Guinea grass, marker-assisted selection, polyploidy

INTRODUCTION

The tropics are home to about a third of the world's population, accounting for 36% of the Earth's landmass, and where most of the global demographic increase takes place (Morales, 2009). The tropical region is the center of origin and domestication of many of the world's most important food crops and is responsible for 50% beef production and 40% milk production worldwide (Morales, 2009; Alexandratos and Bruinsma, 2012). Despite the huge importance of the tropical region, there is an evident gap in technological development between the tropical nations and industrialized temperate countries. The tropical region has great potential to meet the growing global demand for food requirements with intensification through improved management and technologies. One important step toward intensification is the acceleration of breeding programs.

Plant breeding has evolved from a rudimentary process in its early stages to a modern and sophisticated system in the past few decades. The rediscovery of Mendel's laws of genetics in 1900 is one of the pillars of modern plant breeding (Hallauer, 2011), which can also count on modern techniques such as high-throughput sequencing, bioinformatics, and automated phenotyping (Barabaschi et al., 2016). Breeding pipelines can apply these techniques to increase the rates of genetic gain taking into account the parameters found in the "breeder's equation" (Lush, 1937; Eberhart, 1970), which states that the genetic gain is directly proportional to the accuracy of the observed phenotype in relation to the true phenotype and genotype, selection intensity, and genetic variation, but inversely proportional to the time of the breeding cycle. Manipulating variables in the "breeder's equation" can increase genetic gain as well as reduce the timeframe to develop new cultivars (Pereira et al., 2018).

Pastures are the main food source for animal feeding in the tropics, as in Brazil, where approximately 90% of the livestock are solely grass-fed (Silva et al., 2016). An increase in productivity and quality of tropical forage grasses (TFGs) will have a significant impact on livestock production. Although, cattle and buffaloes already contribute to the largest proportion of global animal protein supply, increased quantities of milk and beef are necessary due to growing demands. Production will have to increase by 57% for beef and 48% for milk by 2050 compared to that in 2005, as projected by the FAO (Alexandratos and Bruinsma, 2012), while other estimates indicate that the global demand for livestock products will double by 2050 (Bajželj et al., 2014; Rao et al., 2015). This higher production needs to take into account scenarios where the land destined for pastures may have to be reduced, as has been happening in Brazil (IBGE, 2016). Efforts to breed TFGs focus on increasing productivity and quality while also reducing losses due to biotic and abiotic stresses. However, breeding efforts have been hindered by many features of tropical forages that render the implementation of more dynamic breeding programs difficult. TFGs encompass perennial monocotyledonous plants from the family Poaceae, mostly polyploid, with a C4 photosynthetic pathway and showing both sexual and apomictic reproductive systems. Breeding programs of TFGs face challenges such as different ploidy levels and reproductive modes, evaluation of perennial plants over different cuts, distribution of efforts among different species, the evaluation of traits being laborious and expensive, and most breeding programs being held by public institutions. The development and release of a new cultivar can take up to 10 years (Jank et al., 2014).

Genomic selection (GS) offers the opportunity to increase agricultural production and reduce the breeding interval cycle to at least half of the conventional time (Crossa et al., 2017). Reduction of the breeding cycle is the main advantage of GS in forage breeding (Simeão-Resende et al., 2014). GS and genome-wide association studies (GWAS) have enormous potential for use in the selection of complex traits such as yield, disease, and insect resistance, facilitating the rapid selection of new cultivars to meet the future demand for food and fodder (Heffner et al., 2010; Talukder and Saha, 2017). However, breeding programs of TFGs are still behind those of grain and fiber crops, and even those of temperate/sub-tropical forages, regarding the application of genomic tools as a strategy to accelerate cultivar development. Challenges in applying GS in tropical forages include designing and obtaining adequately sized training populations; developing high-quality, low-cost, and reproducible marker panels; dealing with polyploidy; and gaining knowledge of the genetic architecture of target traits.

This article provides an overview of GS in TFGs focusing on the current scenario, recent advances, and prospects for the effective application of tools and strategies to accelerate TFG breeding. We have focused on elephant grass (*Cenchrus purpureus* syn. *Pennisetum purpureum*), Guinea grass (*Megathyrsus maximus* syn. *Panicum maximum*), and brachiaria (*Urochloa brizantha* syn. *Brachiaria brizantha*, *U. decumbens* syn. *B. decumbens*, and *U. ruziziensis* syn. *B. ruziziensis*), which account for most of the pastures in many parts of the world, including Africa, Asia, Australia, and Latin America. Specific features of breeding programs, availability of genomic resources, statistical methods for GS, and gaps in the application of GS in TFG breeding are discussed here. This discussion is essential for the initiation and practical implementation of GS in TFG breeding programs.

TFG BREEDING

TFG breeding began relatively recently (Valle et al., 2009). For example, EMBRAPA, the Brazilian Agricultural Research Corporation, started its breeding programs for *Urochloa* and

Abbreviations: BLUP, best linear unbiased prediction; CIAT, International Center for Tropical Agriculture; EMBRAPA, Empresa Brasileira de Pesquisa Agropecuária; GBS, genotyping-by-sequencing; GEBV, genomic breeding values; GETV, total (genotypic) genomic values; GLS, generalized least squares; GS, genomic selection; GV, genomic value; GWAS, genome-wide association studies; IBERS, Institute of Biology, Environmental and Rural Sciences at Aberystwyth University; ICARDA, International Center for Agricultural Research in the Dry Areas; ILRI, International Livestock Research Institute; LD, linkage disequilibrium; NIRS, near-infrared spectroscopy; MAF, minor allele frequency; MAS, marker-assisted selection; QTL, quantitative trait loci; QTN, quantitative trait nucleotide; REML, restricted maximum likelihood; SARDI, South Australian Research and Development Institute; SNP, single-nucleotide polymorphism; TFG, tropical forage grass; USDA, United States Department of Agriculture.

Megathyrsus in the 1980s (Jank et al., 2014), while the elephant grass breeding program only began in 1991 (Pereira et al., 2017). One of the first steps toward an effective breeding program is the compilation of a germplasm collection. Approximately 17,000 accessions of TFGs have been preserved in the primary global germplasm banks, such as CIAT, EMBRAPA, IBERS, ICARDA, ILRI, SARDI, and USDA, where *Urochloa* spp., *Cenchrus* spp., and *Megathyrsus* spp. correspond to most of the accessions alongside *Digitaria* spp. and *Paspalum* spp. The genetic variability maintained in these germplasm banks is invaluable, and these banks offer genetic resources adapted to varied edaphoclimatic conditions and diverse purposes. However, there is no corresponding use of this variability during crossings in practical breeding programs. This indicates that these collections are not being used to their full potential, although, initiatives of germplasm exchange between institutions have been taken to increase genetic variability that can be used in breeding programs (Negawo et al., 2018; Habte et al., 2020). Better characterization of germplasm collections will enable quicker utilization to meet the dynamic demands of the production sector with the emergence of diverse limitations due to climate, pests, or changes in production systems.

The limited use of accessions preserved in germplasm banks for crossing is one of the limitations of TFG breeding (Valle et al., 2009; Pereira et al., 2018). Other limitations include the use of a large number of candidate genera and species, insufficient information on the biology of the species, low genetic variability for important traits, polyploidy, a complex mode of reproduction (apomixis), wild characteristics of important species (dehiscence and malformation of seeds, anti-quality factors, and sensitivity to photoperiod), lack of information on the genetic control and heritability of agronomic traits, and little participation of the private sector in the development of cultivars (Valle et al., 2009; Sandhu et al., 2015; Pereira et al., 2018). It is worth noting that brachiaria, Guinea grass, and elephant grass are perennial species, which implies that most traits are evaluated in the field over long periods, including several cuts. It is common for a specific trait to show variation among different cuts (Rocha et al., 2019), which is very different compared to that of annual species. The impact of these limitations, along with the differences in market demands and the ability of producers to absorb new releases, can be seen in the low number of cultivars released through the years when compared to the release of grain and fiber crops (**Figure 1**).

Despite these limitations, the improvement of forage grasses has revolutionized the pastoral systems. In a study conducted in sub-Saharan Africa, legume, and grass cultivars released for different animal production systems increased forage production by 2.65 times when compared to that of traditional cultivars. Production was even higher when only forage grass was used (Paul et al., 2020). Until recently, germplasm introduction was the key method used for forage grass breeding, which involved the evaluation and selection of germplasm accessions as a strategy to obtain cultivars. This method was used to release *U. brizantha* cv. Marandu in 1984 by EMBRAPA (Nunes et al., 1984), which is currently cultivated on fifty-mega hectares of land in Brazil alone (Jank et al., 2014). The germplasm introduction method, albeit simple, rapid, and cost-effective, tends to be prone to exhaustion as it requires the use of accessions collected from nature or accessions obtained from a germplasm bank in other breeding programs (Jank et al., 2011). In addition, natural habitats of species are being increasingly degraded, with loss of variability as well as restrictions in free access to germplasm across different countries and breeding programs, notably due to recent laws for access to genetic diversity and protection of cultivars (Pengelly and Maass, 2019). Since 2000, the use of recombination as a key strategy for cultivar development has intensified. For example, among the new cultivars released by various breeding programs, intra- or interspecific hybrids, especially of *Urochloa*, *Megathyrsus,* and *Cenchrus*, have been highlighted, in which the favorable traits of their progenitors are gathered. Of note, in the last decade, long-term recurrent selection programs have been established for major species, and promising results have been achieved (Miles et al., 2006; Reis et al., 2008; Barrios et al., 2013). The main objectives of TFG breeding are to identify and develop improved genotypes that contribute to increased animal productivity and reduced environmental impact (**Figure 2**). Thus, not only a better agronomic behavior of the plant, but also a more productive performance of the animal is sought, while ensuring minimal environmental impact (Valle, 2001).

Although, TFG breeding programs have been successful in releasing new and important cultivars over the years, there are certain challenges to overcome. In Brazil, these challenges include reducing losses due to biotic stresses (especially spittlebug attacks), increasing adaptation based on expected climate changes, and improving nutritional value to enhance animal performance, resulting in more beef and milk per kilogram of pasture. To address these challenges, research priorities have focused on the development of new capabilities such as the availability of genome sequences, high-throughput genotyping, and germplasm characterization of tropical forage grasses; identification of genes associated with important traits; development and use of large-scale phenotyping tools; and implementation of GS (Pereira et al., 2018).

Therefore, the prospects of applying genomic tools in TFG breeding programs are promising, and these tools, coupled with adequate pasture management, can continue to promote substantial advances in livestock productivity. For each forage species, the objectives of the breeding programs should be well-defined, as highlighted in **Figure 2**. In addition to clear objectives, it is important to use the latest technologies available to accelerate the development of cultivars. In this regard, the use of genomic tools for TFG breeding is fundamental.

GS: AN APPRAISAL IN TFG BREEDING

Peculiarities in Breeding Perennial TFGs: Polyploidy and Apomixis

Perennial forages require selection methods that consider the effects of both between families and within family individuals for higher selection gains, mostly in lower magnitudes of narrow-sense heritability (NSH) (Simeão-Resende et al., 2013, 2014). In forage breeding, a combination of GS methods is expected to be useful for predicting genomic breeding values (GEBVs)

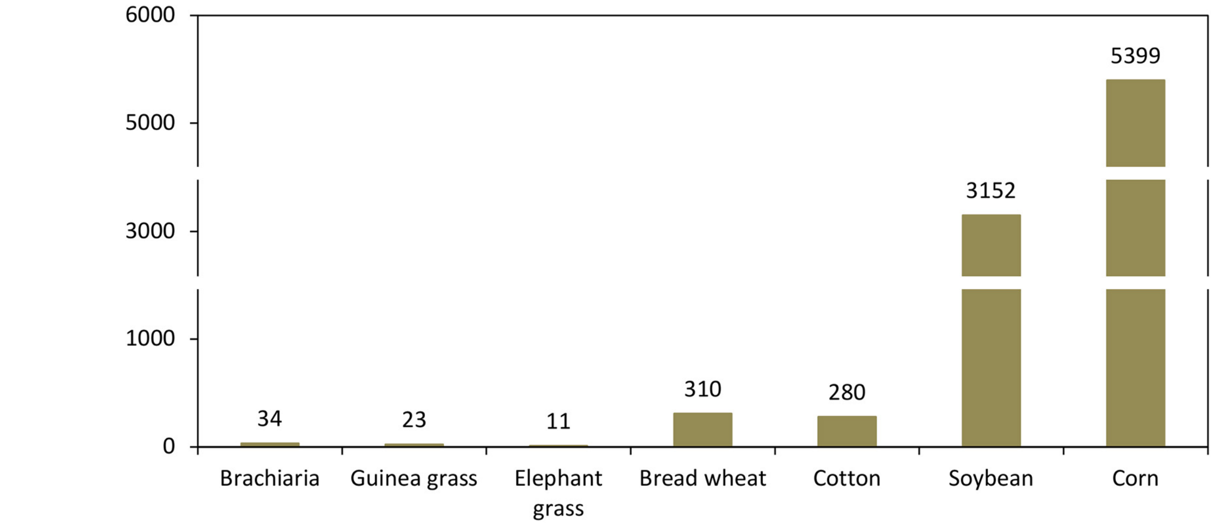


FIGURE 1 | Number of cultivars of brachiaria, Guinea grass, and elephant grass registered in Brazil in comparison with grain and fiber crops. The list was obtained from the National Cultivar Registry Data Bank (*Registro Nacional de Cultivares – RNC*) that is a requirement from the Brazilian Ministry of Agriculture, Livestock and Food Supply since 1997. The numbers shown here were retrieved on 19 November 2020 (http://sistemas.agricultura.gov.br/snpc/cultivarweb/cultivares_registradas.php). Because the difference in the number of cultivars is high, the Y-axis has been adjusted.

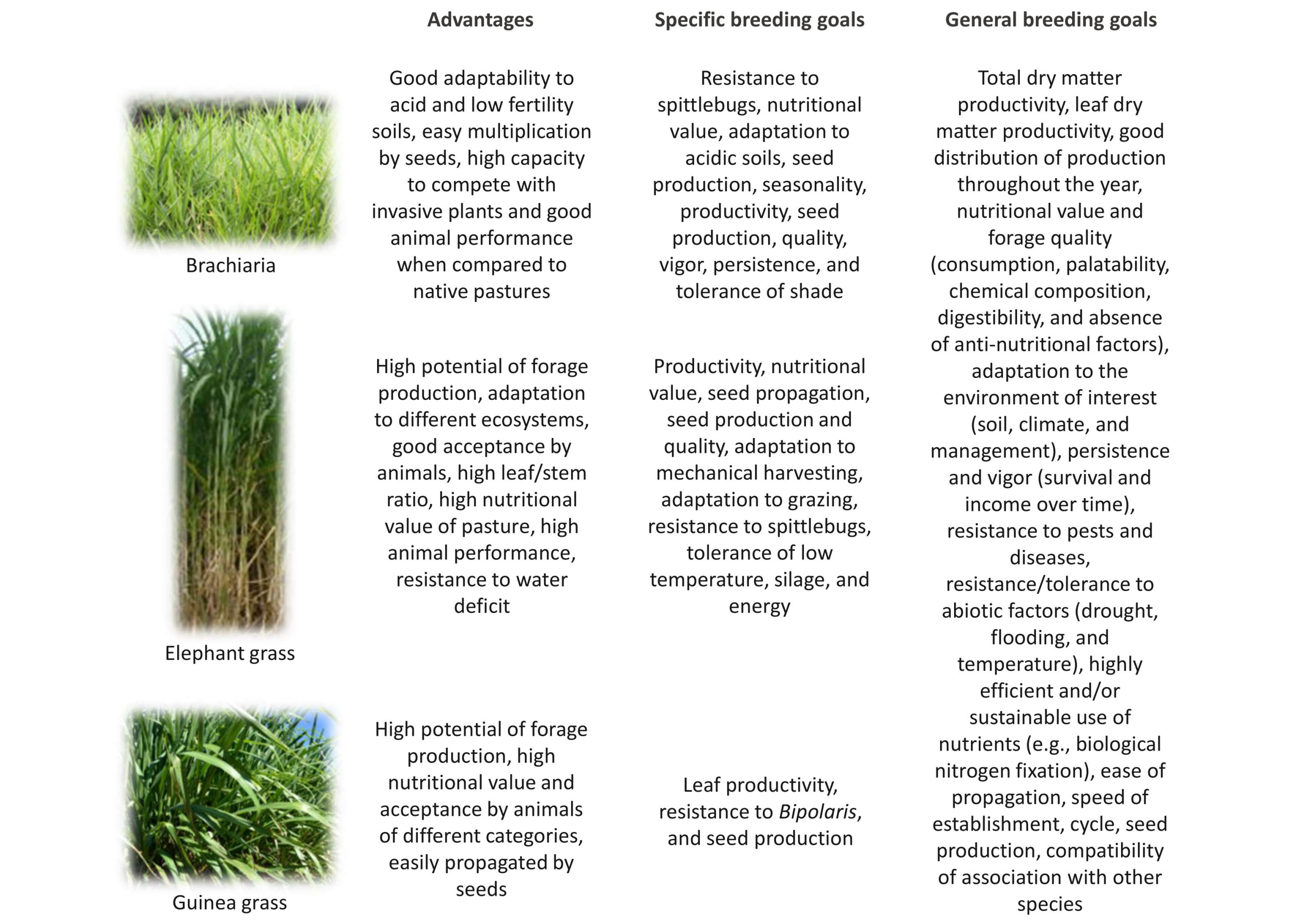
and total (genotypic) genomic values (GETVs), for clonally propagated cultivars. The estimation of marker effects and genomic values should enable an increase in selection accuracy and may reduce the time required for completing a breeding cycle and the evaluation cost per genotype. Forage breeding methods associated with GS show differential accuracy and gains, as demonstrated by Simeão-Resende et al. (2014).

Considering all these premises, and the methods of the current breeding programs, the uses, and advances of GS in tropical forages will be presented. Firstly, facts about perennial TFGs must be pointed out, as most are polyploids and reproduce apomictically. Apomixis is asexual reproduction by seed (Barcaccia et al., 2020) producing genetically identical progeny (Hand and Koltunow, 2014). In both important genera of TFG, *Megathyrsus* and *Urochloa*, gametophytic apomixis subtype apospory occurs (Ozias-Akins and van Dijk, 2007) which is a mode of reproduction in which the embryo originates from a polyploid nucellar cell as a maternal clone by the seed (Valle and Savidan, 1996). Therefore, the commercial cultivars of these species are generally both polyploid and apomictic.

Autotetraploid individuals have been developed in *U. ruziziensis* and *M. maximus* by artificially duplicated chromosomes from diploid sexual individuals (Jank et al., 2014). These sexual individuals are essential for hybridization with apomictic ones in breeding programs of both genera to increase genetic variability and enhance selection. Therefore, cytogenetic analysis is constantly performed in parents and hybrids and should be evaluated by considering the importance of the species targeted by GS.

According to Bourke et al. (2018), the knowledge of the meiotic behavior of a species is sometimes required to analyze polyploid data using dosage calling software that uses the expected segregation ratios in the F_1 autotetraploid population. Unlike allotetraploids, autotetraploids do not behave like diploids during meiosis and require specialized methods and tools for genetic studies and mapping (Gallais, 2003). Autotetraploid plants exhibit polysomic inheritance, which can be detected during a cytogenetic analysis by visualization of tetravalent formation and segmental pairing among "partially homologous" chromosomes (Stebbins, 1947) as well as by molecular inference (Worthington et al., 2016). The consequence of chromosome pairing in a tetravalent is the generation of unbalanced gametes and individuals with non-Mendelian inheritance. Even in recent autotetraploids induced by colchicine, chromosome pairing may not show tetravalent formation or other meiotic abnormalities (Pagliarini et al., 2008); however, the four alleles per locus are always present. This may generate errors in genetic mapping, haplotype designation, and the estimation of marker effects, which are important factors in genomic prediction.

Diploid sexual individuals of *M. maximus* were collected in Korogwe, Tanzania, and artificially duplicated (Jank et al., 2014). The cytogenetic evaluation of autotetraploid ($2n = 4x = 32$) sexual and supposedly segmental allopolyploid apomictic plants revealed a low-to-moderate rate of meiotic abnormalities among sexual (5%–31%) and apomictic (7%–11%) parents (Pessim et al., 2010, 2015). Hybrids originating from a single cross showed abnormal cells at a rate ranging from 16% to 52% (Pessim et al., 2010, 2015). The frequency of meiotic abnormalities found in *M. maximus* is lower than that reported in the tetraploid *Urochloa* ($2n = 4x = 36$) interspecific hybrids, which ranged from 18% to 82% (Risso-Pascotto et al., 2005; Mendes-Bonato et al., 2006, 2007; Fuzinatto et al., 2007). Pagliarini et al. (2008) found that the mean occurrence of meiotic abnormalities in five induced autotetraploid *U. ruziziensis* accessions ranged from 5% to 10% and only one accession reached 55% abnormalities. However, contrary to expectation, in all the autotetraploidized



	Advantages	Specific breeding goals	General breeding goals
Brachiaria	Good adaptability to acid and low fertility soils, easy multiplication by seeds, high capacity to compete with invasive plants and good animal performance when compared to native pastures	Resistance to spittlebugs, nutritional value, adaptation to acidic soils, seed production, seasonality, productivity, seed production, quality, vigor, persistence, and tolerance of shade	Total dry matter productivity, leaf dry matter productivity, good distribution of production throughout the year, nutritional value and forage quality (consumption, palatability, chemical composition, digestibility, and absence of anti-nutritional factors), adaptation to the environment of interest (soil, climate, and management), persistence and vigor (survival and income over time), resistance to pests and diseases, resistance/tolerance to abiotic factors (drought, flooding, and temperature), highly efficient and/or sustainable use of nutrients (e.g., biological nitrogen fixation), ease of propagation, speed of establishment, cycle, seed production, compatibility of association with other species
Elephant grass	High potential of forage production, adaptation to different ecosystems, good acceptance by animals, high leaf/stem ratio, high nutritional value of pasture, high animal performance, resistance to water deficit	Productivity, nutritional value, seed propagation, seed production and quality, adaptation to mechanical harvesting, adaptation to grazing, resistance to spittlebugs, tolerance of low temperature, silage, and energy	
Guinea grass	High potential of forage production, high nutritional value and acceptance by animals of different categories, easily propagated by seeds	Leaf productivity, resistance to *Bipolaris*, and seed production	

FIGURE 2 | Characteristics of brachiaria, Guinea grass, and elephant grass and breeding goals to improve their use as tropical forage grasses. The advantages and breeding goals are based on Machado et al. (2019). Source of the pictures: Embrapa.

accessions, chromosome pairing was preferentially bivalent (Pagliarini et al., 2008).

Regardless of the low rate of tetravalent formation in tetraploid and sexual *M. maximus* and *U. ruziziensis*, the issues of allele dosage and compatibility between apomictic and sexual genomes still remains unresolved, as explained by Gallais (2003) for recently doubled genotypes. Therefore, efficient cytogenetic identification of the best crosses at early stages would allow for the identification of the best and most cytogenetically stable parents and progenies. Consequently, all subsequent stages of breeding programs will certainly benefit from genomic prediction and the unbiased estimation of marker effects.

Availability of TFG Breeding Populations for GS

In practice, three populations must be defined for GS: estimation, validation, and breeding populations (Goddard and Hayes, 2007; Meuwissen, 2007). These populations may be as follows: i) physically distinct (three different populations), ii) with two simultaneous functions (only one population used for estimation and validation), or iii) with three simultaneous functions (only one population used for estimation, validation, and selection). **Figure 3** illustrates strategy ii.

Estimation Population

The estimation population is also called the discovery, training, or reference population. This dataset includes a large number of markers assessed in a moderate number of individuals (1,000 to 2,000 depending on the desired accuracy), which should have their phenotypes assessed for various traits of interest. Equations for predicting genomic values (random multiple regression) are obtained for each trait. These equations associate each marker or interval with its effect (predicted by RR-BLUP) on the trait of interest. The markers that explain the loci regulating the traits are identified in this population, and their effects are estimated. Recently, Lara et al. (2019) and Matias et al. (2019a) used estimation populations with 530 individuals of *M. maximus* and 272 individuals of *Urochloa* hybrids, respectively. Predictive abilities were lower than 0.4 for the evaluated traits. This indicates that factors affecting GS efficiency in TFG breeding, such as adequately sized training populations, still need to be improved.

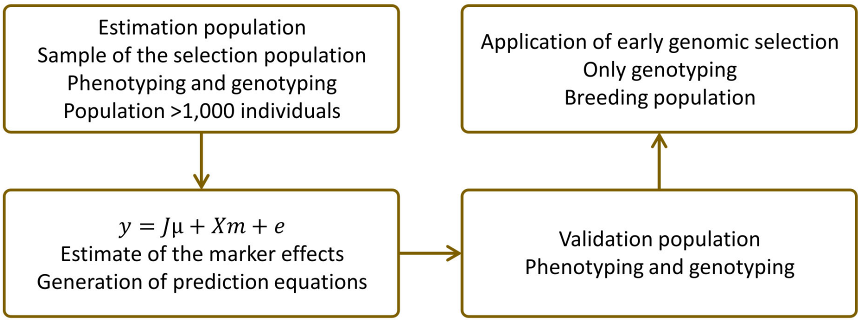

Estimation population
Sample of the selection population
Phenotyping and genotyping
Population >1,000 individuals
Application of early genomic selection
Only genotyping
Breeding population
$y = J\mu + Xm + e$
Estimate of the marker effects
Generation of prediction equations
Validation population
Phenotyping and genotyping

FIGURE 3 | Schematic application of genomic selection (GS) in a genetic improvement program (Resende et al., 2012).

Validation Population

When physically separated from the estimation population, this dataset is smaller than the discovery population and includes individuals that are assessed for SNP markers and various traits of interest. The equations for predicting genomic values are tested to verify their accuracy for this independent sample. To calculate the accuracy, genomic values are predicted, using the estimated effects from the estimation population and subjected to correlation analysis with the observed phenotypic values. As the validation sample is not involved in predicting the marker effects, errors from genomic values and phenotypic values are independent. Correlations between these values are predominantly genetic in nature and equivalent to the predictive ability ($r_{y\hat{y}}$) of GS in estimating phenotypes, which is given by the accuracy of the selection itself ($r_{q\hat{q}}$) multiplied by the square root of the heritability (h), or $r_{y\hat{y}} = r_{q\hat{q}}h$. Thus, to estimate the accuracy, one should obtain $r_{q\hat{q}} = r_{y\hat{y}}/h$. This method is valid when raw phenotypic values are used to calculate the correlations. When using genotypic values predicted based on phenotypes instead of raw phenotypic values, heritability should be replaced by the reliability of the prediction. In general, strategy ii is adopted according to a k-fold scheme for cross-validation. According to Meuwissen (2007), when dozens to hundreds of thousands of haplotypes are estimated, there is a risk of over-parameterization; in other words, errors in the data are explained by the marker effects. Cross-validation is therefore extremely important to address this problem.

Breeding Population

This dataset only contains the markers assessed in the candidates for selection, and the phenotypes do not need to be assessed in this population. Therefore, the prediction equations derived from the estimation population are used to predict the GVs or future phenotypes of the candidates for selection. The associated selection accuracy is calculated for the validation population.

In most TFG breeding programs, the training population is the same as, or part of, the breeding population, and this population may have experienced directional selection for many generations (Simeão-Resende et al., 2014). It is likely that validation may never exist for most breeding programs, firstly because cross-validation (Kohavi, 1995) has been the commonly used method and secondly because it is difficult to find out more than one ongoing breeding program per species and country.

Elephant grass (*Cenchrus purpureus*) is a tetraploid and allogamous species in which open pollinated divergent populations are easy to establish. The two main breeding strategies are: (i) recurrent selection (Reis et al., 2008); and (ii) clonal selection (Pereira et al., 2017; Machado et al., 2019). The recommended number of individuals in estimation populations for GS can be easily reached, in which the effective population size can be previously assumed and effectively estimated after genotyping. Validation can be performed by cross-validation and this may be extended to different environments or even to related populations of breeding programs in other countries.

Guinea grass (*M. maximus*) and brachiaria (*Urochloa*) breeding programs have recently adopted the full-sib reciprocal recurrent selection as a method, in which thousands of hybrids are generated annually (Barrios et al., 2013; Worthington and Miles, 2015). As the intrapopulation recurrent selection is feasible only on sexual populations, because there is no possible crossing between apomictic accessions, selection is performed on sexual individuals as a function of heterosis expressed in crossings with apomictic accessions. The schematic drawing of this procedure was presented for *M. maximus* (Simeão-Resende et al., 2004) and *Urochloa* (Jank et al., 2014). Based on this information, and the fact that these methods have been recently implemented, some limitations for the establishment of estimation populations in these genera must be discussed. Firstly, the N$_e$ of the sexual population is extremely low (N$_e$ <7) in *M. maximus* (Simeão-Resende et al., 2004; Lara et al., 2019), in tetraploid *U. ruziziensis* used as female parents in crossings with apomictic *U. brizantha* (Simeão et al., 2015), and in tetraploid sexual *U. decumbens* (Barrios et al., 2013). Secondly, albeit the genetic diversity within populations of apomictic accessions of *Urochloa* species is high (Vigna et al., 2011), the number of accessions used in crosses is low, because the crosses are performed based only on the adapted and agronomically selected apomictic individuals (Jones et al., 2021). Therefore, the N$_e$ is likewise low (<20). Thirdly, as a result of the intra or interspecific crosses, F$_1$ progeny segregates for mode of reproduction in which individuals in the progeny may vary from 99% apomictic to 99% sexual. While this procedure is efficient to explore the panmictic heterosis (Lamkey and Edwards, 1999) and can readily generate apomictic individuals that are potential cultivar candidates, these hybrid swarms do not constitute a breeding population *per se*. Therefore, the N_e among and within hybrid families must be adjusted and considered for efficient GS use in both genera. Using the equation $N_e = \frac{4N_f n}{n+1}$, in which N_f is the number of full-sib families and n the number of individuals per family, we may suggest the evaluation of 42 families and 10 individuals per family, a total of 420 genotyped individuals, to achieve N_e of approximately 152. For phenotyping, the total number may be approximately 1000 individuals (Resende, 2015). Finally, in practice, the TFG estimation population for GS in the genera *Megathyrsus* and *Urochloa* could be composed of sexual hybrids in an open pollinated population, followed by validation in apomictic hybrids.

Genetic Markers in Selection and Gene Discovery

The use of molecular genetic markers for selection and genetic improvement is based on the genetic linkage between these markers and a quantitative trait locus (QTL) of interest (Resende et al., 2013). Thus, linkage disequilibrium (LD) between markers and QTLs is essential for genomic selection from genomic information (Bourke et al., 2018). It must be made clear that a QTL refers only to the statistical association between a genomic region and a trait.

Recently, molecular genetic markers that consist of SNPs (based on the detection of polymorphisms that arise from a single nucleotide change in the genome) have been widely used in many species (Elshire et al., 2011). Generally, for a SNP to be considered genetically derived, the polymorphism must occur in at least 1% of the population (Resende, 2015). SNPs are the most common type of genomic variation and preferred over other genetic markers because of their abundance, ease of obtainment, and low genotyping cost. Thousands of SNPs can be used to cover the entire genome of an organism with markers not more than 1 cM apart from each other.

LD analysis is based on LD between a marker and a QTL in the whole population and not only within a family, as performed in linkage analysis (Würschum, 2012). For this to occur, the marker and QTL must be closely linked. When this occurs, the association between them is a property of the entire population and persists for many generations.

Association analysis is used for fine mapping and is based on population-level LD (Resende et al., 2013). Linkage can occur when the gene directly affects a trait, and when there is an LD between the marker and the gene controlling the trait. In the first case, the effect of the gene is directly evaluated, and the marker is classified as functional. The functional mutations are known as quantitative trait nucleotides (QTNs). In the second case, the linkage test requires LD between the marker and QTL. When a mutation occurs on a given chromosome, it creates a haplotype with adjacent loci on the chromosome. In the subsequent generations, this mutation tends to occur within the same haplotype unless there is recombination, which creates the LD used for association mapping (Resende et al., 2013).

In recent times, efforts to use molecular markers in genetic improvement research have evolved into two approaches: (1) GWAS for QTL identification and mapping; and (2) genome-wide selection (GWS) or GS (Resende, 2015). GS was proposed by Meuwissen et al. (2001) to increase breeding efficiency and accelerate genetic improvement. GS emphasizes the simultaneous prediction (without the use of significance tests for individual markers) of the genetic effects of thousands of SNP markers dispersed throughout the genome of an organism to capture the effects of all loci (both small and large effects) and identify the overall genetic variation of a quantitative trait (Resende and Alves, 2020). In this case, the sum of the estimated genetic effects of the markers present in an individual provides the genetic value of the individual for selection purposes. Meuwissen et al. (2001) obtained a complete array of estimates of haplotype effects using the ridge regression best linear unbiased prediction (RR-BLUP), BayesA, and BayesB methods. RR had already been used by Whittaker et al. (2000) for marker selection. Haley and Visscher (1998) suggested the name GS for selection on a whole genome scale (Resende and Alves, 2020).

The conceptual development of GS coincides with the technology associated with SNPs, which is accurate and relatively affordable. GS uses the associations between many SNP markers throughout the genome along with phenotypes and takes advantage of LD between markers and QTLs in close linkage (Resende et al., 2013). The predictions derived from phenotypes and SNP genotypes with high density in a generation are thus used to obtain genomic values (GVs) for individuals in any subsequent generations based on their genotypic markers, in which the genetic effects have been estimated.

When LD between markers is incomplete, the joint allele frequencies for the two loci can change markedly across generations, thereby leading to changes in haplotypes. In this case, the marker effects would need to be estimated again to maintain the accuracy of GS for various generations (Dekkers, 2007). In the case of a complete or close LD, the estimated effects remain constant across different families and generations within the same environment.

The number of markers used are directly associated to the genome size, extent of LD and population structure (Ballesta et al., 2020). Larger genomes, rapid breakdown of LD and greater effective population size imply that a higher density of SNP loci would be needed.

Factors Affecting the Efficiency of GS

The accuracy of GS depends on five factors (Resende et al., 2014): i) heritability of the trait; ii) number of loci regulating the trait (also given by $2N_eL$) and the distribution of their effects; iii) number of individuals in the discovery population; iv) effective population size (N_e); and v) marker density, which depends on the number and genome size (L, in Morgans). The first two factors are beyond the breeder's control, and the latter three factors can be modified by the breeder to increase the accuracy of GS.

An increase in the selection efficiency using GS can be achieved by changing the four components of the expression for genetic progress, given by $SG = (kr_{g\hat{g}}\sigma_g)/T$, where k

is the standardized selection differential (dependent on the selection intensity), $r_{g\hat{g}}$ is the accuracy of selection, σ_g is the genetic standard deviation (genetic variability) of the trait in the population, and T is the time required to complete a selection cycle.

In perennial and vegetatively propagated plant species, the benefit of GS is from an increase in $r_{g\hat{g}}$ and a reduction in T. The increase in $r_{g\hat{g}}$ is due to the use of an actual kinship matrix (Resende, 2007). This increase depends on the size of the estimation population and marker density. Factor T is greatly reduced by GS because genomic prediction and selection can be performed at the seedling stage. Thus, even if $r_{g\hat{g}}$ shows the same magnitude as obtained by phenotypic selection, GS is still better than selection based on phenotypes due to the reduction of T.

Inferring the Quality and Efficiency of GS

The quality of GS is inferred by correlation and regression among the predicted genetic values and phenotypes in the validation population, as well as by the accuracy of the prediction. The correlation and regression coefficients involving observed and predicted values are practical measures of the ability of the methods to make predictions that are accurate and unbiased, respectively. Correlation provides predictive ability, which is equivalent to the product of accuracy and the square root of heritability (Resende et al., 2014). The regression coefficient is algebraically equal to 1. Regression coefficients of less than 1 indicate that the genetic values are overestimated and exhibit greater variability than expected; coefficients greater than 1 indicate that the estimated genetic values exhibit variability lower than expected. A lack of bias is important when selection involves individuals from many generations using the estimated marker effects from a single generation. Regression coefficients near 1 indicate that the assessments are unbiased and effectively predict the actual magnitudes of differences among the individuals assessed.

The expected value of the regression coefficient is 1, which indicates an unbiased prediction. Thus, the regression coefficient can also be used to estimate the heritability of the markers (Resende et al., 2013). Various heritability values are assessed, and those that provide a regression equal to 1 should be selected as the best estimate. If the regression yields a result less than 1, the magnitude of the assessed heritability value is too high and should be reduced until the regression coefficient is converged to 1. If the regression yields a result greater than 1, the magnitude of the assessed heritability value is too low and should be increased until it converges to 1.

CURRENT GENOMIC RESOURCES FOR TFG BREEDING

The availability of a large number of high-quality single-nucleotide polymorphisms (SNPs) that can be genotyped at a reasonable cost is a prerequisite for implementing GS in a crop of choice (Hayes et al., 2013). Gold-standard methods for SNP discovery rely on resequencing individual samples at minimum coverage and mapping reads to a reference genome (McKenna et al., 2010); this is also true for low-coverage, genotyping-by-sequencing (GBS) methods (Elshire et al., 2011). Although, reference-free pipelines are available for GBS, such as UNEAK (Lu et al., 2013), reference-based methods allow for more informed decisions during the selection of high-quality SNPs (McCormick et al., 2015). Alternative methods have been recently tested and reported in *Urochloa*, for which the GBS-SNP-CROP pipeline (Melo et al., 2016) was used to generate a "mock" reference from GBS data (Matias et al., 2019b). This led to the discovery of a larger number of biallelic SNPs, when compared to mapping reads to the available genome of the closest related species (*Setaria viridis* and *Setaria italica*).

Assembly quality, which is measured by its accuracy, haplotype phasing, and contiguity, is an important factor influencing marker discovery, GWAS, and GS. In tetraploid blueberries, the selection of probes for targeted SNP genotyping, investigation of the genetic architecture of fruit traits, identification of candidate genes, and genomic prediction benefited from a more complete, chromosome-scale, haplotype-phased genome assembly (Benevenuto et al., 2019). A huge reduction in sequencing costs and increased throughput brought about by second-generation sequencing have allowed unprecedented access to crop genomic information (Shamshad and Sharma, 2018). However, the genomic complexity of most TFG species is still challenging and has kept them recalcitrant to sequencing efforts targeting reference-grade assemblies using mainly second-generation short reads. Forage grass species with the largest breeding programs in Latin America are mostly polyploid and highly heterozygous and have genomes with a high repeat content (**Table 1**). Third-generation long reads can circumvent some of these problems, and, combining with a technology such as optical mapping or chromosome conformation capture can potentially allow chromosome-scale reference-grade assemblies (Belser et al., 2018; Shi et al., 2019). The use of these tools to accelerate genomic research on TFGs is promising and can thus advance the use of GS in breeding programs. We argue that the availability of high quality genome assemblies is the starting point for the development of genotyping systems that will be useful for successfully deploying GS in TFG breeding programs.

TABLE 1 | Reproductive system and genomic information of economically important tropical forage grasses used in livestock production.

Scientific name	Predominant reproductive system	Genome size (Gpb)	Chromosome number and ploidy level	WGS[#]
Urochloa brizantha	Apomictic	1.4	2n = 4x = 36	No
Urochloa humidicola	Apomictic	1.9	2n = 6x, 9x = 36 to 54	No
Urochloa decumbens	Apomictic	1.6	2n = 4x = 36	No
Urochloa ruziziensis	Sexual	0.6	2n = 2x = 18	Yes
Megathyrsus maximus	Apomictic	1.0	2n = 4x = 32	No
Cenchrus purpureus	Sexual	2.1	2n = 4x = 28	Yes

[#] *WGS, availability of whole-genome sequence.*

Genome Assembly in *Urochloa* spp.

Urochloa P. Beauv. Grasses, many of which were previously included in *Brachiaria* (Trin.) Griseb., are the most widely used forage in Latin America. The main species are *U. ruziziensis*, *U. brizantha*, *U. decumbens*, and *U. humidicola*. The first genome assembly of a forage grass species in the genus *Urochloa* has recently been reported for *U. ruziziensis* (Worthington et al., 2020). The diploid genotype CIAT26162 was sequenced using short reads (approximately 100 × coverage). Assembly and scaffolding were performed and PacBio RSII reads were used for gap filling. This is an invaluable resource for an orphan species with no previously published genome information available. However, it also highlights the huge challenges in the assembly of highly heterozygous, repetitive genomes with short-read technologies. The publicly available assembly was fragmented into 102,577 scaffolds, with an N50 of 27.8 kbp. Completeness metrics based on benchmarking universal single-copy orthologs (BUSCOs) indicate that 86.7% are complete (1,248 out of 1,440 in the Embryophyta_odb9 dataset), suggesting that there is still room for improvement. Because of the lack of linkage maps for an intraspecific diploid cross in *U. ruziziensis*, which could be used to cluster and order scaffolds based on linkage groups, the study relied on anchoring scaffolds based on synteny with the *Setaria italica* genome to obtain pseudo-molecules. This potentially affected the anchored assembly due to undetected chromosomal rearrangements that might be present between *S. italica* and *U. ruziziensis,* as phylogenetic information and evolutionary relationships were not taken into account in the anchoring process.

Genomic Resources for Guinea Grass (*M. maximus*)

The placement of Guinea grass in the phylogeny of Paniceae has changed over the years, however, plant breeders, seed producers, and farmers in Latin America mostly refer to it as *Panicum maximum*, or "Panicum" as a common name for the species. There is no reference genome assembly available for *M. maximus*. The fact that Guinea grass was once included in *Panicum* and is still mostly regarded as a *Panicum* species in Latin America may lead to the assumption that the available genome assemblies for panic grasses such as *P. hallii* (two publicly available chromosome-scale assemblies; Lovell et al., 2018) and *P. milliaceum* (two chromosome-scale assemblies; Shi et al., 2019; Zou et al., 2019) would provide suitable shortcuts for the development of genomic resources for *M. maximus*. Lara et al. (2019) illustrated how this approach might limit the genomic distribution and number of SNPs available when genotyping *M. maximus* breeding populations using *Panicum* genome assemblies as references. In this study, a multi-parental population of *M. maximus* half-sib progenies was genotyped using GBS, and six different assemblies were tested as references for read mapping. Two of them comprised the genome assemblies for *P. hallii* and *P. virgatum*, while the remaining included genome assemblies for *Setaria* (*S. italica* and *S. viridis*) and transcriptome assemblies for *M. maximus*. The alignment rates ranged between 19.05% for *P. hallii* and 24.24% for *M. maximus* transcriptomes, showing that a very large proportion of reads were not used for SNP discovery and genotyping. The sets of allele-dosed SNPs containing up to 5% of missing data from each of the reference assemblies ranged between 5,032 for one of the *M. maximus* transcriptomes and 8,112 for *S. viridis*. Although, this was the first report on the development and assessment of prediction models, considering the allele dosage, for GS in *M. maximus*, increased marker density and prediction accuracies may be expected when a high-quality genome assembly for the species is available for SNP discovery and genotyping.

Genomic Resources and Genome Assemblies for Elephant Grass (*C. purpureus*)

Elephant grass (*C. purpureus* Schumach. Morrone, syn. *Pennisetum purpureum* Schumach.), also called napier grass, merker grass, or Uganda grass, is a tropical grass native to Eastern and Central Africa. It is an allotetraploid species with a chromosome constitution of 2n = 4x = 28 A'A'BB and an average amount of DNA per G1 nucleus of 4.58 pg (Hanna, 1981; Taylor and Vasil, 1987). Elephant grass is used as animal fodder and is a promising lignocellulosic biofuel feedstock due to its high growth rate, high biomass yield, and persistence (Morais et al., 2009; Singh et al., 2013; Daud et al., 2014; Rocha et al., 2019).

Wang et al. (2018) conducted a genome survey of elephant grass and estimated its genome size to be 2.01 Gb with 71.36% of repetitive elements and a heterozygosity of 1.02%. A total of 114.36 Gb of raw data, (approximately 57-fold coverage) was generated using the Illumina HiSeq sequencing platform for the Zise genotype (purple elephant grass). A partial draft assembly was obtained using SOAPdenovo. As expected for such a complex genome, this effort allowed a preliminary investigation of the repetitive content of elephant grass and the identification of thousands of genomic SSR markers, 30 of which were tested for genotyping a set of 28 elephant grass accessions. Another genome survey of Merkeron and UF1 cultivars was conducted by Paudel et al. (2018), and they also developed a high-density linkage map using GBS.

More recently, two chromosome-scale assemblies of elephant grass were reported (Yan et al., 2020; Zhang et al., 2020). The initial assembly of the cv. Purple (Yan et al., 2020) was obtained using Nanopore long reads and was then polished with Illumina short reads and scaffolded with Hi-C data. Approximately 2,000 contigs were grouped and oriented into 14 chromosome-scale scaffolds, with a total size of 1.9 Gbp, 66.3% of which were annotated as repetitive elements. The assembly showed high contiguity at the contig level (N50 1.8 Mbp) and 97.8% completeness using BUSCOs. The predicted protein-coding gene set also showed high BUSCO completeness (97.1%).

The second assembly (Zhang et al., 2020, available as a preprint), for the CIAT6263 accession, was obtained with Nanopore reads and ultra-long reads, which were assembled to a total size of 2.07 Gbp with a contig N50 of approximately 2.9 Mbp. The authors used a combination of BioNano optical maps and Hi-C to obtain a final chromosome-scale assembly of approximately 2 Gbp in 14 pseudomolecules. This assembly was also 97.8% BUSCO complete, and repetitive elements accounted for 60.7% of the genome. The difference in repetitive content between the two assemblies might be explained by the different methods applied for *de novo* identification of repeats.

These two assemblies place elephant grass in a unique position among the TFGs. Current research trends indicate the benefits of having not only one but multiple genome assemblies for a species of interest (Della Coletta et al., 2021). Large sequencing projects now target pan-genomes, instead of a single reference that does not capture the full diversity of a species (Zhao et al., 2018). Future efforts of variant discovery and association of phenotypes with genomic locations will be possible using these assemblies as anchors for read mapping, opening up the possibility of revisiting previous datasets that were generated without a reference genome. However, while both efforts resulted in chromosome-scale scaffolds, the fact that the SMARTDENOVO assembler does not generate haplotype-resolved contigs indicates that the high heterozygosity levels of elephant grass were not represented in the assemblies.

STATISTICAL METHODS IN GS

An ideal method for estimation of SNP effects in GS should accommodate the genetic architecture of the trait in terms of genes of small and large effects and their distributions, regularize the estimation process in the presence of multicollinearity and a larger number of markers than individuals, using shrinkage estimators, and perform the selection of covariables (markers) that affect the trait under analysis. The main problem with GS is the estimation of a large number of effects from a limited number of observations and the collinearities arising from LD between the markers. Shrinkage estimators deal with this appropriately by treating the effects of markers as random variables and estimating them simultaneously (Resende, 2007; Resende et al., 2008; Azevedo et al., 2015; Resende and Alves, 2020).

If the effects of markers are taken as fixed, it is not possible to consider the covariance between the effects of the markers. With a high density of markers, more than one marker will be in LD with a segregating QTL, which will result in covariance between the marker effects. Most markers will have no effect on a trait, and the estimated effects of these empty markers will be false. This problem is greater when the markers are considered to have fixed effects, because in that case, these pseudo effects will not shrink toward zero (Resende and Alves, 2020).

In the context of marker-assisted selection (MAS) and genomic prediction, the method of least squares (LS) has serious drawbacks. According to Gianola et al. (2003), the selection index (calculated as the regression involving molecular scores) presented by Lande and Thompson (1990) for MAS fails when formulated vectorially. This failure occurs because the covariance matrix for the molecular scores is singular, as the distribution of fitted regression values is defined only in the p-dimensional space (number of covariables) and not in the n-dimensional space (number of individuals with molecular scores). Therefore, the selection index leads to an infinite number of solutions.

Another difficulty arises when the number of markers is equal to or greater than the number of genotyped individuals. In this case, the collinearity of the predictor variables causes parametric identification problems, and thus, some type of dimensional reduction, such as singular value decomposition, should be used. Another problem is the inadmissibility (unable to provide the minimum mean square error) of LS estimators, a result that collapses estimates by LS and generalized LS (GLS). Thus, the LS method is not recommended for the MAS and GS analyses. In summary, the LS method is inefficient because it is impossible to simultaneously estimate all effects when the number of effects to be estimated is greater than the number of data points; thus, estimating one effect at a time and testing its significance leads to an overestimation of significant effects, and the accuracy of the method becomes low. In addition, only QTLs with large effects will be detected and used, and consequently, not all genetic variations will be captured by the markers. The LS method assumes *a priori* QTL distribution, with an infinitely large variance that disagrees with the known total genetic variance.

Because the number of markers in GS is greater than the number of individuals, there is a lack of degrees of freedom to estimate the effects of all markers. A solution to this problem is to use the RR method (Whittaker et al., 2000) or to consider the marker effects as random instead of fixed. Fitting random effects does not expend degrees of freedom, and the effects of all markers can be estimated simultaneously. This method leads to RR-BLUP, which considers the effects of QTLs with normal distributions and equal variance through chromosomal segments.

The main problem for GS is estimating a large number of effects from a limited number of observations, in addition to collinearities resulting from LD between markers. The shrinkage estimators adequately address this issue by treating the marker effects as random variables and estimating them simultaneously (Resende et al., 2008).

The main methods for GS are based on Random Regression and can be divided into three major classes: explicit, implicit, and dimensionally reduced regression. In the first class, the RR-BLUP, Lasso, BayesA, and BayesB methods stand out among others. In the class of implicit regression, the Reproducing Kernel Hilbert Spaces (RKHS) method, which is semiparametric, is the most popular. The Independent Components, Partial Least Squares, and Principal Components stand out among the regression methods with dimensional reduction. Two new non-parametric approaches for GS proposed by Resende (2015) and Lima et al. (2019a,b) have proven to be efficient (Resende and Alves, 2020) and are called triple categorical regression (TCR) and Delta-p, respectively.

The explicit regression methods are divided into two groups: (i) penalized estimation methods (RR-BLUP, Lasso) and (ii) Bayesian estimation methods (including BayesA, BayesB, fast BayesB, BayesCπ, BayesDπ, Bayesian regression, BayesRR, BayesRS, BLasso, and IBLasso). Among these, the best and most effective in practice are RR-BLUP and BayesB (Visscher et al., 2006, 2008, 2010; Mrode et al., 2010; Mrode, 2014). Each method without covariate selection has a similar method with covariate selection. Thus, the following are the pairs without -

with covariate selection: BayesA - BayesB; BayesRR - BayesCπ; BLasso - IBLasso (Resende and Alves, 2020).

The RR-BLUP is a model equivalent to genomic best linear unbiased prediction (G-BLUP), which is the BLUP method at an individual level with the genealogical relationship matrix A changed to a genomic relationship matrix G. The equivalence between these two methods was given by Habier et al. (2007) and Van Raden (2008). The G-BLUP and RR-BLUP are equivalent when the number of QTLs is large, and no major QTL is present. The use of matrix G based on markers had already been established by Bernardo (1994); Nejati-Javaremi et al. (1997), and Fernando (1998). A single-step BLUP simultaneously using phenotypic, genotypic, and genealogical information, called H-BLUP single-step, was proposed by Misztal et al. (2009), using an H matrix composed of the A and G matrices (Resende and Alves, 2020). The idea of H-BLUP was given by Fernando (1998).

The traditional quantitative genetics rely on random mating populations. Nowadays, with the availability of SNP markers, random mating does not need to be assumed, because breeders can track the transmission of chromosomal segments. Another assumption is linkage equilibrium in the breeding population. Once linkage among markers is accounted for in the G coefficient matrix in RR-BLUP, this circumvents the need to assume linkage equilibrium (Resende and Alves, 2020).

A refinement of GS can be achieved by using QTNs instead of SNPs. The evolution of genomic technology is predictable and the causal mutation of a genetic variation at the nucleotide level (QTN) can be accessed soon. Thus, GS can be improved by the direct use of QTNs instead of SNPs. The use of QTNs will bring the following advantages (Weller, 2016): GS will not depend on the LD as the QTN will be accessed directly and not via markers and, this will increase the robustness of the genomic prediction, which will also be useful in the long run; the genomic prediction may have transferability across different populations and species in the same genus; genomic prediction will use specific QTNs for each trait, unlike G-BLUP by means of SNPs, which uses the same G relationship matrix for all traits; the multiple-trait selection indices will directly weigh the QTNs and not the phenotypic traits; GS may use a smaller number of generations (only the last ones) for the composition of the G matrix, which will bring greater genetic gain and lesser mass of data to be processed; the allele frequencies of the QTNs will be accessed directly and not through LD with SNPs (Resende and Alves, 2020).

Single-Environment RR-BLUP and G-BLUP Models

The parametric regression model for a single environment j^{th} ($j = 1, \ldots, m$) is defined as $y_j = 1_{nj}\mu_j + X_j\beta_j + \varepsilon_j$, where the vector y_j represents nj independent centered observations of the response variable in the j^{th} environment; 1_{nj} is a vector of ones of order nj; μ_j is the overall mean of the j^{th} environment; X_j is the matrix for the p centered and standardized molecular markers in the j^{th} environment; vector β_j represents the effect of each of the p markers in the j^{th} environment, and ε_j is the vector of random errors in the j^{th} environment with normal distribution and common variance $\sigma^2_{\varepsilon_j}$. The RR-BLUP assumes that the effects of the markers have a multivariate normal distribution $\beta_j \sim N(0, I\sigma^2_{\beta\ j})$.

Assuming that the effects of the markers β_j and ε_j are independent, and that $u_j = X_j\beta_j$, then the above model for the j^{th} environment can be written as $y_j = 1_{nj}\mu_j + u_j + \varepsilon_j$, where u_j, and ε_j are independent random variables with $u_j \sim N(0, \sigma^2_{u_j}K_j)$, and $\varepsilon_j \sim N(0, \sigma^2_{\varepsilon}I)$, respectively; $\sigma^2_{u_j}$ is the variance of u_j (to be estimated), and K_j is a symmetric matrix representing the covariance of the genetic values. Thus, for a single-environment where the K_j is of the linear form $K_j = G_j = X_jX_j^{'}/p$ the G-BLUP is equivalent to RR-BLUP (Van Raden, 2008).

Genetic Parameterization of Additive, Dominance, and Total Genotype Effects

Additive Model

The following linear mixed model can be fitted to estimate the marker effects $y = J\mu + Xm + e$, where y is the vector of phenotypic observations, μ is the vector of the fixed effect of the general mean, m is the vector of random marker effects and e is the vector of random residuals. J and X are the incidence matrices for μ and m, respectively. The incidence matrix X contains functions of the values 0, 1, and 2 for the number of alleles for the marker (or the supposed QTL) in a diploid individual. A similar coding method uses the values of -1, 0, and 1. The genomic mixed-model equations for predicting m using the RR-BLUP method are equivalent to $\begin{bmatrix} J'J & J'X \\ X'J & X^{'}X + I\frac{\sigma^2_e}{(\sigma^2_a/n_Q)} \end{bmatrix}\begin{bmatrix} \widehat{\mu} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} J'y \\ X'y \end{bmatrix}$. The total additive genomic value for an individual j is given by $AGV_j = \hat{y}_j = \sum_i X_i\hat{m}_i$, where X_i is equal to 0, 1, or 2 for the genotypes mm, Mm, and MM, respectively, for biallelic and co-dominant markers, such as SNPs.

These prediction equations assume *a priori* that all loci explain equal amounts of genetic variation. Thus, the genetic variation explained by each locus is given by σ^2_a/n_Q, where σ^2_a is the total genetic variation and n_Q is the number of loci (when each locus is perfectly marked by a single marker), which can be given by $n_Q = 2\sum_i^n p_i\left(1 - p_i\right)$, where p_i is the frequency of the allele of the type M in locus i. The genetic variation σ^2_a can be estimated by restricted maximum likelihood (REML) on the phenotypic data in a traditional manner or by the variation among markers or QTL chromosomal segments.

There is no need to use the kinship matrix with the RR-BLUP method. The pedigree-based kinship matrix used for traditional BLUP was replaced by a kinship matrix estimated by the markers. This kinship matrix is a function of *X'X* present in the equations of the mixed model described above. This procedure is more efficient because it effectively captures the kinship produced for each individual and not an average kinship matrix associated with the pedigree.

The parameterization of the incidence matrix X uses the values 0, 1, and 2 for the number of alleles of a marker (or supposed QTL) in a diploid individual and $2p$ for individuals with missing marker data. These values should be centered around 0 so that the effects of co-dominant markers are effects of allelic substitution with a mean of 0 in the population. In this case, assuming Hardy-Weinberg equilibrium, the additive genetic variation of the trait in the population is equal to $\sigma_a^2 = 2\sum_i^n p_i\left(1-p_i\right)\sigma_m^2$. Thus, the values of X_i should be replaced by 0 – 2p, 1 – 2p, and 2 – 2p, to obtain a variable with a mean of 0. Thus, with centralization, $n_Q = 2\sum_i^n p_i\left(1-p_i\right)$, should be used for the RR-BLUP method, and the additive genetic effects of individuals are given by $\hat{a} = X\hat{m}$.

Additionally, the data for markers in matrix X can be standardized as follows for each matrix element X_i corresponding to locus i:

$X_i = (0 - 2p_i)/(\mathrm{Var}(X_i))^{1/2}$ if the individual is homozygous for the first allele (mm).
$X_i = (1 - 2p_i)/(\mathrm{Var}(X_i))^{1/2}$ if the individual is heterozygous (Mm).
$X_i = (2 - 2p_i)/(\mathrm{Var}(X_i))^{1/2}$ if the individual is homozygous for the second allele (MM).
$X_i = 0$ if the individual has missing marker data. The quantity p_i is the frequency of the second marker allele.

The cut-off point for including a marker in the analysis can be determined by the minor allele frequency (MAF), which is a measure related to the variation of alleles in the population, given by $MAF = (1/2N)^{1/2}$ which comes from the standard deviation of a proportion, given by $(pq)^{1/2}/(2N)^{1/2}$, where N is the number of genotyped individuals, meaning that the lower the N value, the greater the MAF needs to be for accurate estimation of the marker effect (Resende, 2015; Resende and Alves, 2020).

Coding and Additive Kinship Matrix in Polyploids

The incidence matrix X contains the values 0, 1, 2, 3, and 4 for the number of alleles for the marker (or the supposed QTL) in a tetraploid individual. Analysis by G-BLUP uses the kinship matrix given by $G = \frac{(X^*X^{*'})}{[2\sum_i^n p_i(1-p_i)]^{1/2}}$, where X* is the X matrix after centralization.

Additive-Dominance Model

According to the marker model $y = J\mu + W\alpha + S\delta + e$, (where coefficients of α and δ are the additive and dominance effects, respectively), the most appropriate parameterization to estimate the effects on the additive-dominance model (Vitezica et al., 2013; Azevedo et al., 2015) is:

Additive effects (W):

$$W = \begin{cases} If\ MM;\ 2 \rightarrow 2-2p = 2q \\ If\ Mm;\ 1 \rightarrow 1-2p = q-p \\ If\ mm;\ 0 \rightarrow 0-2p = -2p \end{cases}.$$

The values of W must be centered at zero so that the effects of the codominant markers are effects of allelic substitution (α) with a mean of 0 in the population.

Dominance effects (S):

$$S = \begin{cases} If\ MM;\ 0 \rightarrow -2q^2 \\ If\ Mm;\ 1 \rightarrow 2pq \\ If\ mm;\ 0 \rightarrow -2p^2 \end{cases}.$$

G-BLUP for the Additive-Dominance Model

The individual mixed model is given by $y = J\mu + Za + Zd + e$, where a is the additive genetic vector of the individuals, and d is the dominance genetic vector of the individuals; $a \sim N(0, G_a\sigma_a^2)$, $d \sim N(0, G_d\sigma_d^2)$, and $e \sim N(0, I\sigma\,_e^2)$.

The mixed-model equations for the additive-dominance model are equivalent to $\begin{bmatrix} J'J & J'Z & J'Z \\ Z'J & Z'Z + G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'J & Z'Z & Z'Z + G_d^{-1}\frac{\sigma_e^2}{\sigma_d^2} \end{bmatrix}\begin{bmatrix} \widehat{\mu} \\ \hat{a} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} J'y \\ Z'y \\ Z'y \end{bmatrix}$, where $G_a = \frac{WW'}{\sum_{i=1}^n (2p_iq_i)}$, and $G_d = \frac{SS'}{\sum_{i=1}^n (2p_iq_i)^2}$; p_i and q_i are the allelic frequencies; $\sigma_a^2 = \sum_{i=1}^n\left[2p_i(1-p_i)\right]\sigma_\alpha^2$, and $\sigma_d^2 = \sum_{i=1}^n\left[2p_i(1-p_i)\right]^2\sigma_\delta^2$; and σ_a^2 and σ_d^2 are the additive and dominance genetic variances, respectively.

Adjusting an individual genomic model is equivalent to adjusting an individual traditional model but with the pedigree-based matrices A and D replaced by the genomic kinship matrices G_a and G_d for additive and dominance effects, respectively.

H-BLUP and Single-Step BLUP

In a simultaneous analysis of genotyped and non-genotyped individuals via G-BLUP, for a global evaluation of the three classes of individuals in a single step, the same additive model $y = J\mu + Za + e$ can be fitted with one alteration (replacing matrix G with matrix H) to the mixed-model equations, according to Misztal et al. (2009) $\begin{bmatrix} J'J & J'Z \\ Z'J & Z'Z + H^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix}\begin{bmatrix} \widehat{\mu} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} J'y \\ Z'y \end{bmatrix}$.

Matrix H includes both the relationships, based on pedigree (A) and differences between those and the genomic relationships (A_δ), such that $H = A + A_\delta$. Thus, H is given by $H = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & G \end{bmatrix} = A + \begin{bmatrix} 0 & 0 \\ 0 & \mathrm{G{-}A_{22}} \end{bmatrix}$, where the subscripts 1 and 2 represent non-genotyped and genotyped individuals, respectively.

The inverse of H, which allows simpler calculations, is given by $H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & G^{-1} + A^{22} - A_{22}^{-1} \end{bmatrix}$, where A_{22}^{-1} is the inverse of the kinship matrix based on pedigree for only genotyped individuals.

From the estimation of genetic values ($\hat{a}$) by G-BLUP, the estimated marker effects ($\hat{m}$) can be obtained by: $\hat{m} = (X'X)^{-1}X'\hat{a}$. Models with dominance effects (d) can also be fitted.

Another important application of this analysis is the estimation of total heritability explained by all the markers simultaneously. With the kinship matrix given by

$G = (XX')/\left[2\sum_i^n p_i(1-p_i)\right]$, total heritability can be estimated by REML using the mixed-model equations to estimate the variance components σ_a^2 and σ_e^2. The elements of matrix G represent the average multilocus kinship and are given by $G_{jk} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}$. Another favorable feature of G-BLUP is the possibility of directly estimating (by prediction error variance (PEV)) the accuracy of GS. For individuals with known phenotypes, this accuracy is valid for the estimation population without cross-validation. In G-BLUP, the phenotypes of the validation population are replaced by missing data. Therefore, individuals from this validation population will have a validated accuracy estimate.

Models at the level of individuals, including genotype × environment (*ae*) interactions, can also be fitted if there are related individuals within the same environment and across environments. In this case, the model is equal to $y = Wb + Za + Zae + e$, where *ae* is the vector of effects from the interaction between additive genetic effects and environmental effects (random), and Z is the incidence matrix for a and ae. The mixed-model equations for predicting a and ae using the BLUP method are

$$\begin{bmatrix} W'W & W'Z & W'Z \\ Z'W & Z'Z + G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'W & Z'Z & Z'Z + G_d^{-1}\frac{\sigma_e^2}{\sigma_{ae}^2} \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{a} \\ \widehat{ae} \end{bmatrix} = \begin{bmatrix} W'y \\ Z'y \\ Z'y \end{bmatrix},$$

where $G_{ae} = G$ for pairs of individuals in the same environment, and $G_{ae} = 0$ for pairs of individuals in different environments. The variance of the interaction between the additive genetic and environmental effects is denoted by σ_{ae}^{2}.

Additive, Dominance, and Total Genotype Effects in Polyploids

As SNP markers are biallelic, the inference of dosage allelic effect is dependent on the genetic effects of interactions (Gallais, 2003; Mackay et al., 2019). Exclusive additivity may create more classes of genotypic values than any other first-degree interaction among alleles and must be studied using allele dosage in GS. In autotetraploids or populations derived from their "allotetraploids," such as those evidenced in *M. maximus* and *U. ruziziensis*, the additive genetic variance and NSH cannot be estimated based solely on testing half-sib progenies or regression of offspring on the progenitors, because half-sib families may have fractions of the dominant genetic variance (Gallais, 1989). G-BLUP analysis of half-sib polyploid data allows the estimation of broad-sense heritability (BSH) using the information of all genetic relationships available in the kinship marker matrix since some "identity-by-state" dominance relationships allow the estimation of dominance effects, which along with estimated additive genetic effects, provide the estimation of the total genotypic value and then the BSH. This procedure was used in *M. maximus* hybrids (Lara et al., 2019) using a low number of parents and in *Urochloa* interspecific hybrids (Matias et al., 2019a).

Experimental crossing that provides simultaneous half-sib and full-sib progenies should be preferentially designed to estimate additive and dominance genetic effects simultaneously (Simeão-Resende et al., 2004), aiming at the total genotypic-genomic value prediction. In this case, there is more information about dominance relationships, thus generating better estimates. The BSH is estimated by the additive-dominance model, in which g = a + d and var(g) = var(a) + var(d). Making estimations of GEBV and GETV simultaneously in full-sibs and half-sibs with some progenitors in common in a training/validation population will allow the summation of the family effect in both predictions as well as the prediction of crosses that have not been performed.

In the case of similar magnitudes of NSH and BSH, there is no need for dominance adjustment, and only half-sibs can be used. In *M. maximus*, the estimated NSH and BSH for important traits showed a remarkably low and high magnitude, respectively, based on the use of phenotypic data (Simeão-Resende et al., 2004) or genomic data (Lara et al., 2019). In this species, GS based on additive and dominant effects needs to be performed to obtain the highest levels of genetic gain. In this way, it is important to work with tetrasomic inheritance more than disomic inheritance (Lara et al., 2019) and more genetically diverse synthetic populations to elevate the heritability and accuracy of GEBV prediction. The higher dominance effect evidenced in tetraploid *M. maximus* cannot be simply extended to other species unless the effect is previously known, or simply tested by different models of GS. de Bem Oliveira et al. (2019) predicted GEBV in blueberries by comparing diploid (data coded as 0, 1, and 2), tetraploid (data coded as 0, 1, 2, 3, and 4), and continuous (data coded as continuous parameterization assuming values between 0 and 1 and a cumulative additive effect) data models at the individual level. The researchers concluded that the use of continuous data generated estimated genetic gain values that were not significantly different from the best models of all traits. As diploid and tetraploid inferences of data did not affect the predictive ability, we can infer that simplified models can perform adequately.

Ridge, Bayes, and Lasso Methods

Bayesian methods are associated with systems of nonlinear equations, and non-linear predictions can be more efficient when the QTL effects are not normally distributed owing to the presence of genes with major effects. The linear predictions associated with RR-BLUP assume that all markers with the same allele frequency contribute equally to genetic variation (lack of genes with major effects). In Bayesian estimation, the shrinkage of effect estimates for the model is controlled by the *a priori* distribution assumed for these effects. Different distributions produce different shrinkages. Methods for penalized and Bayesian estimation may include (BayesB, Fast BayesB, BayesCπ, BayesDπ, Lasso, BLasso, and IBLasso) or lack (RR-BLUP, EN, RR-BLUP-Het, and BayesA) direct covariable selection. Bayesian methods are more efficient when the distribution of QTL effects is leptokurtic (positive kurtosis) because of the presence of genes with large effects. The RR-BLUP method is equally efficient when the QTL effects are normally distributed.

Comparisons among the methods for predicting genomic breeding values have been performed. Meuwissen et al. (2001) concluded that the BayesB method is theoretically best because it is slightly superior to RR-BLUP. However, the author simulated genotypic data with the same *a priori* distribution used for the estimation. This approach yielded greater accuracy for this method, although, such accuracy is unattainable in practice if the actual distribution associated with genetic effects differs from the *a priori* distribution assumed for analysis. In general, there is no method that is best under any circumstances because each method may yield significantly different results depending on the population structure and nature of the trait. However, the results obtained by Guo et al. (2012) indicate that the RR-BLUP method is easier to apply and equal to or better than the others for most applications in plants.

The assumed distributions for the genetic effects of markers in the different GS methods are Gaussian normal with common variance for RR-BLUP, Student's t-distribution given chi-square priori for variances for Bayesian methods, and Double Laplace exponential for Lasso. **Figure 4** illustrates the forms of the normal (RR-BLUP), t (BayesA), and double exponential (Lasso) distributions.

It is observed that, in relation to RR-BLUP, the prior density used in Bayesian Lasso shows a greater density mass at zero point and more robust tails providing greater shrinkage on regression coefficients close to zero and lower shrinkage on regression coefficients away from zero. The prior density used in BayesA also has a higher density mass at the zero point and more robust tails than the normally used RR-BLUP. Bayesian Lasso has greater shrinkage on regression coefficients close to zero than BayesA. However, the distribution tails are similar between the two methods (**Figure 4**).

The BayesA method implies a large number of markers with small effects or a few markers with moderate to large effects. BLasso implies a large number of markers with effects close to zero or a few markers with moderate to large effects. RR-BLUP implies a large number of markers with small effects.

Deep Learning

Machine-learning algorithms (random forest, bagging, support vector machine, and others) have been successful in recognizing complex patterns and making correct decisions based on data. Machine learning is a science of creating and studying algorithms that improve their own behavior in an iterative manner by design (Beysolow, 2017). Recent developments in machine learning enable the implementation of high-dimensional regression using nonlinear methods (Bellot et al., 2018). Another class of models, indeed a subfield of machine learning that became more used to prediction in recent times is deep learning. This theory is devoted to building algorithms that explain and learn a high and low level of abstractions of data that traditional machine learning algorithms often cannot (Beysolow, 2017).

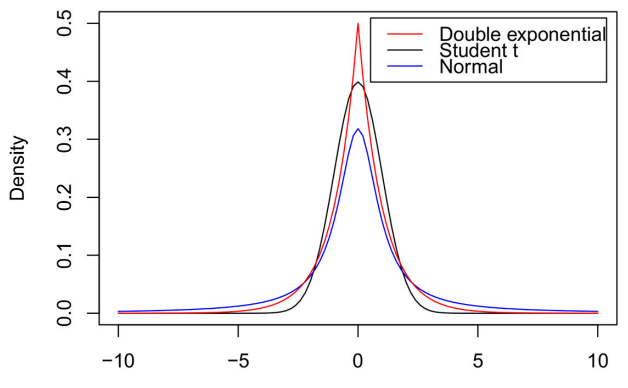


FIGURE 4 | Probability density functions of the double exponential, Student's t, and normal distributions, all with means equal to zero and variances equal to the unit.

Bellot et al. (2018) present an application of deep learning for the prediction of complex traits comparing the Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) with commonly used linear regression methods (BayesB and BayesRR). The deep learning under the linear regression, in some cases was very competitive. The MLP and CNN are very heterogeneous classes of predictor that depend on the number of layers, number of neurons per layer, and the activation function. However, the predictive accuracy of Bayesian linear methods is highly dependent on the heritability and this is not a main factor in MLP and CNN.

GWAS via the BayesCπ and BayesDπ Methods

Gene discovery or GWAS, which can be accomplished by the BayesCπ and BayesDπ methods (described by Habier et al., 2011), are advantageous because they provide information on the genetic architecture of the quantitative trait and identify the QTL positions by modeling the frequencies of SNPs with nonzero effects. They are advantageous over the regression analysis of single markers because they simultaneously account for all markers. However, care needs to be taken whenever the number of markers is larger than the number of individuals genotyped and phenotyped. Gianola (2013) showed that in such cases, the prior in Bayesian approaches such as BayesC and BayesD, is always influential, which could affect the inference of whether a marker is associated with the trait.

In the BayesC method, a common variance is specified for all loci. The BayesD method maintains the specific variances for each locus. Additionally, π is treated as an unknown with a uniform *a priori* distribution (0,1), thus producing the BayesCπ and BayesDπ methods. The modeling of π is interesting in the association analysis. The majority of the markers are not in LD with the genes; therefore, a set of markers associated with a trait must be identified. In contrast, the BayesB method determines π subjectively. Using the indicator variable δ_i, the BayesCπ and BayesDπ methods model the additive genetic effect of individual j as $a_j = \sum_{i=1}^{n} \beta_i x_{ij} \delta_i$, where $\delta_i = (0, 1)$. The distribution of $\delta = (\delta_1, \delta_n)$ is binomial with a probability of π. This mixed model is more parsimonious than the BayesB method. According to the model hierarchy, a distribution must be postulated for π, and there must be a beta distribution, which when appropriately specified, becomes a uniform distribution (0,1) (Legarra et al., 2011).

The quantities for x_{ij} are elements of the codominant marker genotype vector and are generally coded as 0, 1 or 2, depending on the number of copies of one of the alleles at the marker locus i, and β_i is defined as the element of the vector of the regression coefficients, which includes the marker effects on a phenotypic trait y by means of the LD with the genes that control the trait (Resende et al., 2013).

Sample Size for GS and GWAS

Genomic data are especially useful for GS, which allows selection at the seedling stage to increase genetic gain in the adult stage. With a high density of markers, the expected squared accuracy of GS is given by Daetwyler et al. (2008); Resende et al. (2008); Goddard et al. (2011); Grattapaglia and Resende (2011) $r^2_{\hat{g}g} = \frac{Nh^2}{(Nh^2+n_{QTL})} = \frac{Nh^2}{(Nh^2+m_e)} = \frac{Nh^2}{(Nh^2+2N_eL)} = \frac{Nh^2}{(Nh^2+\frac{L}{F})}$, where N is the number of genotyped and phenotyped individuals, L is the genome size (in Morgans) of the species, m_e is the number of independent chromosomal segments, Ne is the effective population size, and F is the inbreeding coefficient of the population. For a desired $r^2_{\hat{g}g}$, h^2, and n_{QTL}, N can be determined.

The reliability of GS is given by the expression $r^2_{gg} = \frac{Nh^2}{Nh^2+N_{QTL}}$, where r_{gg} equals GS accuracy, N is the number of individuals in the population, N_{QTL} is the number of QTLs that control each trait, and h^2 is the individual heritability. The estimate of the number of individuals that must be evaluated to obtain the desired accuracy can be obtained by the following expression, derived from the previous one, $N = \frac{r^2_{gg}N_{QTL}}{(1-r^2_{gg})h^2}$ (Resende et al., 2014).

Figure 5 shows the curve graphs with N in various scenarios (functions of h^2, N_{QTL}, and r_{gg}). Based on these graphs and the genetic information of the traits, breeders can adequately size their studies on inheritance and maximize genetic gain with the improvement made by selection.

Various kinds of information can be obtained from **Figure 5**. For example, considering scenario 3, it appears that for a trait with individual heritability equal to 0.30 and that is controlled by 100 QTLs, an accuracy of 90% can be obtained if the sample size is equal to 1,500 genotyped and phenotyped individuals.

From the first equation, the estimate of the number of QTLs that control each trait can be calculated based on the expression $N_{QTL} = \frac{\left(1-r^2_{gg}\right)Nh^2}{r^2_{gg}}$. Once the selective accuracy and heritability are estimated, given the N practiced in a study, the N_{QTL} can be estimated for several traits.

A possible exercise is the theoretical determination of N_{QTL}, given the N and the estimated h^2 while varying r^2_{gg}. For a case of h^2 equal to 0.30, and N equal to 1,500, the N_{QTL} values can be inferred according to **Figure 6**. The same figure shows the case of h^2 equal to 0.20, and N equal to 1,500.

Based on **Figure 6**, it can be seen that for N = 1,500 genotyped individuals, the QTL numbers vary from 49 to 468 (Scenario 1, $h^2 = 0.30$) and from 32 to 312 (Scenario 2, $h^2 = 0.20$), when the accuracy varies from 0.95 to 0.70, respectively.

Sample Size for Gene Detection

The sample size (N), with power of detection at a significance level of 10^{-5} according to the h^2_{mi} magnitude of the QTL (considered as random effect), is given by Resende (2015) $N \approx \frac{\left(Z_{\left(1-\frac{\alpha}{2}\right)}+Z_{(1-\beta)}\right)^2(1-h^2)}{h^2_{mi}}$, where $Z_{\left(1-\frac{\alpha}{2}\right)}$ and $Z_{(1-\beta)}$ are the values of the cumulative distribution function of the standard normal distribution, associated with the probabilities of error type I (α) and error type II (β) for bilateral hypothesis tests.

The quantity (1 − β) is the probability that the experiment will exhibit a statistically significant difference between the treatment averages. Values of 0.80 and 0.90 are common and appropriate in practice.

Table 2 and **Figure 7** show that the sample sizes ($<$1,000) commonly used in plant breeding only detect QTL when the QTL explains 5% or more of the phenotypic variation, a fact that is unlikely under polygenic inheritance (total trait $h^2 < 0.50$). The power of 0.90 is more appropriate because it leads to an 81% = 0.90^2 probability that two independent studies will detect the same QTL.

GS APPLICATIONS IN BRACHIARIA, GUINEA GRASS AND ELEPHANT GRASS BREEDING

In TFG breeding, combining conventional breeding efforts and GS has not been a simple task. As opposed to advances in animal breeding and crop commodities, they have been slow and challenging in tropical forages. The number of candidate TFG species is high, and decisions about investments need to be made considering the effective benefits of GS, the potential profit that can be achieved by the new cultivars, and the real impact of new forage on livestock production. The three tropical genus/species brachiaria (*Urochloa* spp.), Guinea grass (*M. maximus*) and elephant grass (*C. purpureus*) are very important and extensively used as pastures in tropical America, Asia, and sub-Saharan Africa.

Marker number and density are important factors influencing the efficient use of GS in TFG breeding. One of the reasons for the low accuracy of GS is the exceptionally low number of effective markers, which may result from a non-adequate reference genome. Before the *U. ruziziensis* genome assembly was publicly available, the genomes of *S. viridis* and *P. virgatum* were frequently used as reference genomes for SNP calling and linkage map construction in *Urochloa* species with 1,000 SNPs (Ferreira et al., 2019) and *M. maximus* with 1,322 SNPs (Deo et al., 2020). For *C. purpureus*, 20,144 SilicoDArT and 28,610 SNP markers have been mapped onto the pearl millet (*C. americanus*) reference (Muktar et al., 2019). Compared with other agronomically important *Poaceae* species, such as maize, for which the 50 K Illumina MaizeSNP50 BeadChip (Ganal et al., 2011) and the 600 K Affymetrix Axiom Maize Genotyping Array (Unterseer et al., 2014) are available, the number of markers in TFG needs to be significantly enhanced. Genome calling using reference genomes of other grasses improves the number of SNPs, as shown by Matias et al. (2019a), who reported $>$26k SNPs in *Urochloa* hybrids. However, the minimum allele depth used was $\leq$2 reads considering allele dosage and resulted in a low predictive ability ($<$0.31) in GS for agronomic traits. Similar

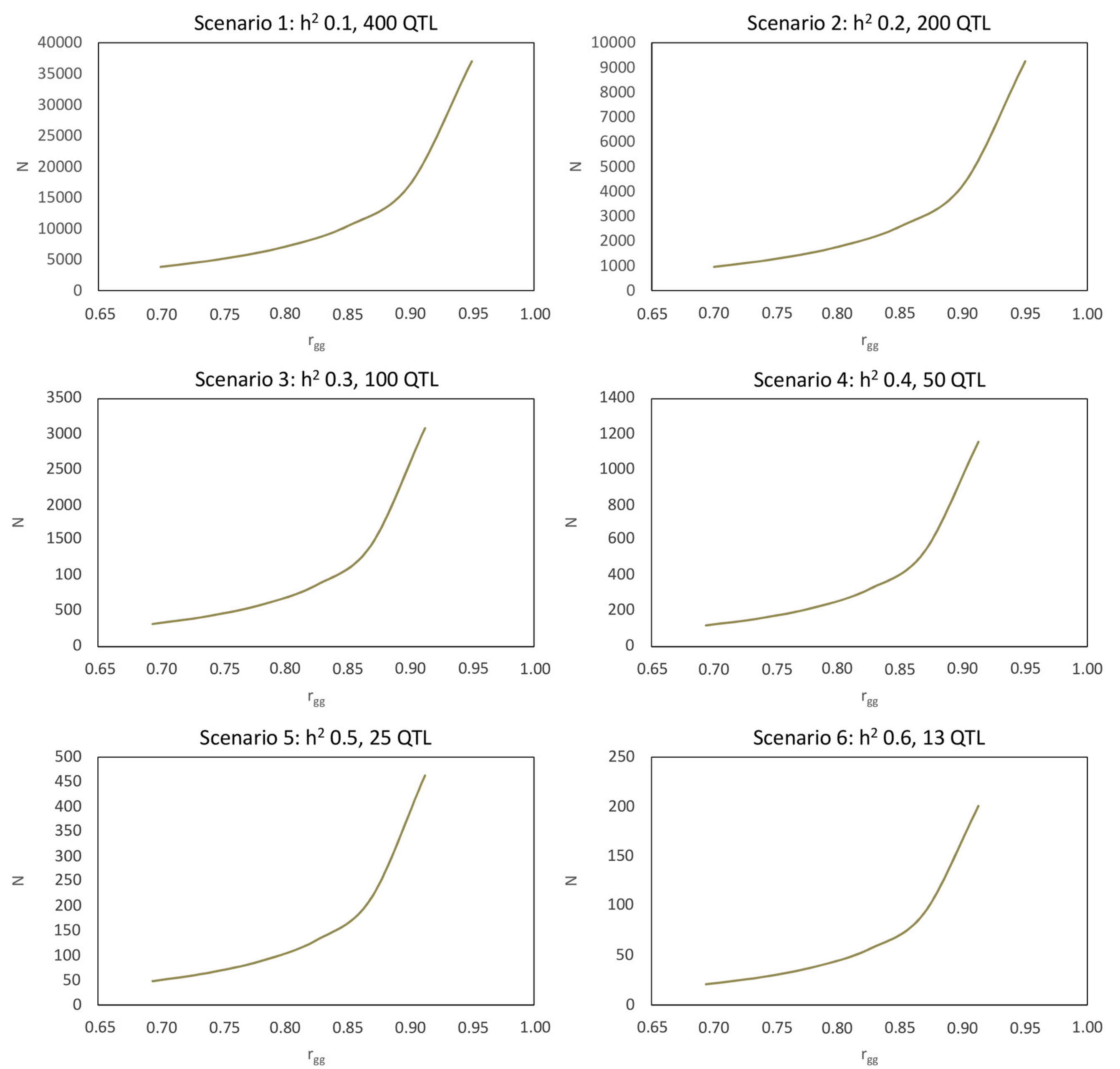


FIGURE 5 | Sample size for genomic selection with desired accuracy ranging from 0.70 to 0.95 in six scenarios in terms of heritability and quantitative trait locus (QTL) number.

results were obtained by Lara et al. (2019) in *M. maximus* in which >32k SNPs were classified as unique and used in GS, although, the maximum value of predictive ability using tetraploid dosage of 0.3955 was achieved for the trait organic matter that displayed secondary importance in forage breeding.

Current TFG breeding programs lack the important information that could improve and allow the efficient use of GS. Firstly, a major impact will be obtained by increasing the number of markers per genome size by sequencing and generating reference genomes for the target species or more closely related species. Secondly, we need to improve our knowledge about the inheritance of target traits in tropical forages including the genetic effects of biallelism in (auto) tetraploids. Thirdly, we need to work with training populations connected with validation and breeding populations and testing environments that must be correlated with the environment of the target population (Burgueño et al., 2012; Jarquín et al., 2014; Santantonio et al., 2020). Finally, we need to work with half-sib and full-sib progenies to improve the predictive ability for traits in which the dominance effects are significant, aiming to predict crosses

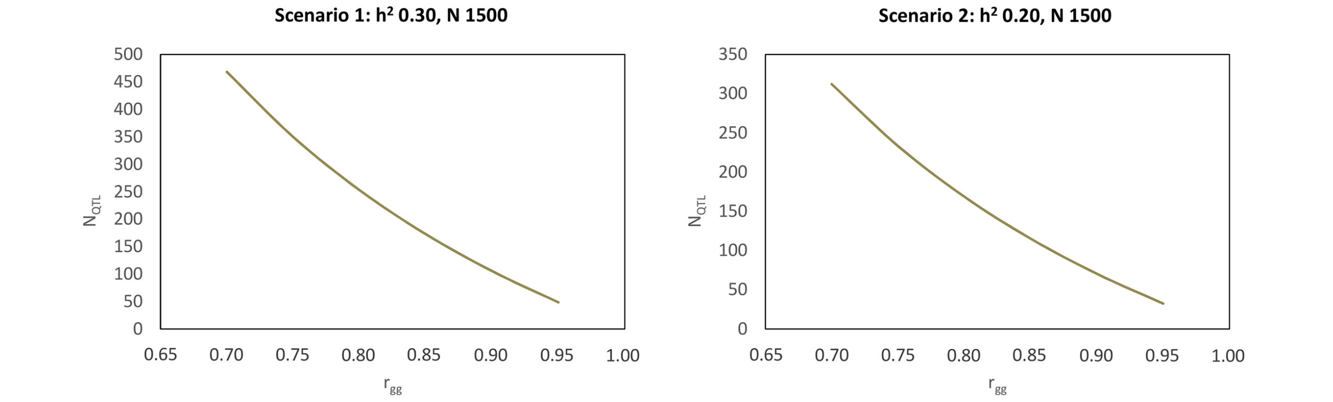


FIGURE 6 | Number of quantitative trait loci (N_{QTL}) for genomic selection with accuracy ranging from 0.70 to 0.95 in two scenarios in terms of heritability and individual sample size.

TABLE 2 | Sample size (*N*) and power for detection of significance level 10^{-5} according to the h^2_{mi} magnitude of the quantitative trait locus, considered as having a random effect: $N \approx \frac{\left(Z_{\left(1-\frac{\alpha}{2}\right)}+Z_{(1-\beta)}\right)^2(1-h^2)}{h^2_{mi}}$.

h² = 0.30					h² = 0.50				
Z for β = 0.90	**Z for α = 10^{-5}**	$\left(Z_{\left(1-\frac{\alpha}{2}\right)}+Z_{(1-\beta)}\right)^2$	h^2_{mi}	***N***	**Z for β = 0.90**	**Z for α = 10^{-5}**	$\left(Z_{\left(1-\frac{\alpha}{2}\right)}+Z_{(1-\beta)}\right)^2$	h^2_{mi}	***N***
1.28	3.99	27.7729	0.001	19441	1.28	3.99	27.7729	0.001	13886
1.28	3.99	27.7729	0.005	3888	1.28	3.99	27.7729	0.005	2777
1.28	3.99	27.7729	0.01	1944	1.28	3.99	27.7729	0.01	1389
1.28	3.99	27.7729	0.05	389	1.28	3.99	27.7729	0.05	278
1.28	3.99	27.7729	0.1	194	1.28	3.99	27.7729	0.1	139
1.28	3.99	27.7729	0.2	97	1.28	3.99	27.7729	0.2	69
1.28	3.99	27.7729	0.3	65	1.28	3.99	27.7729	0.3	46

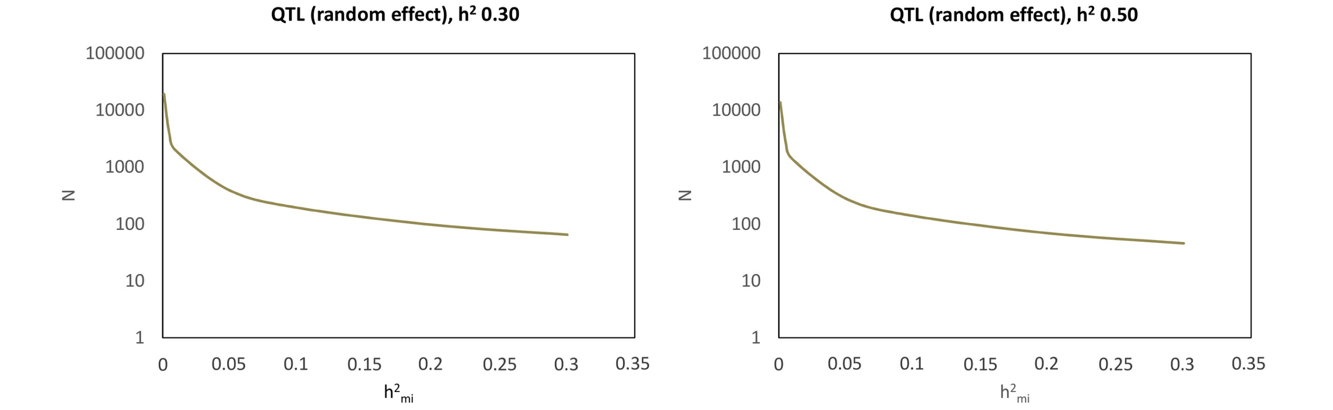


FIGURE 7 | Sample size (N) required to detect genetic effects of markers (assumed to be random effects) with different marker heritability ($h^2{}_{mi}$) and total heritability (h^2): N values as a function of $h^2{}_{mi}$. The plotted N values were obtained via logarithmic transformation to improve visualization.

that are not performed in tetraploid *Urochloa* and *M. maximus* hybrid development programs.

The current *M. maximus* and *Urochloa* breeding programs generate thousands of hybrids annually, and those hybrids must pass through several steps of selection until they finally achieve the status for evaluation under animal feeding pressure to prove their value in animal production and be released as new cultivars. However, it is mandatory for hybrids to show apomixis and resistance/tolerance to spittlebugs (mostly in *Urochloa*) and diseases (mostly in *M. maximus* and *C. purpureus*). These traits are a great bottleneck slowing the subsequent evaluation steps in the breeding program since phenotyping of individuals demands significant labor, time (two or more years), and the dedication of trained technicians. As a result, only a small number of individuals can be evaluated annually reducing the rate of genetic gain. Methods such as GWAS and MAS may help to speed up the identification of individuals showing these important traits and increased rates of genetic gain. Recently published data on mapping genomic regions associated with apospory emphasize the routine application of markers for selection in *Urochloa* and *Megathyrsus* (Worthington et al., 2016; Deo et al., 2020).

Finally, TFG breeding programs can be sped up by the application of GS. However, this demands greater investments; collaboration among breeders, molecular biologists, and bioinformaticians; integration of research teams from different institutions and countries; and most importantly, continuity associated with critical course corrections.

ADDITIONAL METHODS APPLIED TO TFG BREEDING

The use of new efficient high-throughput methodologies in addition to GS should be discussed according to their accuracies and potential use in higher numbers of individuals at initial stages of selection. The first obvious application is its use when no genotyping tool is available at a reasonable cost. A second application would be to use phenomics to screen nearly fixed genetic materials which is likely to capture non-additive genetic effects. Nonetheless, phenomics could deliver breeding innovations, and the challenge represented by the breeding target scenario (Reynolds et al., 2020).

Phenomic selection (PS) using high-throughput phenotyping methods less expensive than genotyping by sequencing is an opportunity for tropical forage breeding. Rincent et al. (2018) proposed using near-infrared spectroscopy (NIRS) variables generated as regressors or to estimate kinship in the same statistical models used in GS to perform PS. The results were promising and cost affordable for wheat and poplar when compared to GS. TFGs are a probable candidate for this method because phenotyping using NIRS to obtain bromatological data is routine in research programs and may be studied and amplified to other spectra to be performed as a routine method of PS. Biomass measuring in TFG is a laborious, time-consuming, and biased task, because of the necessity of several annual evaluations (4 to 7) during selection. It also limits the number of individuals in experiments (300 to 2,000). Sensor-based images enabled high-throughput non-invasive phenotyping throughout the growing cycles of forage grasses, and the models established a high correlation between images and the biomass yield in *M. maximus* (Castro et al., 2020) as well as for crude protein percentage and chlorophyll concentration in *Urochloa* (Jiménez et al., 2020). Deep learning-based neural network studies demonstrated that accuracies must be increased by pre-trained models and data augmentation (Castro et al., 2020). Nevertheless, deep learning progress is accelerating and will be able to perform better predictions than ever (Montesinos-López et al., 2021). Although it has been the subject of debate in the past, extra investment in phenotyping technologies is becoming more accepted to capitalize on recent developments in crop genomics and prediction models. In this context the different strategies for phenotyping can be built from phenomic selection (Rincent et al., 2018), high-throughput phenotyping, and detailed characterization or 'precision' phenotyping (Reynolds et al., 2020).

THE FUTURE OF GS IN FORAGE BREEDING

The availability of genome-wide, high-throughput, and cost-effective flexible markers, across the genome, suitable for large populations with or without a reference genome sequence, is the most important factor for the effective and efficient implementation of GS. Recent advances in long read quality and sequence throughput, in addition to other technologies such as Hi-C or optical maps, make it possible for virtually any research group with reasonable funding to obtain reference-grade genome assemblies for their crop of choice. While not necessarily easy, the generation of high-quality genome assemblies should be considered as a starting point for any orphan species that would benefit from the use of genomic tools for crop improvement. These assemblies could be extremely useful for resequencing and variant discovery, which can lead to genotyping platforms for association studies and GS. When coupled with well-designed and thoroughly phenotyped training populations, these genomic resources could serve as the basis for implementing GS steps in the breeding of TFGs.

As discussed by Lin et al. (2014) and Bhat et al. (2016), the cost of identifying and genotyping a large number of SNPs is still a barrier for TFGs, although, second-generation sequencing technology has provided new SNP genotyping platforms, particularly GBS. In addition, phenotyping large representative reference populations is expensive. Reduction of phenotype assessment costs per individual and new phenomic approaches are essential to take advantage of the true benefits of GS. Marker technologies must be combined with high-throughput phenotyping to achieve significant genetic gains for complex traits.

Furthermore, the considerations stated by Simeão-Resende et al. (2014) are still valid. GS will allow an increase in the early-generation of number of individuals evaluated considering the large number of targeted traits. However, when we deal with GS in the improvement of tropical forages, we realize that there is still a long way to go. Theoretically, by models and methods already developed and successfully applied in commodity species, the procedures could be easily incorporated into the routine of breeding programs. Nevertheless, in orphan species, all knowledge needs to be built on solid molecular bases. In principle, the evaluation of a large number of individuals for selection purposes increases the probability of the best allelic combinations for traits of economic importance without narrowing the genetic basis for selection. This should be considered in the improvement of polyploid and apomictic *Urochloa* spp. and *M. maximus*. Performing inter- and intraspecific crosses with sexual plants in these genera increases the variability available for selection and allows the generation of genetic combinations not found in apomictic accessions. The spectrum of possibilities for GS expands considerably for these species; however, the identification of markers is narrowed to large-scale phenotyping and genotyping. The conformation of the discovery population should be carefully considered in terms of the number of hybrid families to be evaluated, the number of individuals per half-sib and full-sib families, and the distribution of markers on the chromosomes of the paternal and maternal genomes. The mother plants to be used in crosses should be exclusively sexual, so that they do not generate, in addition to hybrids, their own clones (by apomixis) in the progeny, which would cause an incorrect bias in the determination of GEBV and the identification of markers and their effects.

Finally, designing forage breeding programs, mainly for polyploid and apomictic grasses, and proposing breeding schemes that make optimum use of GS is a significant task for plant breeders. Although, this is a challenge, it is also a great opportunity to accelerate genetic gain in TFG breeding.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

We acknowledge the financial support from Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Instituto Nacional de Ciência e Tecnologia do Café (INCT Café), and Associação para o Fomento à Pesquisa de Melhoramento de Forrageiras (UNIPASTO).

ACKNOWLEDGMENTS

We are thankful to João Romero do Amaral Santos de Carvalho Rocha and Marco Antonio Machado for valuable comments and discussions.

REFERENCES

Alexandratos, N., and Bruinsma, J. (2012). *World Agriculture Towards 2030/2050: The 2012 Revision. ESA Working Paper 12-03*. Rome: FAO, doi: 10.22004/ag.econ.288998

Azevedo, C. F., Resende, M. D. V., Silva, F. F., Viana, J. M. S., Valente, M. S. F., Resende, M. F. R. Jr., et al. (2015). Ridge, Lasso and Bayesian additive dominance genomic models. *BMC Genet.* 16:105. doi: 10.1186/s12863-015-0264-2

Bajželj, B., Richards, K. S., Allwood, J. M., Smith, P., Dennis, J. S., Curmi, E., et al. (2014). Importance of food-demand management for climate mitigation. *Nat. Clim. Change* 4, 924–929. doi: 10.1038/nclimate2353

Ballesta, P., Bush, D., Silva, F. F., and Mora, F. (2020). Genomic predictions using low-density SNP markers, pedigree and GWAS information: a case study with the non-model species *Eucalyptus cladocalyx*. *Plants* 9:99. doi: 10.3390/plants9010099

Barabaschi, D., Tondelli, A., Desiderio, F., Volante, A., Vaccino, P., Valè, G., et al. (2016). Next generation breeding. *Plant Sci.* 242, 3–13. doi: 10.1016/j.plantsci.2015.07.010

Barcaccia, G., Palumbo, F., Sgorbati, S., Albertini, E., and Pupilli, F. (2020). A reappraisal of the evolutionary and developmental pathway of apomixis and its genetic control in angiosperms. *Genes* 11:859. doi: 10.3390/genes11080859

Barrios, S. C., Valle, C. B., Alves, G. F., Simeão, R. M., and Jank, L. (2013). Reciprocal recurrent selection in the breeding of *Brachiaria decumbens*. *Trop. Grassl Forrajes Trop.* 1, 52–54.

Bellot, P., De Los Campos, G., and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics* 210, 809–819. doi: 10.1534/genetics.118.301298

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F. C., Falentin, C., et al. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* 4, 879–887. doi: 10.1038/s41477-018-0289-4

Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R., and Munoz, P. (2019). How can a high-quality genome assembly help plant breeders? *GigaScience* 8:giz068. doi: 10.1093/gigascience/giz068

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

Beysolow, T. II (2017). "Introduction to deep learning," in *Introduction to Deep Learning Using R*, (Berkeley, CA: Apress), 1–9.

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., et al. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front. Genet.* 7:221. doi: 10.3389/fgene.2016.00221

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., and Maliepaard, C. (2018). Tools for genetics studies in experimental populations of polyploids. *Front. Plant Sci.* 9:153. doi: 10.3389/fpls.2018.00513

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

Castro, W., Marcato-Junior, J., Polidoro, C., Osco, L. P., Gonçalves, W., Rodrigues, L., et al. (2020). Deep learning applied to phenotyping of biomass in forages with UAV-based RGB imagery. *Sensors* 20:4802. doi: 10.3390/s20174802

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

Daetwyler, H. D., Villanueva, B., Bijma, P., 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

Daud, Z., Hatta, M., Kassim, A., Mohd Aripin, A., and Awang, H. (2014). Analysis of Napier grass (*Pennisetum purpureum*) as a potential alternative fiber in paper industry. *Mater. Res. Innov.* 18, 18–20. doi: 10.1179/1432891714Z.000000000925

de Bem Oliveira, I., Resende, M. F. R., Ferrão, L. F. V., Amadeu, R. R., Endelman, J. B., Kirst, M., et al. (2019). Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3* 9, 1189–1198. doi: 10.1534/g3.119.400059

Dekkers, J. C. M. (2007). Prediction of response to marker assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet.* 124, 331–341. doi: 10.1111/j.1439-0388.2007.00701.x

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B., and Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biol.* 22:3. doi: 10.1186/s13059-020-02224-8

Deo, T. G., Ferreira, R. C. U., Lara, L. A. C., Moraes, A. C. L., Alves-Pereira, A., de Oliveira, F. A., et al. (2020). High-resolution linkage map with allele dosage allows the identification of regions governing complex traits and apospory in Guinea Grass (*Megathyrsus maximus*). *Front. Plant Sci.* 11:15. doi: 10.3389/fpls.2020.00015

Eberhart, S. A. (1970). Factors affecting efficiencies of breeding methods. *Afr. Soils* 15, 655–680.

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:e19379. doi: 10.1371/journal.pone.0019379

Fernando, R. L. (1998). "Genetic evaluation and selection using genotypic, phenotypic and pedigree information," in *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, Armidale, NSW, 329–336.

Ferreira, R. C. U., Lara, L. A., de, C., Chiari, L., Barrios, S. C. L., do Valle, C. B., et al. (2019). Genetic mapping with allele dosage information in tetraploid *Urochloa decumbens* (Stapf) R. D. Webster reveals insights into spittlebug (*Notozulia entreriana* Berg) resistance. *Front. Plant Sci.* 10:92. doi: 10.3389/fpls.2019.00092

Fuzinatto, V. A., Pagliarini, M. S., and Valle, C. B. (2007). Microsporogenesis in sexual *Brachiaria* hybrids (Poaceae). *Genet. Mol. Res.* 6, 1107–1117.

Gallais, A. (1989). Concepts of varietal value and of test value in autotetraploids: application to genetic advance in population improvement. *Genome* 32, 420–424. doi: 10.1139/g89-465

Gallais, A. (2003). *Quantitative Genetics and Breeding Methods in Autopolyploid Plants*. Paris: Institut National de la Recherche Agronomique.

Ganal, M. W., Durstewitz, G., Polley, A., Berard, A., Buckler, E. S., Charcosset, A., et al. (2011). A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One* 6:e28334. doi: 10.1371/journal.pone.0028334

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

Gianola, D., Perez-Enciso, M., and Toro, M. A. (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* 163, 347–365.

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

Goddard, M. E., Hayes, B. J., and Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.* 128, 409–421. doi: 10.1111/j.1439-0388.2011.00964.x

Grattapaglia, D., and Resende, M. D. V. (2011). Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7, 241–255. doi: 10.1007/s11295-010-0328-4

Guo, Z., Tucker, D. M., Lu, J., Kishore, V., and Gay, G. (2012). Evaluation of genome-wide selection efficiency in maize nested association mapping populations. *Theor. Appl. Genet.* 124, 261–275. doi: 10.1007/s00122-011-1702-9

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship 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

Habte, E., Muktar, M. S., Abdena, A., Hanson, J., Sartie, A. M., Negawo, A. T., et al. (2020). Forage performance and detection of marker trait associations with potential for napier grass (*Cenchrus purpureus*) improvement. *Agronomy* 10:542. doi: 10.3390/agronomy10040542

Haley, C. S., and Visscher, P. M. (1998). Strategies to utilize marker-quantitative trait loci associations. *J. Dairy Sci.* 81, 85–97. doi: 10.3168/jds.S0022-0302(98)70157-2

Hallauer, A. R. (2011). Evolution of plant breeding. *Crop Breed. Appl. Biotechnol.* 11, 197–206. doi: 10.1590/S1984-70332011000300001

Hand, M. L., and Koltunow, A. M. G. (2014). The genetic control of apomixis: asexual seed formation. *Genetics* 197, 441–450. doi: 10.1534/genetics.114.163105

Hanna, W. W. (1981). Method of reproduction in napiergrass and in the 3X and 6X alloploid hybrids with pearl millet. *Crop Sci.* 21, 123–126. doi: 10.2135/cropsci1981.0011183X002100010033x

Hayes, B. J., Cogan, N. O. I., Pembleton, L. W., Goddard, M. E., Wang, J., Spangenberg, G., et al. (2013). Prospects for genomic selection in forage plant species. *Plant Breed.* 132, 133–143. doi: 10.1111/pbr.12037

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

IBGE (2016). *Mudanças na Cobertura e Uso da Terra do Brasil 2000 – 2010 – 2012 – 2014*. Rio de Janeiro: IBGE, 29.

Jank, L., Barrios, S. C., do Valle, C. B., Simeão, R. M., and Alves, G. F. (2014). The value of improved pastures to Brazilian beef production. *Crop Pasture Sci.* 65, 1132–1137. doi: 10.1071/CP13319

Jank, L., Valle, C. D., and Resende, R. M. S. (2011). Breeding tropical forages. *Crop Breed. Appl. Biotechnol.* 11, 27–34. doi: 10.1590/S1984-70332011000500005

Jarquín, D., Crossa, J., Lacaze, X., Cheyron, P. D., 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

Jiménez, J. C., Leiva, L., Cardoso, J. A., French, A. N., and Thorp, K. R. (2020). Proximal sensing of *Urochloa* grasses increases selection accuracy. *Crop Pasture Sci.* 71, 401–409. doi: 10.1071/CP19324

Jones, C., Vega, J., Worthington, M., Thomas, A., Gasior, D., Harper, J., et al. (2021). A comparison of differential gene expression in response to the onset of water stress between three hybrid Brachiaria genotypes. *Front. Plant Sci.* 12:637956. doi: 10.3389/fpls.2021.637956

Kohavi, R. (1995). "A study of cross-validation and bootstrap for estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, ed. C. S. Mellish (San Francisco, CA: Morgan Kaufmann Publishers), 1137–1143.

Lamkey, K. R., and Edwards, F. (1999). "Quantitative genetics of heterosis," in *The Genetics and Exploitation of Heterosis in Crops*, eds J. G. Coors and S. Pandey (Madison, WI: ASA/CSSA/SSSA), 29–43.

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743–756.

Lara, L., de, C., Santos, M. F., Jank, L., Chiari, L., Vilela, M. M., et al. (2019). Genomic selection with allele dosage in *Panicum maximum* Jacq. *G3* 9, 2463–2475. doi: 10.1534/g3.118.200986

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., and Fritz, S. (2011). Improved Lasso for genomic selection. *Genet. Res.* 93, 77–87. doi: 10.1017/S0016672310000534

Lima, L. P., Azevedo, C. F., Resende, M. D. V., Silva, F. F., Suela, M. M., Nascimento, M., et al. (2019a). New insights into genomic selection through population-based non-parametric prediction methods. *Sci. Agric.* 76, 290–298. doi: 10.1590/1678-992x-2017-0351

Lima, L. P., Azevedo, C. F., Resende, M. D. V., Silva, F. F., Viana, J. M. S., and Oliveira, E. J. (2019b). Triple categorical regression for genomic selection: application to cassava breeding. *Sci. Agric.* 76, 368–375. doi: 10.1590/1678-992x-2017-0369

Lin, Z., Hayes, B., and Daetwyler, H. (2014). Genomic selection in crops, trees and forages: a review. *Crop Pasture Sci.* 65, 1177–1191. doi: 10.1071/CP13363

Lovell, J. T., Jenkins, J., Lowry, D. B., Mamidi, S., Sreedasyam, A., Weng, X., et al. (2018). The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nat. Commun.* 9:5213. doi: 10.1038/s41467-018-07669-x

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., et al. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* 9:e1003215. doi: 10.1371/journal.pgen.1003215

Lush, J. (1937). *Animal Breeding Plans*. Ames: Iowa State College Press.

Machado, J. C., Pereira, J. F., Azevedo, A. L. S., Mittelmann, A., Pereira, A. V., Sobrinho, F. S., et al. (2019). "Melhoramento genético de forrageiras e o uso de ferramentas genômicas," in *Melhoramento de Forrageiras na Era Genômica*, eds A. L. S. Azevedo, J. F. Pereira, and J. C. Machado (Brasilia: Embrapa), 11–42.

Mackay, I., Piepho, H. P., and Garcia, A. A. F. (2019). "Statistical methods for plant breeding," in *Handbook of Statistical Genomics*, eds D. J. Balding, I. Moltke, and J. Marioni (Hoboken, NJ: John Wiley & Sons), 501–521.

Matias, F. I., Alves, F. C., Meireles, K. G. X., Barrios, S. C. L., Valle, C. B., Endelman, J. B., et al. (2019a). On the accuracy of genomic prediction models considering multi-trait and allele dosage in *Urochloa* spp. interspecific tetraploid hybrids. *Mol. Breed.* 39:100. doi: 10.1007/s11032-019-1002-7

Matias, F. I., Meireles, K. G. X., Nagamatsu, S. T., Barrios, S. C. L., Valle, C. B., Carazzolle, M. F., et al. (2019b). Expected genotype quality and diploidized marker data from genotyping-by-sequencing of Urochloa spp. tetraploids. *Plant Genome* 12:190002. doi: 10.3835/plantgenome2019.01.0002

McCormick, R. F., Truong, S. K., and Mullet, J. E. (2015). RIG: recalibration and interrelation of genomic sequence data with the GATK. *G* 3, 655–665. doi: 10.1534/g3.115.017012

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Melo, A. T. O., Bartaula, R., and Hale, I. (2016). GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics* 17:29. doi: 10.1186/s12859-016-0879-y

Mendes-Bonato, A. B., Pagliarini, M. S., and Valle, C. B. (2007). Meiotic arrest compromises pollen fertility in an interspecific hybrid between *Brachiaria ruziziensis* x *Brachiaria decumbens* (Poaceae: Paniceae). *Braz. Arch. Biol. Technol.* 50, 831–837. doi: 10.1590/S1516-89132007000500011

Mendes-Bonato, A. B., Risso-Pascotto, C., Pagliarini, M. S., and Valle, C. B. (2006). Cytogenetic evidence for genome elimination during microsporogenesis in an interspecific hybrid between *Brachiaria ruziziensis* and *B. brizantha* (Poaceae). *Genet. Mol. Biol.* 29, 711–714. doi: 10.1590/S1415-47572006000400021

Meuwissen, T. H. E. (2007). Genomic selection: marker assisted selection on genome-wide scale. *J. Anim. Breed. Genet.* 124, 321–322. doi: 10.1111/j.1439-0388.2007.00708.x

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.

Miles, J. W., Cardona, C., and Sotelo, G. (2006). Recurrent selection in a synthetic Brachiariagrass population improves resistance to three spittlebug species. *Crop Sci.* 46, 1008–1093. doi: 10.2135/cropsci2005.06-0101

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.* 92, 4648–4655. doi: 10.3168/jds.2009-2064

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodrígues, P., Barrón-López, J. A., Martini, J. W. R., Fajardo-Flores, S., et al. (2021). A review of deep learning applications for genomic selection. *BMC Genomics* 22:19. doi: 10.1186/s12864-020-07319-x

Morais, R. F. D., Souza, B. J. D., Leite, J. M., Soares, L. H. D. B., Alves, B. J. R., Boddey, R. M., et al. (2009). Elephant grass genotypes for bioenergy production by direct biomass combustion. *Pesq. Agrop. Bras.* 44, 133–140. doi: 10.1590/S0100-204X2009000200004

Morales, F. J. (2009). "Introduction to tropical agriculture and outlook for tropical crops in a globalized economy," in *Tropical Biology and Conservation Management*, eds K. Del Claro, P. S. Oliveira, and V. Rico-Gray (Oxford: EOLSSPublishers Co Ltd), 1–27.

Mrode, R., Coffey, M., and Berry, D. P. (2010). Understanding genomic evaluations from various evaluation methods and GMACE. *Interbull. Bull.* 42, 52–55.

Mrode, R. A. (2014). *Linear Models for the Prediction of Animal Breeding Values*. Wallingford: CAB International.

Muktar, M. S., Teshome, A., Hanson, J., Negawo, A. T., Habte, E., Entfellner, J. B. D., et al. (2019). Genotyping by sequencing provides new insights into the diversity of Napier grass (*Cenchrus purpureus*) and reveals variation in genome-wide LD patterns between collections. *Sci. Rep.* 9:6936. doi: 10.1038/s41598-019-43406-0

Negawo, A. T., Jorge, A., Hanson, J., Teshome, A., Muktar, M. S., Azevedo, A. L. S., et al. (2018). Molecular markers as a tool for germplasm acquisition to enhance the genetic diversity of a Napier grass (*Pennisetum purpureum*) collection. *Trop. Grassl. Forrajes Trop.* 6, 58–69. doi: 10.17138/TGFT(6)58-69

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

Nunes, S. G., Boock, A., Penteado, M. I., de, O., and Gomes, D. T. (1984). *Brachiaria brizantha cv. Marandu. Documentos Embrapa, No. 21.* Campo Grande: Embrapa.

Ozias-Akins, P., and van Dijk, P. J. (2007). Mendelism genetics of apomixis in plants. *Ann. Rev. Genet.* 41, 509–537. doi: 10.1146/annurev.genet.40.110405.090511

Pagliarini, M. S., Riso-Pascotto, C., Sauza-Kaneshima, A. M., and Valle, C. B. (2008). Analysis of meiotic behavior in selecting potential genitors among diploid and artificially induced tetraploid accessions of *Brachiaria ruziziensis* (Poaceae). *Euphytica* 164, 181–187. doi: 10.1007/s10681-008-9697-2

Paudel, D., Kannan, B., Yang, X., Harris-Shultz, K., Thudi, M., Varshney, R. K., et al. (2018). Surveying the genome and constructing a high-density genetic map of napiergrass (*Cenchrus purpureus* Schumach). *Sci. Rep.* 8:14419. doi: 10.1038/s41598-018-32674-x

Paul, B. K., Koge, J., Maass, B. L., Notenbaert, A., Peters, M., Groot, J. C. J., et al. (2020). Tropical forage technologies can deliver multiple benefits in Sub-Saharan Africa. A meta-analysis. *Agron. Sustain. Dev.* 40:22. doi: 10.1007/s13593-020-00626-3

Pengelly, B. C., and Maass, B. L. (2019). Tropical and subtropical forage germplasm conservation and science on their deathbed! 2. Genebanks, FAO and donors must take urgent steps to overcome the crisis. *Outlook Agric.* 48, 210–219. doi: 10.1177/0030727019867955

Pereira, A. V., Lédo, F. J. S., and Machado, J. C. (2017). BRS Kurumi and BRS Capiaçu – New elephant grass cultivars for grazing and cut-and-carry system. *Crop Breed. Appl. Biotechnol.* 17, 59–62. doi: 10.1590/1984-70332017v17n1c9

Pereira, J. F., Azevedo, A. L. S., Pessoa-Filho, M., Romanel, E. A. D. C., Pereira, A. V., Vigna, B. B. Z., et al. (2018). Research priorities for next-generation breeding of tropical forages in Brazil. *Crop Breed. Appl. Biotechnol.* 18, 314–319. doi: 10.1590/1984-70332018v18n3n46

Pessim, C., Pagliarini, M. S., Jank, L., Kaneshima, M. A. S., and Mendes-Bonato, A. B. (2010). Meiotic behavior in Panicum maximum Jacq. (Poaceae: Panicoideae: Paniceae): hybrids and their genitors. *Acta Sci. Agron.* 32, 417–422. doi: 10.4025/actasciagron.v32i3.6461

Pessim, C., Pagliarini, M. S., Silva, N., and Jank, L. (2015). Chromosome stickiness impairs meiosis and influences reproductive success in *Panicum maximum* (Poaceae) hybrid plants. *Genet. Mol. Res.* 14, 4195–4202. doi: 10.4238/2015.April.28.2

Rao, I., Peters, M., Castro, A., Schultze-Kraft, R., White, D., Fisher, M., et al. (2015). LivestockPlus – The sustainable intensification of forage-based agricultural systems to improve livelihoods and ecosystem services in the tropics. *Trop. Grassl. Forrajes Trop.* 3, 59–82. doi: 10.17138/TGFT(3)59-82

Reis, M. C., Sobrinho, F. S., Ramalho, M. A. P., Ferreira, D. F., Ledo, F. J. S., and Pereira, A. V. (2008). Allohexaploid pearl millet x elephantgrass population potential for a recurrent selection program. *Pesq. Agropec. Bras.* 43, 195–199. doi: 10.1590/S0100-204X2008000200006

Resende, M. D. V. (2007). *Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético*. Colombo: Embrapa Florestas, 561.

Resende, M. D. V. (2015). *Genética Quantitativa e de Populações*. Visconde do Rio Branco: Suprema.

Resende, M. D. V., and Alves, R. S. (2020). Linear, generalized, hierarchical, Bayesian and random regression mixed models in genetics/genomics in plant breeding. *Funct. Plant Breed. J.* 2, 1–31.

Resende, M. D. V., Lopes, P. S., Silva, R. L., and Pires, I. E. (2008). Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesq. Flor. Bras.* 56, 63–78.

Resende, M. D. V., Silva, F. F., and Azevedo, C. F. (2014). *Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência*. Visconde do Rio Branco: Suprema.

Resende, M. D. V., Silva, F. F., Lopes, P. S., and Azevedo, C. F. (2012). *Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial*. Viçosa: Universidade Federal de Viçosa.

Resende, M. D. V., Silva, F. F., Resende, M. F. R. Jr., and Azevedo, C. F. (2013). "Genome-wide selection (GWS)," in *Biotechnology and Plant Breeding*, eds A. Borém and R. Fritsche-Neto (Amsterdam: Elsevier), 105–134.

Reynolds, M., Chapman, S., Crespo-Herrera, L., Molero, G., Mondal, S., Pequeno, D. N. L., et al. (2020). Breeder friendly phenotyping. *Plant Sci.* 295:1103962. doi: 10.1016/j.plantsci.2019.110396

Rincent, R., Charpentier, J. P., Faivre-Rampant, P., Paux, E., Gouis, J. L., 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. *G3* 8:3961. doi: 10.1534/g3.118.200760

Risso-Pascotto, C., Pagliarini, M. S., and Valle, C. B. (2005). Meiotic behavior in interspecific hybrids between Brachiaria ruziziensis and *Brachiaria brizantha* (Poaceae). *Euphytica* 145, 155–159. doi: 10.1007/s10681-005-0893-z

Rocha, J. R. A. S. C., Marçal, T. S., Salvador, F. V., da Silva, A. C., Carneiro, P. C. S., de Resende, M. D. V., et al. (2019). Unraveling candidate genes underlying biomass digestibility in elephant grass (*Cenchrus purpureus*). *BMC Plant Biol.* 19:548. doi: 10.1186/s12870-019-2180-5

Sandhu, J. S., Kumar, D., Yadav, V. K., Singh, T., Sah, R. P., and Radhakrishna, A. (2015). "Recent trends in breeding of tropical grass and forage species," in *Proceedings of the 23rd International Grassland Congress*, eds D. Vijay, M. K. Srivastava, C. K. Gupta, D. R. Malaviya, M. M. Roy, S. K. Mahanta, et al. (Jhansi: Range Management Society of India), 337–348.

Santantonio, N., Atanda, S. A., Beyene, Y., Varshney, R. K., Olsen, M., Jones, E., et al. (2020). Strategies for effective use of genomic information in crop breeding programs serving Africa and South Asia. *Front. Plant Sci.* 11:353. doi: 10.3389/fpls.2020.00353

Shamshad, M., and Sharma, A. (2018). The usage of genomic selection strategy in plant breeding. *Next Gener. Plant Breed.* 26:93. doi: 10.5772/interchopen.76247

Shi, J., Ma, X., Zhang, J., Zhou, Y., Liu, M., Huang, L., et al. (2019). Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. *Nat. Commun.* 10:464. doi: 10.1038/s41467-018-07876-6

Silva, R. O., Barioni, L. G., Hall, J. A. J., Matsuura, M. F., Albertini, T. Z., Fernandes, F. A., et al. (2016). Increasing beef production could lower greenhouse gas emissions in Brazil if decoupled from deforestation. *Nat. Clim. Change* 6, 493–497. doi: 10.1038/nclimate2916

Simeão, R., Silva, A., Valle, C., Resende, M. D., and Medeiros, S. (2015). Genetic evaluation and selection index in tetraploid Brachiaria ruziziensis. *Plant Breed.* 135, 246–253. doi: 10.1111/pbr.12353

Simeão-Resende, R. M., Casler, M. D., and Resende, M. D. V. (2013). Selection methods in forage breeding: a quantitative appraisal. *Crop Sci.* 53, 1925–1936. doi: 10.2135/cropsci2013.03.0143

Simeão-Resende, R. M., Casler, M. D., and Resende, M. D. V. (2014). Genomic selection in forage breeding: accuracy and methods. *Crop Sci.* 54, 143–156. doi: 10.2135/cropsci2013.05.0353

Simeão-Resende, R. M., Jank, L., Valle, C. B., and Bonato, A. L. V. (2004). Biometrical analysis and selection of tetraploid progenies of *Panicum maximum* using mixed model methods. *Pesq. Agropec. Bras.* 39, 335–341. doi: 10.1590/S0100-204X2004000400006

Singh, B. P., Singh, H. P., and Obeng, E. (2013). "Elephant grass," in *Biofuel Crops: Production, Physiology and Genetics*, ed. B. P. Singh (Fort Valley, GA: CAB International), 271–291.

Stebbins, G. L. (1947). Types of polyploids: their classification and significance. *Adv. Genet.* 1, 403–429. doi: 10.1016/S0065-2660(08)60490-3

Talukder, S. K., and Saha, M. C. (2017). Toward genomics-based breeding in C3 cool-season perennial grasses. *Front. Plant Sci.* 8:1317. doi: 10.3389/fpls.2017.01317

Taylor, M. G., and Vasil, I. K. (1987). Analysis of DNA size, content and cell cycle in leaves of napier grass (*Pennisetum purpureum* Schum.). *Theor. Appl. Genet.* 74, 681–686. doi: 10.1007/BF00247541

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., et al. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* 15:823. doi: 10.1186/1471-2164-15-823

Valle, C. B., Jank, L., and Resende, R. M. S. (2009). O melhoramento de forrageiras tropicais no Brasil. *Revist. Ceres* 56, 460–472.

Valle, C. B., and Savidan, Y. H. (1996). "Genetics, cytogenetics, and reproductive biology of *Brachiaria*," in *Brachiaria: Biology, Agronomy, and Improvement*, eds J. W. Miles, B. L. Maass, and C. B. do Valle (Colombia: Embrapa), 147–163.

Valle, C. B. D. (2001). "Genetic resources for tropical areas: achievements and perspectives," in *Proceedings of the 19° International Grassland Congress*, (Piracicaba: Fundação de Estudos Agrários Luiz de Queiroz), 477–482.

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

Vigna, B. Z. B., Jungmann, L., Francisco, P. M., Zucchi, M. I., Valle, C. B., and Souza, A. P. (2011). Genetic diversity and population structure of the *Brachiaria brizantha* germplasm. *Trop. Plant Biol.* 4, 157–169. doi: 10.1007/s12042-011-9078-1

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era: concepts and misconceptions. *Nat. Rev. Genet.* 9, 255–266. doi: 10.1038/nrg2322

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., et al. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* 2:e41. doi: 10.1371/journal.pgen.0020041

Visscher, P. M., Yang, J., and Goddard, M. E. (2010). A commentary on 'Common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res. Hum. Genet.* 13, 517–524. doi: 10.1375/twin.13.6.517

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195, 1223–1230. doi: 10.1534/genetics.113.155176

Wang, C., Li, J., Zhou, S., Liu, T., Zhang, X., and Huang, L. (2018). Genome survey sequencing of purple elephant grass (Pennisetum purpureum Schum 'Zise') and identification of its SSR markers. *Mol. Breed.* 38:94. doi: 10.1007/s11032-018-0849-3

Weller, J. I. (2016). *Genomic Selection in Animals*. Hoboken, NJ: John Wiley & Sons.

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

Worthington, M., Heffelfinger, C., Bernal, D., Quintero, C., Zapata, Y. P., Perez, J. G., et al. (2016). A parthenogenesis gene candidate and evidence for segmental allopolyploidy in apomictic *Brachiaria decumbens*. *Genetics* 203, 1117–1132. doi: 10.1534/genetics.116.190314

Worthington, M., and Miles, J. W. (2015). "Reciprocal full-sib recurrent selection and tools for accelerating genetic gain in apomictic Brachiaria," in *Molecular Breeding of Forage and Turf*, eds H. Budak and G. Spangenberg (Cham: Springer International Publishing), 19–30. doi: 10.1007/978-3-319-08714-6_3

Worthington, M., Perez, J. G., Mussurova, S., Silva-Cordoba, A., Castiblanco, V., Cardoso Arango, J. A., et al. (2020). A new genome allows the identification of genes associated with natural variation in aluminum tolerance in *Brachiaria* grasses. *J. Exp. Bot.* 16:eraa469. doi: 10.1093/jxb/eraa469

Würschum, T. (2012). Maping QTL for agronomic traits in breeding populations. *Theor. Appl. Genet.* 125, 201–210. doi: 10.1007/s00122-012-1887-6

Yan, Q., Wu, F., Xu, P., Sun, Z., Li, J., Gao, L., et al. (2020). The elephant grass (*Cenchrus purpureus*) genome provides insights into anthocyanidin accumulation and fast growth. *Mol. Ecol. Resour.* 21, 526–542. doi: 10.1111/1755-0998.13271

Zhang, S., Xia, Z., Zhang, W., Li, C., Wang, X., Lu, X., et al. (2020). Chromosome-scale genome assembly provides insights into speciation of allotetraploid and massive biomass accumulation of Elephant Grass (*Pennisetum purpureum* Schum.). *bioRxiv* [Preprint] doi: 10.1101/2020.02.28.970749

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* 50, 278–284. doi: 10.1038/s41588-018-0041-z

Zou, C., Li, L., Miki, D., Li, D., Tang, Q., Xiao, L., et al. (2019). The genome of broomcorn millet. *Nat. Commun.* 10:436. doi: 10.1038/s41467-019-08409-5

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

Perspectives on Applications of Hierarchical Gene-To-Phenotype (G2P) Maps to Capture Non-stationary Effects of Alleles in Genomic Prediction

Owen M. Powell[1,3]*, *Kai P. Voss-Fels*[1], *David R. Jordan*[2,3], *Graeme Hammer*[1,3] and *Mark Cooper*[1,3]

[1] Queensland Alliance for Agriculture and Food Innovation, Centre for Crop Science, The University of Queensland, St Lucia, QLD, Australia, [2] Queensland Alliance for Agriculture and Food Innovation, Hermitage Research Facility, The University of Queensland, Warwick, QLD, Australia, [3] ARC Centre of Excellence for Plant Success in Nature and Agriculture, The University of Queensland, St Lucia, QLD, Australia

OPEN ACCESS

Edited by:
Johannes W.R. Martini,
International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Torsten Pook,
University of Göttingen, Germany
Diercles Francisco Cardoso,
São Paulo State University, Brazil
Kanwarpal Singh Dhugga,
Consultative Group on International Agricultural Research (CGIAR),
United States

***Correspondence:**
Owen M. Powell
o.powell2@uq.edu.au

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: 03 February 2021
Accepted: 13 April 2021
Published: 04 June 2021

Citation:
Powell OM, Voss-Fels KP, Jordan DR, Hammer G and Cooper M (2021) Perspectives on Applications of Hierarchical Gene-To-Phenotype (G2P) Maps to Capture Non-stationary Effects of Alleles in Genomic Prediction. Front. Plant Sci. 12:663565. doi: 10.3389/fpls.2021.663565

Genomic prediction of complex traits across environments, breeding cycles, and populations remains a challenge for plant breeding. A potential explanation for this is that underlying non-additive genetic (GxG) and genotype-by-environment (GxE) interactions generate allele substitution effects that are non-stationary across different contexts. Such non-stationary effects of alleles are either ignored or assumed to be implicitly captured by most gene-to-phenotype (G2P) maps used in genomic prediction. The implicit capture of non-stationary effects of alleles requires the G2P map to be re-estimated across different contexts. We discuss the development and application of hierarchical G2P maps that explicitly capture non-stationary effects of alleles and have successfully increased short-term prediction accuracy in plant breeding. These hierarchical G2P maps achieve increases in prediction accuracy by allowing intermediate processes such as other traits and environmental factors and their interactions to contribute to complex trait variation. However, long-term prediction remains a challenge. The plant breeding community should undertake complementary simulation and empirical experiments to interrogate various hierarchical G2P maps that connect GxG and GxE interactions simultaneously. The existing genetic correlation framework can be used to assess the magnitude of non-stationary effects of alleles and the predictive ability of these hierarchical G2P maps in long-term, multi-context genomic predictions of complex traits in plant breeding.

Keywords: multi-trait prediction, non-linear relationships, crop growth models, genetic correlation, non-additive genetic effects, epistasis, pleiotropy, GxE interactions

INTRODUCTION

Response to selection in breeding programs relies on predicting the additive genetic merit of new individuals for a target population of environments (Hallauer and Miranda, 1988; Comstock, 1996). Predicting the additive genetic merit of individuals, i.e., breeding values, requires the estimation of allele substitution effects of genetic loci (Falconer and Mackay, 1996). Both functional additive genetic effects and functional non-additive genetic effects, generated by interactions that exist within (dominance) and between (epistasis) genetic loci, contribute to estimates of allele substitution effects (Cheverud and Routman, 1996; Hill et al., 2008; Huang and Mackay, 2016). The contributions of functional additive effects to allele substitution effects are considered stationary as they are not influenced by changes in allele frequencies at genetic loci. However, the contributions of functional non-additive genetic effects (GxG interactions) to allele substitution effects are dependent on the allele frequencies of genetic loci. Therefore, changes in the genetic background can alter the predictions of allele substitution effects. Predictions of allele substitution effects can also change across environments, producing gene-by-environment (GxE) interactions. We refer to the alterations of allele substitution effects, and therefore predictions of the additive genetic merit of individuals in the presence of these interactions as non-stationary effects of alleles. In the most extreme case, allele substitution effects can change sign, i.e., from positive to negative values and *vice versa*, if changes in the value of non-stationary effects exceed the value of stationary effects (Paixão and Barton, 2016; Wientjes et al., 2021). Such sign changes in allele substitution effects change the performance landscape's optimum and influence the breeding target (Wright, 1963; Messina et al., 2011). Therefore, breeding programs need to accurately predict these non-stationary effects of alleles across different contexts to deliver the highest possible response to selection. Beyond the theoretical considerations, we consider three contexts where the potential for change in sign of allele substitution effects was identified to influence genomic prediction accuracy for commercial maize breeding for the United States corn-belt (Cooper et al., 2014a,b): breeding cycles, populations, and environments. We anticipate these considerations will also be relevant for other plant breeding situations.

Non-stationary effects of alleles decrease the accuracy of genomic predictions across breeding cycles. The accuracy of genomic prediction decreases with an increase in breeding cycles between the training and prediction set (Clark et al., 2012; Pszczola et al., 2012; Daetwyler et al., 2013; Habier et al., 2013). Changes in genetic relationships, linkage disequilibrium, and causal loci's cosegregation have been identified as important factors (Habier et al., 2013). These factors can impact GxG interactions due to changes in allele frequencies. A practical approach to account for GxG interactions in the decrease in genomic prediction accuracy over breeding cycles is periodic retraining of the genomic prediction equation (Podlich et al., 2004). However, this is costly and may exclude smaller breeding operations. The ability to estimate non-stationary effects of alleles can create opportunities to increase the persistence of prediction accuracy across breeding cycles and widen the application of genomic prediction in plant breeding.

Non-stationary effects of alleles decrease the accuracy of genomic predictions across populations. Genomic prediction across populations is important as the germplasm accessed for breeding applications is often organized in many different populations (Melchinger and Gumber, 1998; Technow et al., 2020; White et al., 2020). Across population prediction often suffers from lower accuracy than prediction across breeding cycles due to more considerable differences in allele frequencies of causal genetic loci (de Roos et al., 2009; Hayes et al., 2009). Along with mutations and redundancy of causal genetic loci, extreme differences in allele frequencies can cause discrepancies in segregation patterns of causal genetic loci between populations, which can cause large differences in allele substitution effects between populations (Rio et al., 2020). Empirical and simulation studies have shown that GxG interactions primarily determine these large changes in allele substitution effects between populations (Duenk et al., 2020; Legarra et al., 2020). Therefore, the ability to accurately capture GxG interactions in genomic prediction will be necessary to effectively utilize diverse germplasm (Tanksley and McCouch, 1997; Jordan et al., 2011; Mace et al., 2013, 2020; Gorjanc et al., 2016; Halewood et al., 2018).

Non-stationary effects of alleles decrease the accuracy of genomic predictions across environments. Genomic prediction across environments has allowed faster identification of stable performing varieties. Most methods that predict performance across environments, including GxE interactions, have been purely statistical (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966; Piepho, 1997; Burgueño et al., 2012; Crossa, 2012). With implicit knowledge of environmental effects, these methods have been shown to increase prediction accuracy within specific datasets or a well-defined target population of environments. Still, they are sensitive to changes in the target population of environments. Explicit knowledge of environmental effects can make genomic prediction across environments more robust. More recent methods have demonstrated improved prediction accuracy by explicitly including environmental covariates in genomic prediction (Heslot et al., 2014; Jarquín et al., 2014; Costa-Neto et al., 2021; Jarquin et al., 2021). However, all of these methods generate predictions conditional on current environments and therefore represent short-term predictions. Improved long-term predictions of response to selection in plant breeding, including effects of GxE interactions, will require methods to generate predictions of "best-bet" synthetic future environments (Hammer et al., 2020).

Despite the challenge of non-stationary effects of alleles, plant breeding has accurately predicted short-term response to selection to accumulate genetic gain over the long term (Duvick, 2005; Mackay et al., 2011). Short-term predictions of response to selection can mitigate non-stationary effects of alleles by conditioning predictions on current genetic backgrounds and environments. However, with the introduction of genomic prediction (Meuwissen et al., 2001), plant breeding now seeks to re-design breeding programs to further accelerate the pace of varietal development (Bernardo and Yu, 2007; Heffner et al., 2009; Gaynor et al., 2017). The increased speed of selection trajectories of new breeding strategies deploying genomic prediction places a stronger focus on plant breeding programs' ability to predict long-term response to selection. Long-term predictions of response to selection struggle to mitigate the non-stationary effects of alleles, as predictions conditional on the current genetic background and environment become increasingly uninformative into the future. An illustrative simulation example to explore these concepts is provided in the **Supplementary Information**.

In this perspective, we discuss a few lessons learned from applying hierarchical gene-to-phenotype (G2P) maps in predictive breeding and our view of promising future research directions to realize improvements in the prediction of long-term response to selection in plant breeding.

PERSPECTIVE

Improvements in prediction from the specification of interactions require thorough interrogation of the underlying G2P maps of complex traits (Houle et al., 2010; Marjoram et al., 2014). The genetic architecture of traits, which details the number, distribution of effect sizes, and "behavior" of these causal genetic variants, can be viewed as a G2P map. Therefore, the G2P map defines the complete paths from causal genetic variants to the phenotype of complex traits (Waddington, 1957; Burns, 1970; Lewontin, 1974). The dominant G2P map used to investigate the role of interactions in response to selection is a single complex trait underpinned by the infinitesimal model (Robertson, 1960; Carlborg et al., 2006; Hill et al., 2008; Mäki-Tanila and Hill, 2014; Goodnight, 2015; Paixão and Barton, 2016; Wientjes et al., 2021). The infinitesimal model allows breeders to consider complex phenotypes in a single trait context, with underlying genetic variation associated directly with the phenotypic variation of complex traits within a reference population of genotypes (**Figure 1A**). The infinitesimal model, embedded within the breeders equation (Lush, 1937), has been successful in plant breeding (Hallauer and Miranda, 1988; Comstock, 1996). However, alternative G2P maps have been developed. Here we consider their potential for breeding applications.

Hierarchical G2P maps provide a multi-trait context for investigations into the importance of interactions in genomic prediction. Complex trait phenotypes, such as grain yield, can be viewed as the product of multiple component traits. The hierarchical structure allows intermediate processes (**Figure 1B**), such as other traits and environmental factors and their interactions, to contribute to complex trait variation (Wright, 1934; Waddington, 1957; Houle et al., 2010; Liu et al., 2019; Cooper et al., 2020a).

In quantitative genetics, hierarchical G2P maps have been developed based on path analysis (Wright, 1934). The specification of intermediate processes in hierarchical G2P maps allows the decomposition of total effects, captured by the infinitesimal G2P map, into path specific direct and indirect effects (Wright, 1934). Lande and Arnold (1983) demonstrated that hierarchical G2P maps could be used to separate direct response to selection from indirect response to selection of multiple correlated traits. Valente et al. (2013) provide an overview of the breeding applications of Structural Equation Models (Gianola and Sorensen, 2004; Pearl, 2012) and highlight their ability to allow prediction across a broader range of livestock and crop management practices than standard multi-trait models without requiring frequent re-estimation of the G2P map. Recently, there has been an increase in the use of Structural Equation Models for prediction and inference in both animal and plant breeding (Tiezzi et al., 2015; Momen et al., 2018; Campbell et al., 2019; Pegolo et al., 2020; Abdalla et al., 2021). However, due to a lack of prior knowledge of the underlying relationships, most studies have used Structural Equation Models to estimate linear relationships between traits. The assumption of linear relationships restricts the range and magnitude of non-stationary effects and, therefore, the frequency of rank changes in additive genetic merit.

In plant science, decades of experiments led to the development of hierarchical G2P maps for plant breeding that allow predictions across a wide range of growing conditions (Holzworth et al., 2014; Hammer et al., 2019). Crop Growth Models are hierarchical mechanistic models of plants that simulate trajectories of multiple trait phenotypes over time for the growing season determined by environmental conditions. Crop Growth Models explicitly quantify the relationships, both linear and non-linear, between traits, physiological "meta-mechanisms" and complex trait phenotypes such as grain yield. These "meta-mechanisms" are measurable via high-throughput phenotyping and resulting in robust and stable equations with heritable genotype-dependent parameters (Tardieu et al., 2020). This has allowed Crop Growth Models to be linked to underlying genotypic variation for plant breeding applications (Chapman et al., 2003; Chenu et al., 2009; Messina et al., 2011). More recently, Crop Growth Model – Whole Genome Prediction methods have connected an underlying "infinitesimal" genetic architecture to key components of Crop Growth Models via a hierarchical Bayesian estimation procedure (**Figure 2**; Technow et al., 2015; Cooper et al., 2016). The inclusion of Crop Growth Models in genomic prediction enables the prediction of trait-trait and trait-environment interactions in the hierarchy's upper levels, which are directly associated with the estimates of allele substitution effects of genetic parameters for traits in the lower levels of the crop growth model hierarchy. This correction of phenotypes can lead to improved estimates of genetic correlations between traits and increased prediction accuracies across the different contexts discussed above. Crop Growth Model – Whole Genome Prediction methods, and subsequent variations, have been shown to improve short-term predictions of genetic merit in the presence of GxE interactions (Bustos-Korts et al., 2019; Millet et al., 2019; Robert et al., 2020; Toda et al., 2020; Diepenbrock et al., 2021) and genotype-by-environment-by-management interactions in plant breeding. The success of hierarchical G2P maps in capturing non-stationary effects in predictions across diverse environments has seen growth models being revisited in animal breeding (Doeschl-Wilson et al., 2007; Puillet et al., 2016, 2021).

However, the prediction of long-term response to selection remains a significant challenge (Reeve, 2000; Goddard, 2009; Hill, 2017). For example, long-term selection experiments in maize have often produced results not predictable *a priori* or from simulation (Lamkey, 1992; Dudley and Lambert, 2003), such as continued selection response after 100 years

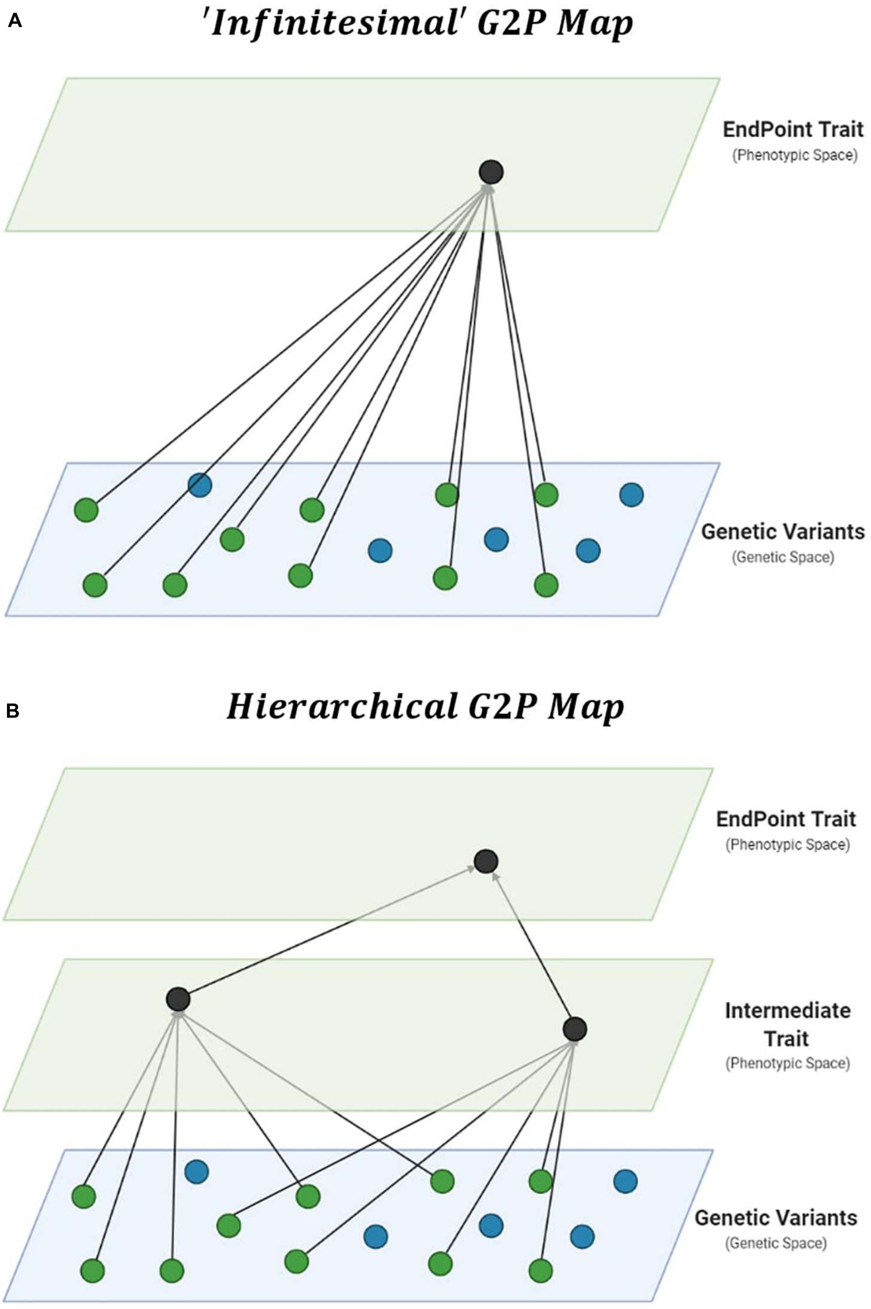


FIGURE 1 | Gene-to-Phenotype (G2P) Maps. **(A)** Representation of an additive infinitesimal G2P map, assuming direct effects of causal genetic variants (green circles) on complex trait phenotypes. **(B)** Representation of an additive hierarchical G2P map, decomposing total effects into direct effects of causal genetic variants on intermediate traits, and phenotypic effects of multiple intermediate traits on complex trait phenotypes.

(Dudley and Lambert, 2003). Long-term predictions of response to selection, based on the classical versions of the infinitesimal model (Walsh and Lynch, 2018), struggle to accurately predict the non-stationary effects of alleles as information from current genetic backgrounds and environments become increasingly uninformative into the future. A key paper by Paixão and Barton (2016), extending Robertson's (1960) work with only functional additive effects, has clarified the importance of non-stationary effects of alleles generated by GxG interactions for long-term response to selection. They describe two explicit scenarios: (i) when drift dominates selection, i.e., when the selection pressure at individual functional loci is weak, the initial variance components will determine the increase in response to selection over breeding cycles due to interactions; (ii) when selection dominates drift, i.e., when the selection pressure at individual functional loci is strong, the initial variance components are

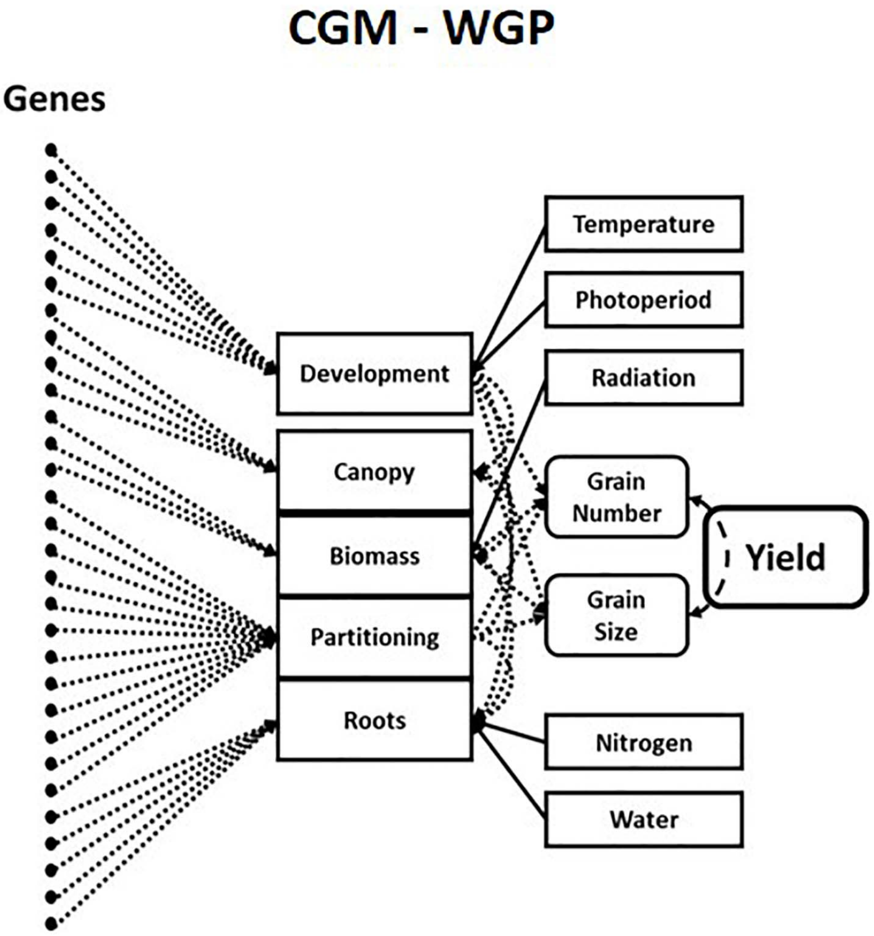


FIGURE 2 | Schematic representation of a hierarchical crop growth model whole genome prediction (CGM-WGP) G2P map. Taken from Figure 2b of Cooper et al. (2020a). Genetic variants are associated with traits or "meta-mechanisms" at lower levels in the crop growth model hierarchy to predict traits at higher levels in the hierarchy.

poor predictors of the response to selection over breeding cycles and details of the G2P map need to be explicitly considered. Therefore, to quantify the importance of non-stationary effects of alleles in predicting long-term response to selection in plant breeding, we should consider two questions:

i. What is the strength of selection operating on the causal loci for traits in breeding programs?
ii. If selection operating on the causal loci is strong, what is the underlying G2P map?

The availability of dense genotype data, sequence data, and advances in phenotyping provide the opportunity to revisit theories about the strength of selection in plant breeding programs. Before the ability to study allelic variation via genotype data, the selection units of breeding programs were breeding values of individuals. It has been shown for complex traits that strong selection at the individual level does not necessarily translate to strong selection at the causal loci (Goddard, 2009; Walsh and Lynch, 2018). However, technologies such as genomic prediction (Meuwissen et al., 2001) are shifting the selection units of breeding programs toward the allele substitution effects of genetic loci. Despite selection still occurring on individuals, genomic selection can distribute selection pressure unevenly across the genome by directing selection pressure to genetic loci with large estimated allele substitution effects (Heidaritabar et al., 2016; Wientjes et al., 2021). Therefore, the use of genomic selection in breeding programs can result in selection dominating drift at specific genetic loci placing greater importance on the G2P map assumed in genomic predictions.

Complete knowledge of the underlying G2P maps of complex traits is unlikely. However, hierarchical G2P maps with partial knowledge of intermediate processes offer promise for predicting long-term response to selection, given their success in improved short-term predictions of non-stationary effects of alleles. An obstacle in the practical applications of such hierarchical G2P modeling approaches is non-identifiability, also referred to as equifinality or the many-to-one property (Lamsal et al., 2018; Barghi et al., 2020; Henshaw et al., 2020; Kruijer et al., 2020; Tsutsumi-Morita et al., 2021). Effects can be non-identifiable due to unmeasured confounders that generate correlated errors between effects, which results in multiple, equally likely hierarchical G2P maps for experimental data sets. As an example, a multi-trait G2P map involving GxG interactions and the summation of Trait 1 and Trait 2 (**Figure 3A**) could equally be parameterized as the simplified Crop Growth Model – Whole Genome Prediction G2P map of two traits with purely additive functional genetic effects and non-linear relationships between traits (**Figure 3B**). Therefore, the level of detail required in hierarchical G2P maps to overcome non-identifiability is still an active research area.

FUTURE DIRECTIONS

In recent times, genomic prediction across multiple contexts has received increased focus in breeding (de Roos et al., 2009; Hayes et al., 2009; Windhausen et al., 2012; Gorjanc et al., 2016; Montesinos-López et al., 2019). In a multi-context setting, the genetic correlation naturally provides a measure to quantify predictive accuracy (Falconer, 1952; Robertson, 1959; Bohren et al., 1966). To maximize the benefits of using the genetic correlation framework, plant breeding requires hierarchical G2P maps that include the explicit specification of interactions (**Figure 3C**). Specification of gene-gene interactions would allow the assessment of changes in the genetic background on GxG interactions and prediction accuracy. Specification of gene-trait and trait-trait interactions would allow the assessment of changes in the environment and agronomic management on GxE interactions and prediction accuracy. Breeding programs are often organized in many different populations or regions to limit these impacts of GxG and GxE interactions, respectively, while assuming a single performance optimum and single breeding target. However, GxG or GxE interactions can generate a performance landscape with multiple optima (Wright, 1963; Cooper et al., 2005; Messina et al., 2011; Technow et al., 2020). Prior specification of this multiple optima landscape, via hierarchical G2P maps, would allow more comprehensive explorations of the impact of such interactions on the long-term response to selection of plant breeding programs.

Complementary simulation and empirical studies can interrogate the changes of genetic correlations across contexts to quantify the relative magnitude of GxG and GxE interactions and measure their impact on genomic prediction. Recent research, primarily from animal breeding, has renewed the focus on this framework (Wientjes et al., 2015; Dai et al., 2020; Duenk et al., 2020; Legarra et al., 2020). The common theme has been using the genetic correlation to assess likely magnitudes of GxG interactions underpinning complex traits. Duenk et al. (2020) used simulations to show that realistic levels of dominance alone could not drive the genetic correlation between two populations

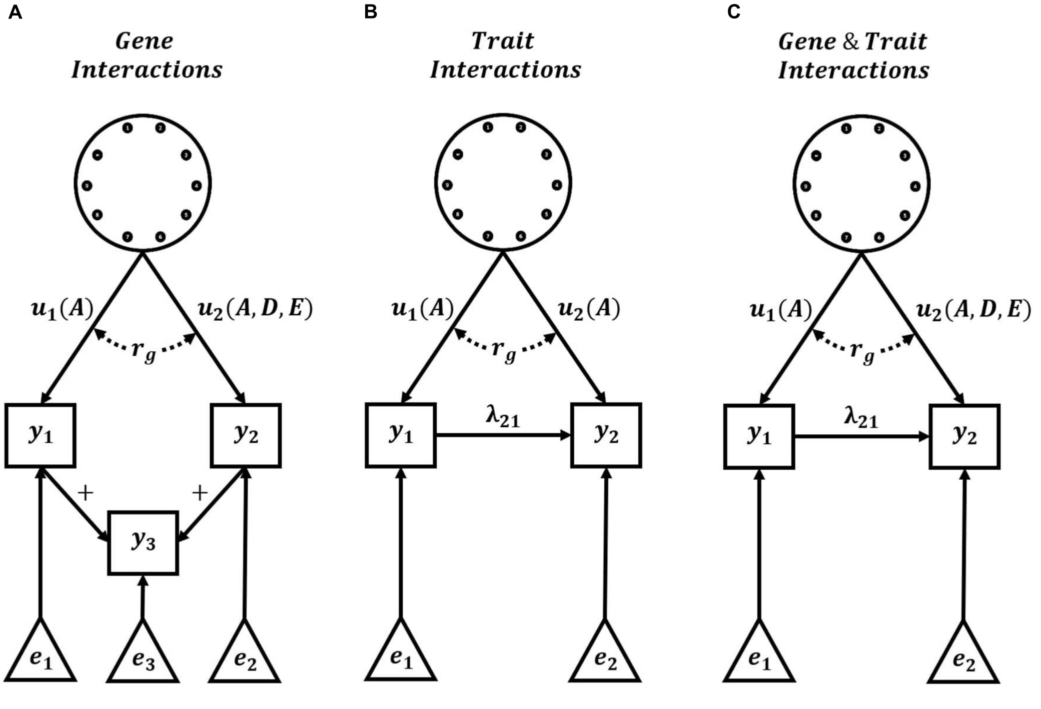


FIGURE 3 | Hierarchical G2P Maps for Plant Breeding. Examples of three multi-trait hierarchical G2P maps with the explicit specification of interactions. Hierarchical G2P maps incorporating knowledge of trait interactions (+, λ) can be used to adjust phenotypes and increase the accuracy of the estimation of gene effects (*u*), gene interactions, and genetic correlations (r_g) between traits. Gene effects (*u*) can be directly assigned to trait phenotypes (*y*) or indirectly assigned via linear trait relationships (+) or non-linear trait interactions (λ). *A, D,* and *E* indicate additive, dominance, and epistatic functional genetic effects, respectively. Non-genetic effects of trait phenotypes are represented by *e*. **(A)** Representation of a G2P map with gene interactions and linear relationship between trait phenotypes, **(B)** Representation of current Crop Growth Model – Whole Genome Prediction (CGM -WGP) G2P maps with additive genetic effects and non-linear trait interactions, and **(C)** Representation of potential G2P maps with both gene interactions and non-linear trait interactions.

below 0.8, but realistic levels of epistasis could drive the genetic correlation as low as 0.45. Legarra et al. (2020) used two regularly intermated populations with similar allele frequencies and an expectation of minimal GxG interactions to speculate on the role of GxE in low across population predictions. They also suggested a genetic correlation threshold of 0.6, below which populations should be classed as distinct. However, these recent animal breeding studies overlooked the inclusion of GxE interaction scenarios. GxE interaction scenarios are of high relevance to plant breeding which regularly predict across a diverse set of target population of environments. Plant breeding is in a prime position to use results from evolutionary genetics (de Villemereuil et al., 2016), multi-environment trial analyses (Piepho, 1997; van Eeuwijk et al., 2005; Malosetti et al., 2013), and Crop Growth Models (Jones et al., 2003; Hammer et al., 2010; Messina et al., 2011; Holzworth et al., 2014) to assess the impact of GxE interactions on genetic correlations and determine their influence on breeding programs designed to utilize genomic prediction. Therefore, we propose that the plant breeding community undertake complementary simulation and empirical studies to quantify the relative magnitude of GxG and GxE interactions across relevant environmental and population contexts to quantify their impact on genomic prediction.

The dominant crop improvement procedure of today is a sequential operation. Breeding programs first develop new varieties with a limited sampling of the full range of farmers' agronomic possibilities. Within this first step, plant breeding programs simultaneously perform population improvement to improve the additive genetic merit of breeding germplasm and product development, to identify new varieties with the highest total genotypic merit (Messina et al., 2011; Powell et al., 2020; Technow et al., 2020; Werner et al., 2020). Then agronomic research programs follow, focusing on developing and optimizing crop management strategies for the handful of new varieties. Hierarchical G2P maps can connect the objectives of plant breeding and quantitative genetics with those of crop agronomy (**Figure 3**; Cooper et al., 2020a,b). The explicit connections between gene and multiple trait levels, embedded in hierarchical G2P maps, can be perturbed experimentally (empirical and simulation) to quantify the impact of agronomic management interventions and changes in the environment. The effects of the perturbations can be investigated to determine how they propagate through the hierarchical G2P map and update estimates of allele effects at both the gene and trait levels. *Ex-ante* predictions of perturbations at the gene level could be used to guide improved prediction of "synthetic" varieties developed through novel gene-editing techniques. *Ex-ante* predictions of perturbations at the trait level could improve the efficiency of breeding new varieties adapted for alternative farming systems and future climate scenarios (Hammer et al., 2020). At the same time, predictions can be extracted from each level of the hierarchical G2P map, allowing the decomposition of individual performance into additive genetic, total genetic, and phenotypic merit. Decomposition of path-specific values in hierarchical G2P maps has been demonstrated in evolutionary and quantitative genetics (Lande and Arnold, 1983; Gianola and Sorensen, 2004; Valente et al., 2010, 2013; Henshaw et al., 2020; Janeiro et al., 2020; Pegolo et al., 2020). Therefore, the ability to exploit different sources of improved crop performance under a single prediction framework could improve crop improvement pipelines' accuracy and flexibility to navigate performance landscapes for current and future environments (Messina et al., 2011, 2020; Technow et al., 2020).

CONCLUSION

Current genomic prediction methods struggle to predict the non-stationary effects of alleles as the genetic background (breeding cycles and populations) and the environment changes. These non-stationary effects of alleles are determined by interactions between genetic loci, traits, and the environment. Non-stationary effects of alleles result in low prediction accuracy across breeding cycles, populations and environments. As discussed above, the development of hierarchical G2P maps has been shown to improve the genomic prediction of non-stationary effects of alleles across breeding cycles and environments. The simultaneous specification of GxG and GxE interactions in hierarchical G2P maps may help to more thoroughly explore the impact of non-stationary effects of alleles on the long-term response to selection of plant breeding programs.

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

OP and MC conceived and designed the perspective. OP wrote the first manuscript draft and developed the supporting simulations. MC, KV-F, DJ, and GH helped to refine the manuscript. All authors read and approved the final manuscript.

FUNDING

Contribution supported by the Australian Research Council Centre of Excellence for Plant Success in Nature and Agriculture (CE200100015) and the Australian Grains Research and Development Corporation project UOQ1903-008RTX. KV-F was supported by an Australian Research Council Discovery Early Career Research Award (DE210101407).

SUPPLEMENTARY INFORMATION

A walkthrough of a simulation, data files, and scripts demonstrating non-stationary effects of alleles over breeding cycles can be accessed at https://powellow.github.io/Interactions_In_Breeding/.

REFERENCES

Abdalla, E. A., Wood, B. J., and Baes, C. F. (2021). Accuracy of breeding values for production traits in turkeys (*Meleagris gallopavo*) using recursive models with or without genomics. *Genet. Sel. Evol.* 53:16. doi: 10.1186/s12711-021-00611-8

Barghi, N., Hermisson, J., and Schlötterer, C. (2020). Polygenic adaptation: a unifying framework to understand positive selection. *Nat. Rev. Genet.* 21, 769–781. doi: 10.1038/s41576-020-0250-z

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

Bohren, B. B., Hill, W. G., and Robertson, A. (1966). Some observations on asymmetrical correlated responses to selection. *Genet. Res.* 7, 44–57. doi: 10.1017/S0016672300009460

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

Burns, J. (1970). "The synthetic problem and the genotype-phenotype relation in cellular metabolism," in *Towards a Theoretical Biology*, ed C. H. Waddington (New Brunswick, NJ: Transaction Publishers), 47–51.

Bustos-Korts, D., Malosetti, M., Chenu, K., Chapman, S., Boer, M. P., Zheng, B., et al. (2019). From QTLs to adaptation landscapes: using genotype-to-phenotype models to characterize G×E over time. *Front. Plant Sci.* 10:1540. doi: 10.3389/fpls.2019.01540

Campbell, M. T., Yu, H., Momen, M., and Morota, G. (2019). Examining the relationships between phenotypic plasticity and local environments with genomic structural equation models. *bioRxiv[perprint]* doi: 10.1101/2019.12.11.873257

Carlborg, Ö, Jacobsson, L., Åhgren, P., Siegel, P., and Andersson, L. (2006). Epistasis and the release of genetic variation during long-term selection. *Nat. Genet.* 38, 418–420. doi: 10.1038/ng1761

Chapman, S., Cooper, M., Podlich, D., and Hammer, G. (2003). Evaluating plant breeding strategies by simulating gene action and dryland environment effects. *Agron. J.* 95, 99–113. doi: 10.2134/agronj2003.9900

Chenu, K., Chapman, S. C., Tardieu, F., McLean, G., Welcker, C., and Hammer, G. L. (2009). Simulating the yield impacts of organ-level quantitative trait loci associated with drought response in maize: a """""" gene-to-phenotype"""""""". *Mod. Approach. Genet.* 183, 1507–1523. doi: 10.1534/genetics.109.105429

Cheverud, J. M., and Routman, E. J. (1996). Epistasis as a source of increased additive genetic variance at population bottlenecks. *Evolution* 50, 1042–1051. doi: 10.1111/j.1558-5646.1996.tb02345.x

Clark, S. A., Hickey, J. M., 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.* 44:4. doi: 10.1186/1297-9686-44-4

Comstock, R. E. (1996). *Quantitative Genetics With Special Reference to Plant and Animal Breeding*, 1st Edn. Ames: Iowa State University Press.

Cooper, M., Gho, C., Leafgren, R., Tang, T., and Messina, C. (2014a). Breeding drought-tolerant maize hybrids for the US corn-belt: discovery to product. *J. Exp. Bot.* 65, 6191–6204. doi: 10.1093/jxb/eru064

Cooper, M., Messina, C. D., Podlich, D., Totir, L. R., Baumgarten, A., Hausmann, N. J., et al. (2014b). 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. W., and Smith, O. S. (2005). Gene-to-phenotype models and complex trait genetics. *Aust. J. Agric. Res.* 56, 895–918. doi: 10.1071/AR05154

Cooper, M., Powell, O., Voss-Fels, K. P., Messina, C. D., Gho, C., Podlich, D. W., et al. (2020a). Modelling selection response in plant breeding programs using crop models as mechanistic gene-to-phenotype (CGM-G2P) multi-trait link functions. *Silico Plants* 3: diaa016. doi: 10.1093/insilicoplants/diaa016

Cooper, M., Tang, T., Gho, C., Hart, T., Hammer, G., and Messina, C. (2020b). Integrating genetic gain and gap analysis to predict improvements in crop productivity. *Crop Sci.* 60, 582–604. doi: 10.1002/csc2.20109

Cooper, M., Technow, F., Messina, C., Gho, C., and Totir, L. R. (2016). Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial. *Crop Sci.* 56, 2141–2156. doi: 10.2135/cropsci2015.08.0512

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2021). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity* 126, 92–106. doi: 10.1038/s41437-020-00353-1

Crossa, J. (2012). From genotype × environment interaction to gene × environment interaction. *Curr. Genom.* 13, 225–244. doi: 10.2174/138920212800543066

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

Dai, Z., Long, N., and Huang, W. (2020). Influence of genetic interactions on polygenic prediction. *G3amp58 Genes Genom. Genet.* 10, 109–115. doi: 10.1534/g3.119.400812

de Roos, A. P. W., Hayes, B. J., and Goddard, M. E. (2009). Reliability of genomic predictions across multiple populations. *Genetics* 183, 1545–1553. doi: 10.1534/genetics.109.104935

de Villemereuil, P., Gaggiotti, O. E., Mouterde, M., and Till-Bottraud, I. (2016). Common garden experiments in the genomic era: new perspectives and opportunities. *Heredity* 116, 249–254. doi: 10.1038/hdy.2015.93

Diepenbrock, C., Tang, T., Jines, M., Technow, F., Lira, S., Podlich, D., et al. (2021). Can we harness digital technologies and physiology to hasten genetic gain in U.S. maize breeding? *bioRxiv[preprint]* doi: 10.1101/2021.02.23.432477

Doeschl-Wilson, A. B., Knap, P. W., Kinghorn, B. P., and Van der Steen, H. A. M. (2007). Using mechanistic animal growth models to estimate genetic parameters of biological traits. *Animal* 1, 489–499. doi: 10.1017/S1751731107691848

Dudley, J. W., and Lambert, R. J. (2003). “""""" 100 Generations of selection for oil and protein in corn""""","” in *Plant Breeding Reviews*, ed. J. Jules (John Wiley & Sons, Ltd), 79–110. doi: 10.1002/9780470650240.ch5

Duenk, P., Bijma, P., Calus, M. P. L., Wientjes, Y. C. J., and van der Werf, J. H. J. (2020). The impact of non-additive effects on the genetic correlation between populations. *G3amp58 Genes Genom. Genet.* 10, 783–795. doi: 10.1534/g3.119.400663

Duvick, D. N. (2005). Genetic progress in yield of united states maizE (*Zea mays* L.). *Maydica* 50, 193–202.

Eberhart, S. A., and Russell, W. A. (1966). Stability parameters for comparing varieties1. *Crop Sci.* 6, 36–40. doi: 10.2135/cropsci1966.0011183X000600010011x

Falconer, D. S. (1952). The problem of environment and selection. *Am. Nat.* 86, 293–298.

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

Finlay, K., and Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Aust. J. Agric. Res.* 14:742. doi: 10.1071/AR9630742

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

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

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

Goodnight, C. (2015). Long-term selection experiments: epistasis and the response to selection. *Methods Mol. Biol. Clifton NJ* 1253, 1–18. doi: 10.1007/978-1-4939-2155-3_1

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* 17:30. doi: 10.1186/s12864-015-2345-z

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

Halewood, M., Chiurugwi, T., Sackville Hamilton, R., Kurtz, B., Marden, E., Welch, E., et al. (2018). Plant genetic resources for food and agriculture: opportunities and challenges emerging from the science and information technology revolution. *New Phytol.* 217, 1407–1419. doi: 10.1111/nph.14993

Hallauer, A. R., and Miranda, J. B. F. (1988). *Quantitative Genetics in Maize Breeding*, 2nd Edn. Ames: Iowa State University Press.

Hammer, G., Messina, C., Wu, A., and Cooper, M. (2019). Biological reality and parsimony in crop modelswhy we need both in crop improvement! *Silico Plants* 1:diz010. doi: 10.1093/insilicoplants/diz010

Hammer, G. L., McLean, G., Oosterom, E., van, Chapman, S., Zheng, B., et al. (2020). Designing crops for adaptation to the drought and high-temperature risks anticipated in future climates. *Crop Sci.* 60, 605–621. doi: 10.1002/csc2.20110

Hammer, G. L., van Oosterom, E., McLean, G., Chapman, S. C., Broad, I., Harland, P., et al. (2010). Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops. *J. Exp. Bot.* 61, 2185–2202. doi: 10.1093/jxb/erq095

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., and Goddard, M. E. (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41:51. doi: 10.1186/1297-9686-41-51

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

Heidaritabar, M., Calus, M. P. L., Megens, H.-J., Vereijken, A., Groenen, M. A. M., and Bastiaansen, J. W. M. (2016). Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. *J. Anim. Breed. Genet.* 133, 167–179. doi: 10.1111/jbg.12199

Henshaw, J. M., Morrissey, M. B., and Jones, A. G. (2020). Quantifying the causal pathways contributing to natural selection. *Evolution* 74, 2560–2574. doi: 10.1111/evo.14091

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. *TAG Theor. Appl. Genet. Theor. Angew. Genet.* 127, 463–480. doi: 10.1007/s00122-013-2231-5

Hill, W. G. (2017). """"Conversion"""" of epistatic into additive genetic variance in finite populations and possible impact on long-term selection response. *J. Anim. Breed. Genet.* 134, 196–201. doi: 10.1111/jbg.12270

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

Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., et al. (2014). apsimevolution towards a new generation of agricultural systems simulation. *Environ. Model. Softw.* 62, 327–350. doi: 10.1016/j.envsoft.2014.07.009

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nat. Rev. Genet.* 11, 855–866. doi: 10.1038/nrg2897

Huang, W., and Mackay, T. F. C. (2016). The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *bioRxiv[preprint]* doi: 10.1101/041434 bioRxiv, 041434,

Janeiro, M. J., Henshaw, J. M., Pemberton, J. M., Pilkington, J. G., and Morrissey, M. B. (2020). Selection of lamb size and early pregnancy in Soay sheep (*Ovies aries*). *bioRxiv [preprint]* doi: 10.1101/2020.09.16.299685

Jarquín, D., Crossa, J., Lacaze, X., Cheyron, P. D., 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, C., Bohn, M., Buckler, E. S., Ciampitti, I., et al. (2021). 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

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., et al. (2003). The DSSAT cropping system model. *Eur. J. Agron.* 18, 235–265. doi: 10.1016/S1161-0301(02)00107-7

Jordan, D. R., Mace, E. S., Cruickshank, A. W., Hunt, C. H., and Henzell, R. G. (2011). Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. *Crop Sci.* 51, 1444–1457. doi: 10.2135/cropsci2010.06.0326

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

Lamkey, K. R. (1992). Fifty years of recurrent selection in the Iowa stiff stalk synthetic maize population. *Maydica* 37, 19–28.

Lamsal, A., Welch, S. M., White, J. W., Thorp, K. R., and Bello, N. M. (2018). Estimating parametric phenotypes that determine anthesis date in Zea mays: Challenges in combining ecophysiological models with genetics. *PLoS One* 13:e0195841. doi: 10.1371/journal.pone.0195841

Lande, R., and Arnold, S. J. (1983). The measurement of selection on correlated characters. *Evolution* 37:1210. doi: 10.2307/2408842

Legarra, A., Garcia-Baccino, C. A., Wientjes, Y. C. J., and Vitezica, Z. G. (2020). The correlation of substitution effects across populations and generations in the presence of non-additive functional gene action. *bioArxiv*. doi: 10.1101/2020.11.03.367227

Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change*. New York, NY: Columbia University Press.

Liu, X., Li, Y. I., and Pritchard, J. K. (2019). Trans effects on gene expression can drive omnigenic inheritance. *Cell* 177, 1022–1034e6. doi: 10.1016/j.cell.2019.04.014

Lush, J. L. (1937). *Animal Breeding Plans*. Iowa: State Press.

Mace, E. S., Cruickshank, A. W., Tao, Y., Hunt, C. H., and Jordan, D. R. (2020). A global resource for exploring and exploiting genetic variation in sorghum crop wild relatives. *Crop Sci.* 61, 150–162. doi: 10.1002/csc2.20332

Mace, E. S., Tai, S., Gilding, E. K., Li, Y., Prentis, P. J., Bian, L., et al. (2013). Whole-genome sequencing reveals untapped genetic potential in ' 'Africa's indigenous cereal crop sorghum. *Nat. Commun.* 4:2320. doi: 10.1038/ncomms3320

Mackay, I., Horwell, A., Garner, J., White, J., McKee, J., and Philpott, H. (2011). Reanalyses of the historical series of UK variety trials to quantify the contributions of genetic and environmental factors to trends and variability in yield over time. *Theor. Appl. Genet.* 122, 225–238. doi: 10.1007/s00122-010-1438-y

Mäki-Tanila, A., and Hill, W. G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics* 198, 355–367. doi: 10.1534/genetics.114.165282

Malosetti, M., Ribaut, J.-M., and van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Front. Physiol.* 4:44. doi: 10.3389/fphys.2013.00044

Marjoram, P., Zubair, A., and Nuzhdin, S. V. (2014). Post-GWAS: where next? More samples, more SNPs or more biology? *Heredity* 112, 79–88. doi: 10.1038/hdy.2013.52

Melchinger, A. E., and Gumber, R. K. (1998). Overview of heterosis and heterotic groups in agronomic crops. Concepts Breed. *Heterosis Crop Plants* 25, 29–44. doi: 10.2135/cssaspecpub25.c3

Messina, C. D., Cooper, M., Hammer, G. L., Berning, D., Ciampitti, I., Clark, R., et al. (2020). Two decades of creating drought tolerant maize and underpinning prediction technologies in the US corn-belt: review and perspectives on the future of crop design. *[preprint]* doi: 10.1101/2020.10.29.361337

Messina, C. D., Podlich, D., Dong, Z., Samples, M., and Cooper, M. (2011). Yield–trait performance landscapes: from theory to application in breeding maize for drought tolerance. *J. Exp. Bot.* 62, 855–868. doi: 10.1093/jxb/erq329

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.

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

Momen, M., Ayatollahi Mehrgardi, A., Amiri Roudbar, M., Kranis, A., Mercuri Pinto, R., Morota, G., et al. (2018). Including phenotypic causal networks in genome-wide association studies using mixed effects structural equation models. *Front. Genet.* 9:455. doi: 10.3389/fgene.2018.00455

Montesinos-López, O. A., Montesinos-López, A., Tuberosa, R., Maccaferri, M., Sciara, G., and Ammar, K. (2019). Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. *Front. Plant Sci.* 10:1311. doi: 10.3389/fpls.2019.01311

Paixão, T., and Barton, N. H. (2016). The effect of gene interactions on the long-term response to selection. *Proc. Natl. Acad. Sci. U.S.A.* 113, 4422–4427. doi: 10.1073/pnas.1518830113

Pearl, J. (2012). *The Causal Foundations of Structural Equation Modeling*.Fort Belvoir, VA: Defense Technical Information Center, doi: 10.21236/ADA557445

Pegolo, S., Momen, M., Morota, G., Rosa, G. J. M., Gianola, D., Bittante, G., et al. (2020). Structural equation modeling for investigating multi-trait genetic architecture of udder health in dairy cattle. *Sci. Rep.* 10:7751. doi: 10.1038/s41598-020-64575-3

Piepho, H.-P. (1997). Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* 53, 761–766. doi: 10.2307/2533976

Podlich, D. W., Winkler, C. R., and Cooper, M. (2004). Mapping as you go: an effective approach for marker-assisted selection of complex traits. *Crop Sci.* 44, 1560–1571. doi: 10.2135/cropsci2004.1560

Powell, O., Gaynor, R. C., Gorjanc, G., Werner, C. R., and Hickey, J. M. (2020). A two-part strategy using genomic selection in hybrid crop breeding programs. *bioArxiv* doi: 10.1101/2020.05.24.113258

Pszczola, M., Strabel, T., Mulder, H. 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

Puillet, L., Ducrocq, V., Friggens, N. C., and Amer, P. R. (2021). Exploring underlying drivers of genotype by environment interactions in feed efficiency traits for dairy cattle with a mechanistic model involving energy acquisition and allocation. *J. Dairy Sci.* 104, 5805–5816. doi: 10.3168/jds.2020-19610

Puillet, L., Réale, D., and Friggens, N. C. (2016). Disentangling the relative roles of resource acquisition and allocation on animal feed efficiency: insights from a dairy cow model. *Genet. Sel. Evol.* 48:72. doi: 10.1186/s12711-016-0251-8

Reeve, J. P. (2000). Predicting long-term response to selection. *Genet. Res.* 75, 83–94. doi: 10.1017/S0016672399004140

Rio, S., Mary-Huard, T., Moreau, L., Bauland, C., Palaffre, C., Madur, D., et al. (2020). Disentangling group specific QTL allele effects from genetic background epistasis using admixed individuals in GWAS: An application to maize flowering. *PLoS Genet.* 16:e1008241. doi: 10.1371/journal.pgen.1008241

Robert, P., Le Gouis, J., Consortium, T. B., and Rincent, R. (2020). Combining crop growth modeling with trait-assisted prediction improved the prediction of genotype by environment interactions. *Front. Plant Sci.* 11:827. doi: 10.3389/fpls.2020.00827

Robertson, A. (1959). The sampling variance of the genetic correlation coefficient. *Biometrics* 15:469. doi: 10.2307/2527750

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

Tanksley, S. D., and McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277, 1063–1066. doi: 10.1126/science.277.5329.1063

Tardieu, F., Granato, I. S. C., Van Oosterom, E. J., Parent, B., and Hammer, G. L. (2020). Are crop and detailed physiological models equally "mechanistic' for predicting the genetic variability of whole-plant behaviour? The nexus between mechanisms and adaptive strategies. *Silico Plants* 2:diaa011. doi: 10.1093/insilicoplants/diaa011

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

Technow, F., Podlich, D., and Cooper, M. (2020). Back to the future: implications of genetic complexity for hybrid breeding strategies. *bioRxiv [preprint]* doi: 10.1101/2020.10.21.349332 bioRxiv, 2020.10.21.349332,

Tiezzi, F., Valente, B. D., Cassandro, M., and Maltecca, C. (2015). Causal relationships between milk quality and coagulation properties in Italian Holstein-Friesian dairy cattle. *Genet. Sel. Evol.* 47:45. doi: 10.1186/s12711-015-0123-7

Toda, Y., Wakatsuki, H., Aoike, T., Kajiya-Kanegae, H., Yamasaki, M., Yoshioka, T., et al. (2020). Predicting biomass of rice with intermediate traits: modeling method combining crop growth models and genomic prediction models. *PLoS One* 15:e0233951. doi: 10.1371/journal.pone.0233951

Tsutsumi-Morita, Y., Heuvelink, E., Khaleghi, S., Bustos-Korts, D., Marcelis, L. F. M., Vermeer, K. M. C. A., et al. (2021). Yield dissection models to improve yield; a case study in tomato. *Silico Plants* 3:diab012. doi: 10.1093/insilicoplants/diab012

Valente, B. D., Rosa, G. J. M., de los Campos, G., Gianola, D., and Silva, M. A. (2010). Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics* 185, 633–644. doi: 10.1534/genetics.109.112979

Valente, B. D., Rosa, G. J. M., Gianola, D., Wu, X.-L., and Weigel, K. (2013). Is structural equation modeling advantageous for the genetic improvement of multiple traits? *Genetics* 194, 561–572. doi: 10.1534/genetics.113.151209

van Eeuwijk, F. A., Malosetti, M., Yin, X., Struik, P. C., and Stam, P. (2005). Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. *Aust. J. Agric. Res.* 56, 883–894. doi: 10.1071/AR05153

Waddington, C. H. (1957). *The Strategy of the Genes*. London: Routledge.

Walsh, B., and Lynch, M. (2018). *Evolution and Selection of Quantitative Traits*. Oxford: OUP.

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. *bioArxiv* doi: 10.1101/2020.06.15.152017

White, M. R., Mikel, M. A., de Leon, N., and Kaeppler, S. M. (2020). Diversity and heterotic patterns in North American proprietary dent maize germplasm. *Crop Sci.* 60, 100–114. doi: 10.1002/csc2.20050

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

Wientjes, Y. C. J., Bijma, P., Calus, M. P. L., Zwaan, B. J., Vitezica, Z. G., and van den Heuvel, J. (2021). The long-term effects of genomic selection: Response to selection, additive genetic variance and genetic architecture. *bioRxiv [preprint]* doi: 10.1101/2021.03.16.435664 bioRxiv, 2021.03.16.435664,

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 Genes Genomes Genet.* 2, 1427–1436. doi: 10.1534/g3.112.003699

Wright, S. (1934). The method of path coefficients. *Ann. Math. Stat.* 5, 161–215.

Wright, S. (1963). *""""""""Discussion: Plant and Animal Improvement in the Presence of Multiple Selective Peaks" """"," in Statistical Genetics and Plant Breeding*. Washington, D.C: National Academies Press, 116–122. doi: 10.17226/20264

Yates, F., and Cochran, W. G. (1938). *The Analysis of Groups of Experiments*. Cambridge: Cambridge University Press, doi: 10.1017/S0021859600050978

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 Plant Science

ORIGINAL RESEARCH
published: 14 June 2021
doi: 10.3389/fpls.2021.676326

Genomic Selection in an Outcrossing Autotetraploid Fruit Crop: Lessons From Blueberry Breeding

*Luís Felipe V. Ferrão [1], Rodrigo R. Amadeu [1], Juliana Benevenuto [1], Ivone de Bem Oliveira [1,2] and Patricio R. Munoz [1]**

[1] Blueberry Breeding and Genomics Lab, Horticultural Sciences Department, University of Florida, Gainesville, FL, United States, [2] Hortifrut North America, Inc., Estero, FL, United States

Blueberry (*Vaccinium corymbosum* and hybrids) is a specialty crop with expanding production and consumption worldwide. The blueberry breeding program at the University of Florida (UF) has greatly contributed to expanding production areas by developing low-chilling cultivars better adapted to subtropical and Mediterranean climates of the globe. The breeding program has historically focused on recurrent phenotypic selection. As an autopolyploid, outcrossing, perennial, long juvenile phase crop, blueberry breeding cycles are costly and time consuming, which results in low genetic gains per unit of time. Motivated by applying molecular markers for a more accurate selection in the early stages of breeding, we performed pioneering genomic selection studies and optimization for its implementation in the blueberry breeding program. We have also addressed some complexities of sequence-based genotyping and model parametrization for an autopolyploid crop, providing empirical contributions that can be extended to other polyploid species. We herein revisited some of our previous genomic selection studies and showed for the first time its application in an independent validation set. In this paper, our contribution is three-fold: (i) summarize previous results on the relevance of model parametrizations, such as diploid or polyploid methods, and inclusion of dominance effects; (ii) assess the importance of sequence depth of coverage and genotype dosage calling steps; (iii) demonstrate the real impact of genomic selection on leveraging breeding decisions by using an independent validation set. Altogether, we propose a strategy for using genomic selection in blueberry, with the potential to be applied to other polyploid species of a similar background.

Keywords: genotyping by sequencing, sequencing depth, allele dosage, plant breeding, molecular marker, fruit quality, independent validation, genomic prediction

Edited by:
Fernando H. Toledo,
International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Gota Morota,
Virginia Tech, United States
Danilo Lyra,
Rothamsted Research, United Kingdom
Zhao-Bang Zeng,
North Carolina State University, United States

****Correspondence:***
Patricio R. Munoz
p.munoz@ufl.edu

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *05 March 2021*
Accepted: *12 May 2021*
Published: *14 June 2021*

Citation:
Ferrão LFV, Amadeu RR, Benevenuto J, de Bem Oliveira I and Munoz PR (2021) Genomic Selection in an Outcrossing Autotetraploid Fruit Crop: Lessons From Blueberry Breeding. Front. Plant Sci. 12:676326. doi: 10.3389/fpls.2021.676326

INTRODUCTION

Blueberry (*Vaccinium corymbosum* and hybrids) is recognized worldwide for its health benefits due to the high content and diversity of polyphenolic compounds (Kalt et al., 2020). Such health-related attributes have resulted in an increased demand for blueberries, as it has become a crop with one of the fastest growths in production trends, with an increase of 142% of its production in the last 10 years (FAOSTAT, 2021). In this sense, the blueberry breeding program at the University of Florida (UF) has had a major contribution to the expansion of production areas. Starting in the

1950s, the UF blueberry breeding program led to pioneering hybridizations between high-quality US northern adapted species (*Vaccinium corymbosu*m) and endemic US southern species (e.g., *Vaccinium darrowii*), selecting for low-chill requirements to break the dormancy of flower buds (Sharpe and Sherman, 1971; Lyrene, 2000). The resulting breeding material and cultivars, known as southern highbush blueberries, established a new industry in Florida and multiple warmer regions worldwide, allowing a year-round supply of fresh blueberries for the global market.

Historically, like many others, the UF program used recurrent phenotypic selection with visual assessment of plants to select both new parents for crossing and genotypes for commercial testing (Cellon et al., 2018). Despite the success of the industry and the release of many cultivars in recent decades, the use of conventional methods results in low genetic gains per unit of time. Moreover, the autopolyploid nature of the crop, long juvenile phase, multi-year evaluations, large experimental areas, and the high sensibility to inbreeding depression make phenotypic selection costly and time-consuming. Remarkably, it can take up to 12 years to release a new cultivar using conventional tools (Lyrene, 2005). As DNA sequencing costs continue to decrease, genomics-based markers present an opportunity to accelerate the breeding process by achieving more accurate selection during earlier breeding stages. Therefore, the UF blueberry breeding program has been leading innovative genomics studies and procedures to fill two primary gaps in the blueberry breeding literature: understanding the genetic architecture of complex traits via genome-wide association studies (GWAS) and quantitative trait loci (QTL) mapping; and, at the practical level, performing genomic prediction based on molecular markers, a methodology popularly referred to as genomic selection (GS).

GWAS and QTL mapping are both tools for providing a biological elucidation of the genetic architecture, in which molecular markers spanning the entire genome are statistically tested for associations with phenotypes (Pritchard et al., 2000). While QTL analyses are usually performed using structured populations, GWAS increases the mapping resolution by using populations with low levels of linkage disequilibrium considering a deep history of recombination events. In blueberry, we recently detected candidate genomic regions and markers associated with different fruit quality traits (Ferrão et al., 2018) and flavor-related volatiles (Ferrão et al., 2020) via GWAS investigations; and we built a high-density linkage map and detected QTL associated to berry firmness (Cappai et al., 2020a). In counterpart, GS aims to predict breeding values by using all genome-wide markers simultaneously (Meuwissen et al., 2001). The underlying rationale is that most QTL will be in linkage disequilibrium with some of the markers used whenever the marker density is high enough. Therefore, the estimated effect of all markers will lead to accurate predictions of the genetic merit for a complex trait. We have recently shown the potential of GS in blueberry breeding under distinct modeling scenarios (de Bem Oliveira et al., 2019, 2020; Amadeu et al., 2020a; Zingaretti et al., 2020).

The autopolyploid nature of blueberry ($2n = 4X = 48$) imposes additional challenges for analyzing and interpreting genetic data. Autopolyploids possess genomes with multiple sets of homologous chromosomes, resulting in non-preferential pairing and potential polysomic inheritance during meiosis. Given the presence of higher allele dosage (i.e., the number of copies of each allele at a particular locus), a higher number of genotypic classes are possible (Gallais, 2003; Garcia et al., 2013; Dufresne et al., 2014). Thus, the inclusion of allelic dosage information on GS models could imply a more accurate estimation of breeding values by considering the additive effect of multiple copies of the same allele and the potential inheritance of dominance effects. However, accurate allele dosage calling on polyploids depends on a higher depth of coverage, increasing genotyping costs when using sequence-based genotyping platforms (Gerard et al., 2018; Caruana et al., 2019). After performing foundational studies on the importance of polyploid models, the inclusion of non-additive effects, and sequencing depth on allele dosage parameterizations, the UF blueberry breeding program is now on track to overcome the barrier a simple promise to make GS a reality.

Motivated by the potential to use GS to reshape traditional blueberry breeding, we herein revisited some of our previous studies and described the current achievements in blueberry. Thus, our contributions in this paper are three-fold: (i) summarize previous results on the relevance of model parametrizations, such as diploid or polyploid methods, and inclusion of additive and non-additive gene actions for prediction; (ii) assess the importance of accurate dosage estimation for genomic prediction under low and high sequencing depth scenarios; (iii) demonstrate the realized impact of GS over breeding cycles by using an independent validation set. Altogether, we anticipate challenges and directions for future studies in blueberry that could be applied to other polyploid and fruit species of a similar breeding background.

MATERIALS AND METHODS

Populations and Phenotypic Data

The southern highbush blueberry populations used in this study were generated as part of the breeding program at the University of Florida. Two phenotypic datasets, referred to as *calibration set* and *testing set*, were used for different purposes.

The *calibration set* comprises a large breeding population already described in previous studies (Ferrão et al., 2018; de Bem Oliveira et al., 2019). Briefly, it consists of 1,837 individuals originating from 117 biparental crosses using 146 distinct parents. The population corresponds to early stages in the breeding scheme, and it was planted in a high-density nursery at the "Plant Science Research and Education Unit" in Citra, Florida. All phenotypic evaluations were conducted on ripe fruits collected from the beginning of April to mid-May. Fruit firmness ($g^{*}mm^{-1}$ of compression force), size (mm), and weight (g) were

Abbreviations: UF, University of Florida; GEBV, genomic estimated breeding value; GWAS, genome-wide association study; QTL, quantitative trait loci; eBLUE, empirical best linear unbiased estimate; SNP, single nucleotide polymorphism; GxE, genotype by environment interaction; MAS, marker-assisted selection; GS, genomic selection; GBLUP, genomic best linear unbiased prediction.

evaluated over two seasons (2014 and 2015), while soluble solid (°Brix) was evaluated only in 2015. Given the large representative population, all genomic prediction models reported in this study were calibrated using this dataset. The empirical best linear unbiased estimates (eBLUEs) were estimated for each genotype based on a linear model. Genotype and year were considered fixed effects, as described by Amadeu et al. (2020a). Hereafter, the eBLUEs for each trait were considered as our response variable in the genomic prediction analyses.

The *testing set* was used for independent validation in genomic prediction analyses. It comprises 280 advanced selections not originally included in the *calibration set*. These genotypes represent materials in advanced stages in the breeding program planted over 2013–2017 under commercial conditions. These genotypes were evaluated over several years (2014–2020), some of them (16 common genotypes) in different locations throughout Florida. As these phenotypes were collected from plants in different physiological phases and multiple environments, we adjusted the phenotypes using a linear model, including separate fixed effects for the year, location, and plant age. The eBLUEs of each genotype per trait were used as the phenotypic value in subsequent genomic prediction analyses. All phenotypic analyses were carried out using the ASReml-R software (Butler et al., 2009). Additional details about the *calibration* and *testing* datasets are reported in **Supplementary Figures 1, 2**.

Genotyping

The *calibration set* was genotyped using the "Capture-Seq" approach described in Benevenuto et al. (2019). The genotyping of the *testing set* was also performed using "Capture-Seq," considering 10,000 biotinylated probes of 120-mer at RAPiD Genomics (Gainesville, FL, USA). Sequencing was carried out in the Illumina HiSeq2000 platform using 150 cycle paired-end runs. To ensure that the same group of single nucleotide polymorphisms (SNPs) will be called in both *calibration* and *testing* sets, we included the next-generation sequence data from both sets under the same SNP calling pipeline. First, raw reads were cleaned and trimmed. Then, the remaining reads were aligned using Mosaik v.2.2.3 (Lee et al., 2014) against the largest scaffolds of each of the 12 homoeologous groups of *Vaccinium corymbosum* cv. "Draper" genome assembly (Colle et al., 2019). SNPs were called with FreeBayes v.1.3.2 using the 10,000 probe positions as targets (Garrison and Marth, 2012). Loci were filtered out applying the following criteria: minimum mapping quality of 10; only biallelic locus; maximum missing data of 50%; minor allele frequency of 1%; and minimum and maximum mean sequence depth of 3,750 across individuals, respectively. A total of 63,552 SNPs were kept after these filtering steps. Sequencing read counts per allele per individual were extracted from the variant call file using vcftools v.0.1.16 (Danecek et al., 2011) and subsequently used to investigate some practical questions implementation of genomic prediction in polyploids.

We first investigated the importance of accurate genotype calling for genomic prediction by testing *ratio* and *dosage* under high and low sequencing depth scenarios. For this purpose, we used the *calibration* set only in a 10-fold cross-validation scheme. For the *ratio* method, each genotypic score was computed as the ratio between the alternative and total read depth, as described by Sverrisdóttir et al. (2017) and applied in de Bem Oliveira et al. (2019). For the *dosage* method, genotypic classes were assigned probabilistically using the updog R package v.2.1.0 considering the "norm model" and prior bias equals zero (Gerard et al., 2018; Gerard and Ferrão, 2020). Both genotyping methods (*ratio* and *dosage*) were compared under scenarios of high sequencing depth (random sampling for the mean number of 60 reads – 60×) and low sequencing depth (random sampling for the mean number of 6 reads – 6×). Specifically, we assumed the sequencing reads of each allele (alternative or reference) for a given marker come from a multinomial distribution, with probability equal to the number of the reads divided by the total number of reads across all the alleles, markers, and individuals (N). Then, we sampled $N/10$ reads from this multinomial distribution. We performed this sampling 10 times, and each sampling result was used in a different cross-validation fold. To avoid an eventual confounding between the number of markers and the predictive ability over the four scenarios, we kept the same number of SNPs (63,552) across all scenarios. Therefore, in total, four scenarios were tested: *ratio_60x*, *ratio_6x*, *dosage_60x*, and *dosage_6x*.

For the real validation and implementation of GS in the blueberry breeding program, we used the actual read counts to estimate the allele dosage in the *calibration* and *test* sets according to the "norm model" in the updog 2.1.0 R package (Gerard et al., 2018; Gerard and Ferrão, 2020). The posterior probability modes were used as our genotypic score. After estimating the posterior mean per genotype, we filtered out markers with a proportion of individuals genotyped incorrectly ("prop_miss" < 10%) and markers with an estimated bias higher than 0.13 and smaller than 7.38. Missing genotypes were imputed by the mean of each locus. A total of 48,829 SNPs were kept and used in genomic prediction for independent validations.

Statistical Analyses

Single-trait linear mixed models were used to predict breeding values using the best linear unbiased prediction (BLUP) and restricted maximum likelihood approach (REML) to estimate variance components, as following: $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Zu} + \boldsymbol{e}$; where $\mathbf{y}$ is a vector of pre-corrected phenotypic records for a particular trait; $\boldsymbol{\mu}$ is the overall mean; $\mathbf{Z}$ is an incidence matrix linking observations in the vector $\mathbf{y}$ to their respective breeding value in the vector $\mathbf{u}$. Normality was assumed for the additive and residual effects, where $\boldsymbol{u} \sim MVN(\mathbf{0}, \boldsymbol{G}\sigma_u^2)$ and the residual variance $\boldsymbol{e} \sim MVN(\mathbf{0}, \boldsymbol{I}\sigma_u^2)$. For the residual, $\mathbf{I}$ is an identity matrix; while σ_u^2 and σ_e^2 are the genetic and residual variance components. The matrix $\mathbf{G}$ denotes the genomic relationship matrix computed using the ratio genotypic score or the tetraploid allele dosages with the different sequencing depths described above. The matrices were estimated in the AGHmatrix v.2.0.0 R package (Amadeu et al., 2016). For the *ratio* implementation, we used the "ratio" option in the software that computes the relationship as $\boldsymbol{G} = \boldsymbol{ZZ}'/h$, where $\mathbf{Z}$ is the mean-centered matrix of the molecular marker information (ratio values); and h is a scale factor, where $h = \sum_{i=0}^{m} s_i^2$ and s_i^2 is the variance of the vector z_i centered marker i (for more details, see de Bem Oliveira et al., 2019). For the *dosage*

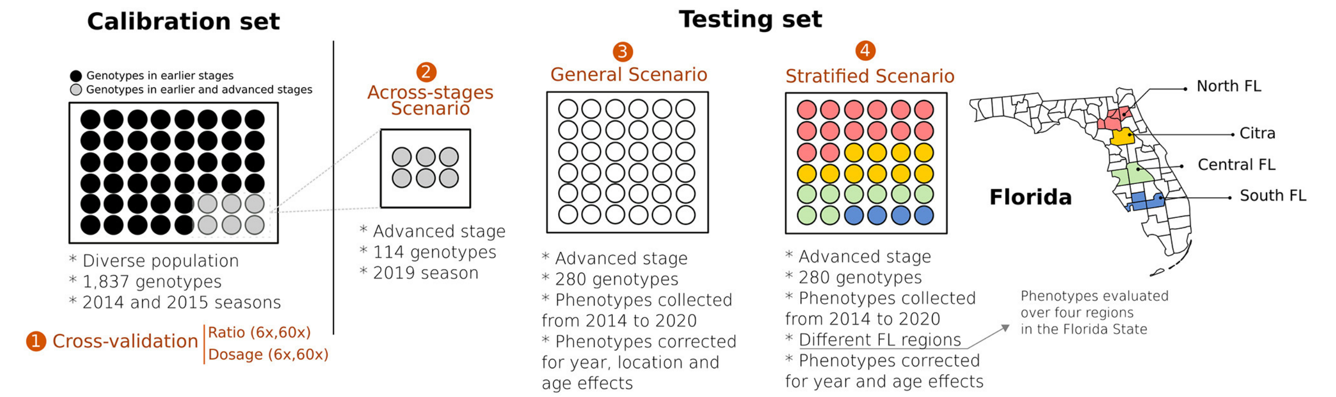


FIGURE 1 | Schematic representation of four validation scenarios tested in blueberries. Calibration set represents a diverse group of genotypes representative of the UF blueberry breeding population. In Scenario 1, we used the calibration set in a 10-fold cross-validation scheme to test the relevance of genotyping calling (ratio vs. dosage) considering two different sequencing depths (6× and 60×). Scenario 2 (across-stages) represents a group of 114 individuals originally presented in the calibration set that were clonally propagated, moved to the advanced Stage of the breeding program, and phenotyped under commercial field conditions. Scenario 3 (general prediction) represents an independent group of 280 genotypes (testing set), evaluated under commercial conditions. The phenotypic values of the target individuals were pre-adjusted for the year, location, and age effects. Finally, in Scenario 4 (stratified prediction), we performed predictions over four regions of the State of Florida. To avoid potential model overfitting, we removed genotypes from the calibration set overlapped with the testing set.

implementation, we used the additive relationship matrix based on VanRaden (2008) as described by de Bem Oliveira et al. (2019). All genomic prediction analyses were carried out using the rrBLUP package (Endelman, 2011). For comparison, predictions were also carried out using pedigree BLUP. Using the same linear mixed model, we computed the numerator pedigree-based relationship considering autotetraploidy and no double reduction (Kerr et al., 2012), using the AGHmatrix v.2.0.0 R package (Amadeu et al., 2016).

Predictive performances were assessed for the *ratio* and *dosage* methods under high (60×) and low (6×) sequencing depth scenarios using only the *calibration set* in a 10-fold cross-validation scheme. To this end, the *calibration set* was randomly divided into 10 groups, where one group was used as a validation test, while the remaining nine groups were used as training. Models were trained in the validation test using the genomic best linear unbiased prediction (GBLUP) approach. For each fold, predictive abilities were estimated using Pearson's correlation between genomic estimated breeding values (GEBVs) and the corresponding eBLUEs. We also evaluated the correspondence between the top 20 groups of individuals ranked using *dosage_60x* and the other scenarios. A *post-hoc* Tukey test (alpha = 0.05) was used for intergroup comparisons between the top 20 ranked genotypes.

For the independent GS validation over the breeding cycles, we assessed the robustness of our predictive model over different scenarios: (i) *across-stages* scenario refers to 114 individuals from the *calibration* set that were clonally propagated in 2014 and planted in a commercial condition in a single location, becoming the *testing* set – prediction accuracy in this scenario can demonstrate the potential losses when models are trained at earlier stages (high density) and used at late stages of selection (commercial condition); (ii) *general* scenario stands for models trained in the *calibration* set and predictions carried out in the *testing* data, in which the target phenotypic values were pre-corrected for year, location, and age fixed effects; (iii) *stratified* scenario comprises models trained in the *calibration set* that were tested for predictions across four regions in Florida (North-FL, Central-FL, South-FL, and Citra-FL) – in contrast to the *general* predictions, in this scenario the target phenotypic values were pre-corrected only for the year effect per region. In all scenarios, predictive performances were assessed via Pearson's correlation.

A summary of all validation scenarios is illustrated in **Figure 1**. We complemented the predictive analysis for the stratified predictions by accessing the importance of genotype-by-environment interaction (GxE) via ANOVA. To this end, we considered 16 genotypes (checks) that were phenotyped over the four regions. We fitted a linear model considering the year, genotype, location, and the interaction between genotype and location (GxE) as fixed effects. ANOVA was performed in R (R Team, 2013) using the native *lm()* function.

RESULTS AND DISCUSSIONS

In the last two decades, GS has become a reality for many animal and plant breeding programs. Despite the optimism and proven efficacy, its wide implementation is still hindered by investment costs and the analytical skills required (Hickey et al., 2017). With that in mind, the UF blueberry breeding program initiated genomic studies on a large scale in 2013. First, we worked closely with genotyping companies to design customized genotyping platforms; we phenotyped and genotyped a large and multi-parental blueberry breeding population; we increased our computational resources; and finally, we adapted our breeding framework to incorporate genomics. During this

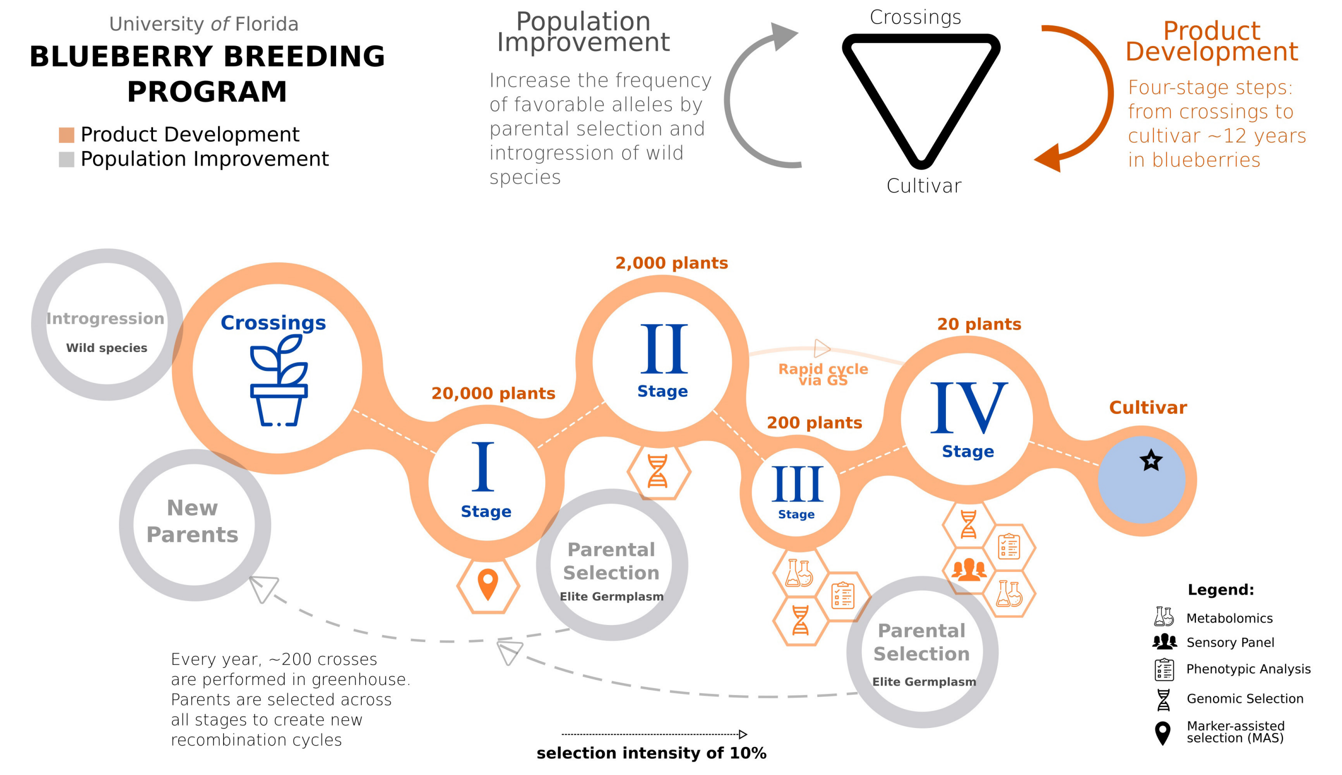


FIGURE 2 | A schematic representation of the UF blueberry breeding program, integrating phenotypic selection and genomic prediction. The breeding process is conventionally organized in two integrated steps: population improvement and product development. A breeding cycle starts with crosses between outstanding parental genotypes. After that, several stages (I–IV) are required to evaluate the genotype performance. At Stage I, we will use marker-assisted selection targeting traits with simple genetic architecture. Genomic selection will be implemented in Stage II when GEBVs are computed. In advanced selections (Stages III and IV), high-quality phenotyping will be performed to leverage the calibration of genomic prediction models. At these stages, metabolomics and sensory panel analyses will also play an important role in flavor-assisted selections. In the end, elite materials are registered as clonally propagated cultivars. In addition, to shortening the time for product development, GS can be applied to move top-ranked plants directly from Stages II to IV, skipping at least 3 years of evaluation at Stage III. For population improvement, GS can assist in more accurate parental selection at early stages.

process, the implementation of GS in a polyploid and outcrossing species proved challenging, particularly regarding the intrinsic biological complexities and the availability of genomic and computational tools (Mackay et al., 2019). In blueberry, for example, a high-quality genome assembly became available only in 2019 (Colle et al., 2019). As a result, about half of the capture-seq genotyping probes originally developed based on a draft genome assembly were discarded afterward based on the high-quality genome, without compromising genetic association and genomic prediction analyses (Benevenuto et al., 2019). We also explored additional optimizations to reduce costs regarding the number of individuals per family, the number of markers, and sequencing depth (de Bem Oliveira et al., 2020). Moreover, new genomics methods and tools have been developed in the last decade for the polyploid community, including allele dosage estimation, haplotype reconstruction, and the use of different relationship matrices (Bourke et al., 2018). Here, we presented the lessons we have learned so far for implementing GS in an autotetraploid and outcrossing species. We summarized previous results and also included novel findings relevant to the blueberry and polyploid community.

Filling the Gaps: Phenotypic and Genotypic Selection in the Same Breeding Framework

Blueberry is an outcrossing and clonally propagated crop, for which the breeding process can be conventionally organized in two central steps: population improvement and product development (Lyrene, 2005). First, population improvement is done to manage the frequency of beneficial alleles over time by selecting and crossing outstanding materials, as conceptualized in recurrent selection designs. Second, in parallel, product development consists of a series of trials in which potential candidates are evaluated over several years and locations, advancing across stages until selecting the best genotypes becomes a registered variety. In **Figure 2**, we illustrated these two key steps and how they are integrated into a four-stage selection design (from Stages I to IV) in the UF blueberry breeding program.

Annually, the blueberry breeding program performs more than 200 crosses, including parents selected among cultivars, elite material, and wild germplasm (Lyrene, 2005). From these crosses, about 20,000 seedlings are planted in non-replicated high-density nurseries (area of 0.2 ha), establishing the so-called

Stage I. After 1 year, plants in Stage I are visually selected based on fruit size, color, scar, and using the breeder's "bite test" for flavor quality attributes. Approximately 10% of the original number of seedlings are kept after this first selection, and the unselected plants are removed from the field. To not exhaust genetic diversity, a minimal number of individuals per family are kept. However, given blueberry's long juvenile period, the availability of few berries, and the high competition in a high-density planting, it is difficult to phenotype for all traits and assess the individuals' full potential stage.

Additionally, the large number of individuals prevents genomic prediction at this stage, given the costs of genotyping. Therefore, at Stage I, we envision that marker-assisted selection (MAS) for traits with simple genetic architecture is a more feasible approach, and it is a current research line of the breeding program. In this regard, the example of MAS implementation in early selection stages is reported in strawberry (Gezan et al., 2017; Osorio et al., 2020).

After the first selection, ∼2,000 genotypes pass to the second stage (Stage II). All plants stay in the same field plot, in high density. Further visual phenotypic evaluations are performed for the next 3 years. At this stage, we are implementing genomic prediction to increase genetic gains by improving phenotyping accuracy and selecting parents at early stages. Therefore, at Stage II, all plants will be genotyped. The GEBVs will be predicted for five fruit quality traits (soluble solids, titratable acidity, weight, size, and firmness), yield, and consumers panel liking scores. Using a selection index according to trait importance (Williams, 1962), we will perform GS to complement standard phenotypic descriptors and rank all genotypes. Different selection indexes are defined every year, depending on the traits and crosses performed, with yield and flavor traits usually receiving the highest weights. As routinely done, 10% of the 2,000 plants will be moved to the next stage (Stage III), where selected plants are clonally propagated and evaluated in a 15-plant clonal plot in a commercial field.

At Stage III, around 200 plants are more accurately phenotyped for more traits, using more fruits, clonal repetitions, and multiple years of evaluations in commercial conditions. Technically, all information collected at this stage will be used to feed the genomic prediction models. The UF blueberry breeding program has included new traits for routine phenotyping to meet the current demand from different marketable demands in recent years. For example, the use of volatiles for flavor-assisted selection has shown the ability to predict sensory perceptions by explaining 55% of the variation in overall liking scores (Colantonio et al., 2020). Given the high costs to perform sensory panels, we are incorporating metabolomics in the breeding pipeline to predict flavor ratings for many genotypes at Stage III (Gilbert et al., 2015; Colantonio et al., 2020; Ferrão et al., 2020).

In the last stage (Stage IV), around 15–20 plants selected from Stage IIIs with consistent and outstanding performances are propagated and planted at commercial trial sites across Florida. The different locations comprise two production systems according to the accumulation of chilling hours: evergreen and deciduous (Fang et al., 2020). To ensure accurate selection, phenotypic data is collected weekly and used to feed our genomic prediction models. Fruits from selected genotypes are also submitted to sensory panels, where blueberry consumers score flavor preferences. Elite selections from this final Stage are ultimately named, patented, and released as clonally propagated cultivars.

Altogether, the conventional breeding pipeline takes up to 12 years to evaluate the genotype merit of an individual to be released as a cultivar. With the implementation of genomic selection at the scope of the breeding program, the selection criteria can be more accurate than the visual phenotypic selection at Stage II. Moreover, it will shorten the time to select genotypes to become a parent in the next breeding cycle and advance to Stage III. In a typical recurrent selection breeding scheme, the parental selection is crucial (Lyrene, 2005). We have optimized this selection by ranking the GEBVs over the breeding cycles and seeking crosses that minimize inbreeding. Among the different tools available for mate allocation, we have recently implemented the algorithm described in the AlphaMate software with default parameters (Gorjanc and Hickey, 2018).

"*Simplicity Is the Ultimate Sophistication*"[1]: On the Relevance of Additive GBLUP Models

When confronting the problem of modeling the relationship between molecular markers and variation in the observed traits, an important question to keep in mind is what statistical method could better describe this relationship (Ferrão et al., 2019). In recent years, we have investigated statistical and biological aspects underlying the implementation of genomic prediction in autopolyploid species, including (i) the importance of accounting for allele dosage in whole-genome statistical models (de Bem Oliveira et al., 2019); (ii) the relevance of multiple gene actions, including additive and non-additive genetic sources (Amadeu et al., 2020a; Zingaretti et al., 2020); and finally, (iii) the impact of sequencing depth of coverage, when sequence-based genotyping approaches are used (de Bem Oliveira et al., 2020).

Among the factors that differentiate diploid and polyploid analyses, resolving the allelic dosage of individual loci is one the most important. While in diploid organisms, only three genotypic classes are possible for biallelic markers, autotetraploids, like blueberry, can have up to five genotypic classes. Therefore, in theory, it is expected that statistical models accounting for the dosage effect could be more informative and provide a more realistic representation of the genetic complexity of a quantitative trait (Garcia et al., 2013). We first tested this hypothesis by contrasting polyploid and diploid parametrizations in GWAS studies (Ferrão et al., 2018), whereby a larger number of associations were observed under polyploid models. In a subsequent study, we investigated a similar assumption for genomic prediction (de Bem Oliveira et al., 2019). We tested GBLUP models using relationship matrices built in a tetraploid (Slater et al., 2016) and diploid (VanRaden, 2008) fashion.

Interestingly, both parametrizations resulted in similar performances for all traits tested. Furthermore, the similar

[1] Quote by Leonardo da Vinci.

predictive ability for diploid and polyploid parametrizations was also reported in other autotetraploid species (Lara et al., 2019; Matias et al., 2019), which ultimately reinforced the robustness of the predictive accuracy of GBLUP regardless of the ploidy parametrization used. These results are explained by the similarity between the genomic relationship matrices computed using diploid and autotetraploid parametrizations. Recently, we presented empirical evidence on this topic by showing that the estimation of molecular pairwise relatedness in both scenarios are highly correlated, in particular, under low-to-middle rates of heterozygosity (Amadeu et al., 2020b).

Besides the potential additive impact of allele dosages, dominance effects can also be heritable in polyploids and could improve the prediction of genetic values. Therefore, it is also reasonable to speculate that a greater number of alleles per locus may increase the range of genetic models to describe one-locus genotypic value by accounting for multiple dominance levels (Gallais, 2003). This is exemplified by the different models addressing the dominance effect proposed in the polyploid literature, including the use of digenic interactions (Endelman et al., 2018), the use of a general effect by assuming that each genotype has its effect (Rosyara et al., 2016; Slater et al., 2016), and the use of heterozygous parametrization (Enciso-Rodriguez et al., 2018). In blueberries, we tested the importance of such different gene actions in predictive studies. Although we have observed an improvement in the statistical goodness of fit when dominance effects are counted, this increment is not directly translated into predictive ability (Amadeu et al., 2020a). Hence, the additive model resulted in performance similar to models accounting for dominance effects, as it has been described for diploid species (Muñoz et al., 2014).

Given the genetic complexity of polyploids and the potentially higher intra- and inter-locus interactions, we also hypothesized that predictions could be improved by using deep learning techniques (Zingaretti et al., 2020). Through deep learning, we could take advantage of non-linearity assumptions to model the whole genetic merit of an individual. We used allo-octoploid strawberry and autotetraploid blueberry as our biological models and compared linear models and deep learning techniques for prediction to test this. We did not observe improvements of deep learning over traditional linear models for traits with presumably different genetic architectures in both species. The only exception was observed in a simulated data set. Deep learning performed better for traits with large epistatic effects and low narrow-sense heritability, which reinforced the high predictive ability of mixed models as prediction machinery.

Our last contribution to the practical implementation of genomic prediction in polyploids is the relevance of sequencing depth of coverage for genotyping methods based on next-generation sequencing. Sequencing depth refers to the number of reads sequenced at a given site in the genome. Low coverage datasets increase the chances of not sampling all homologous chromosomes at a given site for a given individual during sequencing. Thus, it could result in high rates of missing data, miscalled genotypes, and uncertainty of allele copy number in heterozygous genotypes (Clark et al., 2019). Some studies in polyploid crops have recommended increasing the sequencing depth to circumvent this issue, which implies higher costs of genotyping. For example, Bastien et al. (2018) and Uitdewilligen et al. (2013) suggested sequencing depths of 50X−80X for an accurate assessment of allele dosage in autotetraploid potatoes. In a recent study, we demonstrated that such numbers are quite conservative for genomic prediction. By combining a simple genetic parametrization (*ratio*) and low-to-mid sequencing depth (*6x–12x*), we achieved similar predictive accuracies as the ones obtained using higher depths for blueberry traits with different genetic architectures (de Bem Oliveira et al., 2020). In practical terms, reducing the amount of sequencing data will also reduce the costs of implementing GS or potentially genotyping more individuals under a fixed budget.

Despite the considerable advancements previously explored, the relevance of using more sophisticated algorithms for genotype calling and its impact on genomic prediction remains unexplored. Recently, several new methods have been developed to assign accurate allelic dosage of individual loci in polyploids (Garcia et al., 2013; Gerard et al., 2018; Pereira et al., 2018; Clark et al., 2019). In this paper, we compared predictive abilities. We confirmed that low-to-mid sequencing depth and ratio parametrization could be used to rank GEBVs with similar predictive performance (**Figure 3** and **Supplementary Table 2**) and genotypic ranking (**Table 1**). Nonetheless, despite the attractive simplicity of using the ratio and low-sequencing depth, such results are only valid for prediction analysis (de Bem Oliveira et al., 2019, 2020). Importantly, there is no empirical evidence that setting the parameters to these levels could work for inferential studies such as GWAS, population genomics, linkage, and QTL mapping. In this sense, an important counterpoint was recently reported in hexaploid sweet potato. Higher sequencing depths and accurate dosage calling improved the ultra-dense linkage map and posterior QTL analysis (Gemenet et al., 2020; Mollinari et al., 2020). For GWAS, we observed large rates of false-positive associations when analyses were performed using low sequencing depth associated with the ratio parametrization (results not shown). Herein, we systematically observed large biases when relationship matrices were constructed using the *ratio_6x* approach (**Supplementary Figure 4** and **Supplementary Table 4**).

Our results suggest that the use of traditional GBLUP is robust enough for genomic prediction, even under simplistic assumptions. This fact has long been discussed in the specialized literature and has raised questions on the contribution of linkage disequilibrium between QTL and markers vs. the relationship information to GS (Habier et al., 2013).

How Does Genomic Prediction Work in a Real Validation Population?

While we have investigated several statistical and computational aspects related to GS in blueberry, it is still unknown how accurate the predictions will be across breeding cycles, with plants in different phenological stages and locations. This scenario came to be called "true validation" and involves the use of independent populations. We investigate it by dividing our prediction analyses as following: models calibrated in 2014 and 2015 using plants in Stage II were used for genomic predictions

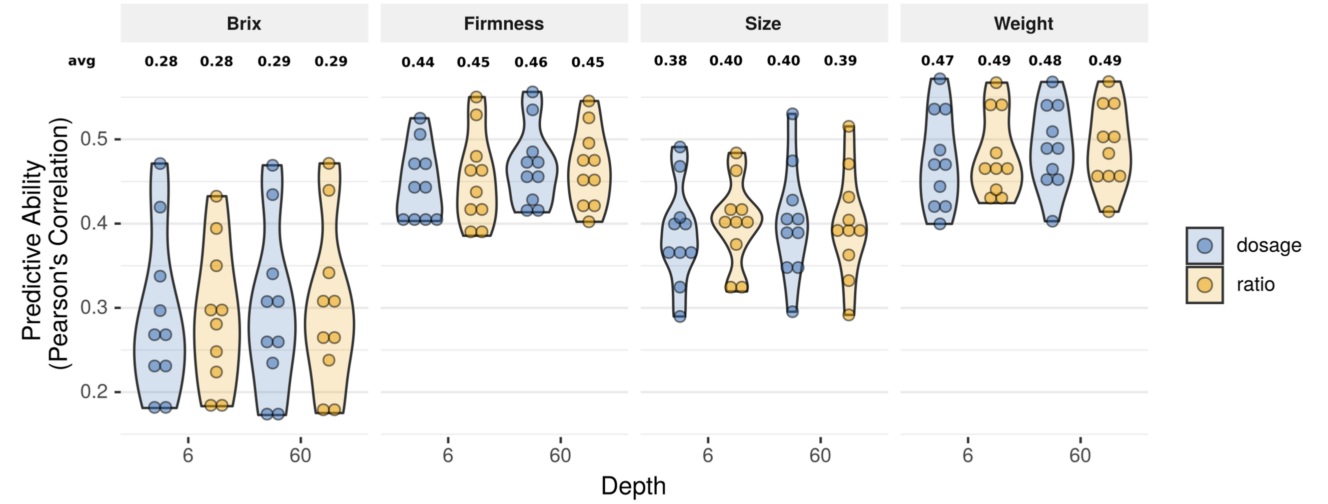


FIGURE 3 | Violin plot with predictive ability considering two genotype calling approaches (dosage and ratio) under two sequencing depth scenarios (6× and 60×) for four fruit quality traits in blueberry using 10-fold cross validation. Each circle represents one cross validation fold result.

TABLE 1 | The number of genotypes matching the top 20 rankings using the dosage_60× method as the benchmark, under 10-fold cross-validation.

Method	Depth	Firmness	Size	Weight	Brix
Dosage	6×	16.5^b	16.9^b	16.3^b	16.4^b
Ratio	6×	16.2^b	15.6^c	16.2^b	15.3^b
Ratio	60×	18.8^a	18.3^a	18.6^a	18.7^a

A post-hoc Tukey test (alpha = 0.05) was used for intergroup comparisons over the scenarios. Cells with the same letter represent non-statistically different groups for the given trait (column).

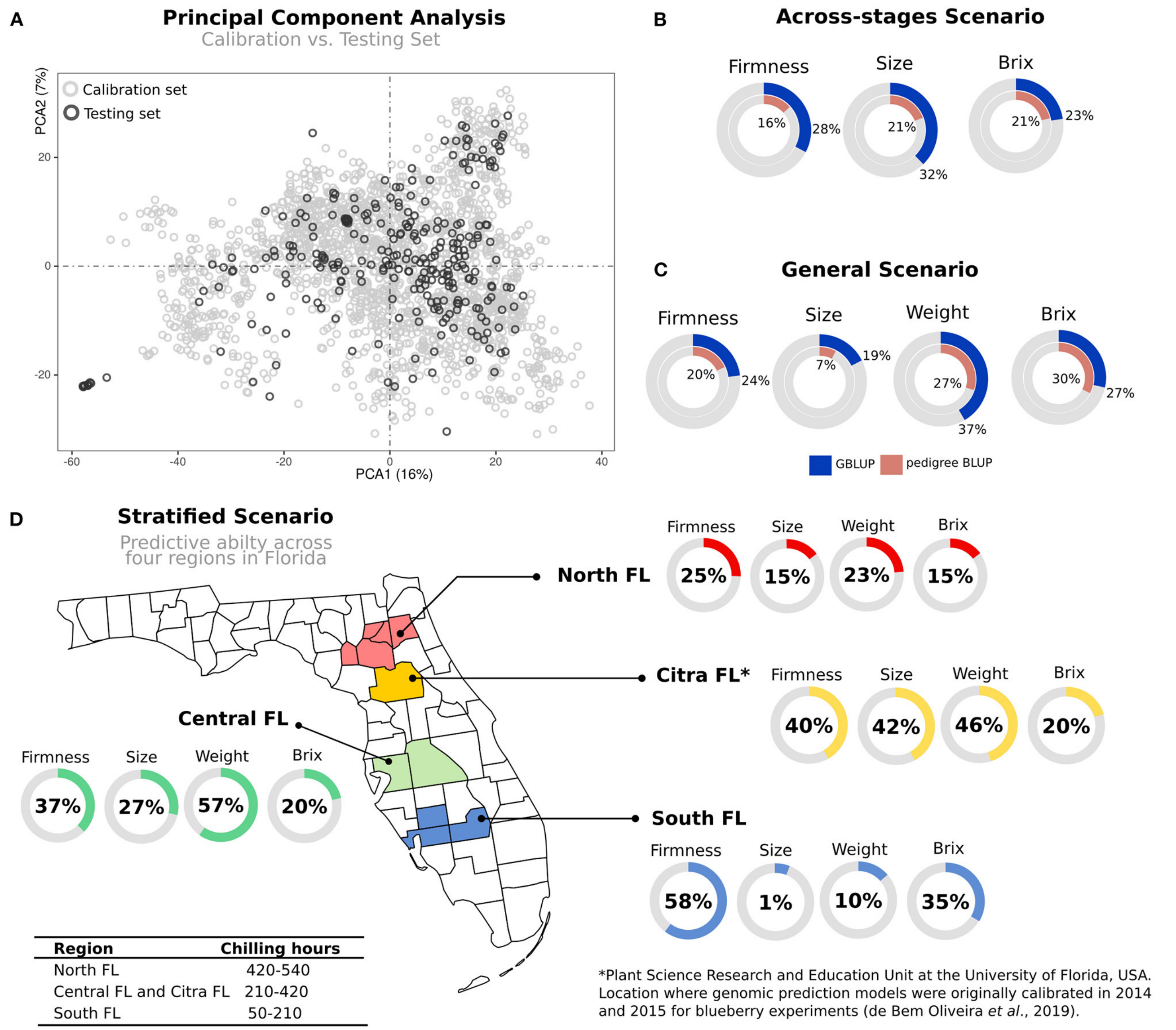
of individuals at Stages III and IV. Both data sets share genetic similarity (**Figure 4A**).

For independent validations, we tested different scenarios in which GS could be applied (**Figure 1**). First, we focused on validations across breeding stages. To this end, we used the *calibration* test—originally evaluated in Stages II—to predict a subset of individuals that were cloned and planted in an advanced stage (Stages III). When compared to within-sample cross-validation schemes, as originally reported by de Bem Oliveira et al. (2019) and Amadeu et al. (2020a), lower predictive accuracies were observed (**Figure 4B**). These results mainly highlight (i) the importance of collecting better phenotypic data and (ii) the influence of plant management. Remarkably, most of the phenotypic traits measured in the *calibration* set were collected from five berries per genotype, while on Stage III, we used 25 berries per genotype. Furthermore, genotypes in Stage II are planted in high-density nurseries with phenotypes collected in plants that are still in their juvenile phase. At the same time, Stages III are grown under commercial conditions and evaluated over several years.

A second predictive scenario tested the relevance of *calibration* tests at early stages to predict independent genotypes in advanced stages that were more extensively phenotyped. The results for most fruit-quality traits confirm the importance of genomic information (*general* predictions) over pedigree-based methods (**Figure 4C**). However, compared with predictions using within-sample cross-validation schemes, we also observed a reduction in the predictive results (**Supplementary Table 3**) (de Bem Oliveira et al., 2019; Amadeu et al., 2020a). This decline in predictive performance in true validation is expected due to differences in the allele frequencies over populations, variation in linkage disequilibrium patterns, and GxE interactions (Habier et al., 2013).

In the third scenario, a more challenging exercise was to measure how predictive ability varies across regions in the State of Florida (*stratified* predictions, **Figure 4D**). Higher predictability was observed for Citra and Central-FL, the closest regions where the models were originally trained. In counterpart, plants evaluated in the South-FL showed, on average, lower predictability performances. Despite the small number of genotypes included in this analysis, these results provide insights into the importance of GxE interaction for GS in blueberry. We further explored this hypothesis by using a group of 16 common genotypes (checks) evaluated over the four regions. The results confirmed the significance of the GxE effect for most of the traits (**Table 2** and **Supplementary Figure 3**), with the plants evaluated in South-FL showing the most contrasting values. It is noteworthy that blueberry locations in South-FL are grown under an evergreen production system, under less chilling hours, and are focused on preventing defoliation during the winter months (Fang et al., 2020). On the other hand, Citra, Central-FL, and North-FL regions are grown under the deciduous production system, where leaves are dropped during the winter. Such differences in the production systems could drive the largest disparity observed at South-FL compared with the other regions.

The results from independent validations allow us to draw some practical conclusions. First, even with low-to-moderate



Region	Chilling hours
North FL	420-540
Central FL and Citra FL	210-420
South FL	50-210

*Plant Science Research and Education Unit at the University of Florida, USA. Location where genomic prediction models were originally calibrated in 2014 and 2015 for blueberry experiments (de Bem Oliveira *et al.*, 2019).

FIGURE 4 | Genomic prediction on independent validation scenarios. **(A)** Principal component analysis (PCA) of two blueberry populations: *calibration* set represents the trained set, where genomic prediction models were originally calibrated; and *testing* set comprising additional 280 individuals used for testing. **(B)** Across-stages prediction: predictive ability was measured in a subset of 114 individuals included in the calibration set, but phenotypes were collected in advanced selection stages. **(C)** General prediction: the predictive ability of the testing set after training the models in the calibration set. **(D)** Stratified predictions: after training the models in the calibration set, individuals in the testing set were predicted using phenotypes collected over four macro-regions in the Florida State, which are under different chilling hour accumulation. All predictive abilities were expressed as percentage values.

predictive accuracies, GS is still encouraging. For example, soluble solids and firmness are both traits treasured by consumers, for which routine phenotyping is expensive and time-consuming for large populations, like Stage IIs. Ranking plants based on their GEBVs proved to be a better alternative than any other criteria historically used throughout UF blueberry breeding program (pedigree or visual selection). More accurate phenotypic data to annually recalibrate the model also has the potential to improve predictability.

Unifying Biological Discoveries and Predictions

Genomic information can also provide new opportunities to integrate biotechnology and quantitative genetics into modern breeding programs, creating platforms for both deliveries of new products and biological discovery (Hickey et al., 2017). For example, in blueberry, biological discoveries have been addressed via QTL mapping (Cappai et al., 2020a) and GWAS studies (Ferrão et al., 2018, 2020) for multiple fruit quality traits.

TABLE 2 | Mean and standard deviation (in parenthesis) of four fruit quality traits were evaluated in advanced stages of the blueberry breeding program at four Florida regions.

Location	Firmness (g * mm^{-1})	Size (mm)	Weight (g)	° Brix
North FL	248 (32.8)	18.0 (1.79)	2.57 (0.641)	11.3 (1.31)
Citra	245 (42.0)	17.0 (2.22)	2.34 (0.691)	10.9 (1.33)
Central FL	244 (29.3)	17.7 (1.46)	2.29 (0.491)	11.8 (1.27)
South FL	251 (33.4)	17.4 (1.34)	2.21 (0.549)	12.0 (1.91)
GxE (*p*-value)*	0.007	0.0002	0.005	0.47

**p-values associated to genotype-by-environment interaction (GxE) were computed using a linear model and ANOVA, where season, genotype, location, and the interaction between genotype and location (GxE) were fitted as fixed effects.*
Values were computed using 16 common genotypes (checks).

Unifying such discoveries with prediction is challenging, but it has been addressed under three different avenues: (i) use of GWAS discovered QTL as fixed effects on GS models; (ii) incorporating markers (or QTL) in MAS designs, and (iii) using genome-editing technology to speed up breeding.

In a strategy called "GS *de novo* GWAS," we explored the importance and applicability of GWAS findings for prediction using the significant GWAS hits as fixed effects in GS models, considering independent datasets. For oligogenic traits, like some flavor-related volatiles, we achieved an increase of more than 20% in predictive ability compared with traditional GS methods (Ferrão et al., 2020). Using a similar strategy, gains in predictive performance have also been reported in other crops, such as maize (Bernardo, 2014; Rice and Lipka, 2019), wheat (Sehgal et al., 2020), and rice (Spindel et al., 2016). Alternatively, we have investigated further modeling strategies to accommodate biological information into the predictive models. For example, the use of Bayesian strategies that could accommodate SNPs with larger effect by using different prior distributions (Erbe et al., 2012; Gianola, 2013; Zhou et al., 2013); and GBLUP models that could weight variants previously selected either via association analysis or using bioinformatic pipelines (Su et al., 2014; Zhang et al., 2016; Liu et al., 2020; Ren et al., 2021).

Another potential strategy is to use target markers associated with important traits for MAS during Stage I of the blueberry breeding program. Such a strategy could be used for the early selection of plants still in the seedling stage. Acknowledged by their simple genetic architecture, we showed that few markers could yield reasonable predictive accuracies of volatile emission and, thus, leverage flavor selection (Ferrão et al., 2020). We envision that MAS can also be implemented for other oligogenic traits. In this regard, we have been conducting other GWAS and QTL mapping studies for disease resistance, such as anthracnose (*Colletotrichum gloeosporioides*) and bacterial wilt (*Ralstonia solanacearum*). A similar strategy has been implemented in strawberries (Gezan et al., 2017; Osorio et al., 2020) and other fruits (Iezzoni et al., 2020). However, for MAS to be applicable for thousands of plants, cheap and fast DNA extraction and targeted SNP genotyping assays should be optimized. We are currently testing high-resolution melting (HRM) and competitive allele-specific PCR (KASP) assays to validate and implement MAS for volatiles.

Gene editing is another attractive technology with the potential to have significant effects on the breeding program. Aside from the use of CRISPR-Cas9 for validating candidate genes identified via GWAS or QTL studies, some simulations have recently shown that genome editing can double the rate of genomic gain when coupled with genomic prediction, compared with GS conducted in isolation (Noman et al., 2016; Hickey et al., 2017). However, to our knowledge, there is only one study of CRISPR-Cas9 targeted mutagenesis in blueberry (Omori et al., 2021). At the UF blueberry breeding program, we have advanced our understanding of the best tissue culture practices and most effective transformation markers (Cappai et al., 2020b), laying the ground for CRISPR/Cas9 genome editing implementation in our breeding program. Using this technique, we can also take advantage of the knowledge accumulated from model crops to introduce novel allelic diversity in orthologs and accelerate the domestication process.

CONCLUSIONS

The implementation of GS has already changed the UF blueberry breeding program routine by reorganizing how we collect genotypic and phenotypic information and analyze data to rank the material to advance stages and breed in the next cycles. Our previous studies on GS were fundamental to define the most cost- and time-effective methods for model parameterization and genotyping. The main lessons learned can be conveniently divided into different areas. Statistically, despite the numerous algorithms for prediction—many of them more elegant at the biological and computational level—the use of additive effects under a linear mixed model framework (GBLUP) showed the best balance between efficiency and accuracy. Considering the particularities of autopolyploid genetic data, we showed that for GS, low depth of sequencing (6×-12×) simplifies the allele dosage information (i.e., diploidization and ratio) resulted in similar prediction accuracies as those obtained using more refined scenarios. Finally, the genomic prediction was incorporated in a recurrent selection breeding scheme at the practical level, whereby variety development and populational improvement run in parallel. So far, GEBVs have been primarily used for parental selection to increase genetic gains while keeping the genetic diversity. A more objective reduction in the number of years to develop a cultivar would be selecting the top-ranked genotypes from Stage II directly to IV, skipping at least 3 years of evaluations at Stage III.

FUTURE DIRECTIONS

Finally, we highlight some challenges and opportunities for further studies in blueberries. First, recalibrating the model with more accurate phenotypic data can yield better predictive ability. In this sense, phenomics is also a cutting-edge area of research that could leverage the number of traits and samples collected during a season and improve the quality of phenotypic data. For example, yield is a complex and time-consuming trait to be phenotyped over the season. We envision that image-based phenotyping may aid in evaluating yield and other traits, such as plant architecture and diseases. For the future, it would also be important to incorporate additional statistical checks (common genotypes) across years and locations to understand better the effects of GxE interaction on genomic predictions and recalibrate our models according to the environmental targets. On integrating multi-omics data, we expect that we will predict flavor preferences through volatile quantification and perform an early selection for more flavorful cultivars. Statistically, testing new algorithms for mate allocation and using haplotypes for prediction and imputation methods are some potential areas that could further improve genomic predictions.

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.

AUTHOR CONTRIBUTIONS

PM and LF conceived and supervised the study. JB coordinated the collection and genotyping of the samples. IB coordinated the data collection for real validation. LF and RA analyzed and interpreted the phenotypic and genomic selection results. LF wrote the paper and included the revision from all authors. All authors read and approved the final version of the manuscript for publication.

FUNDING

This work was supported by the UF royalty fund generated by the licensing of blueberry cultivars.

ACKNOWLEDGMENTS

The authors thank Catherine Cellon and James Olmstead for the calibration set data collection; Werner Collante, Lauren Scott, and Douglas Phillips for the testing set data collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.676326/full#supplementary-material

REFERENCES

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A. F., Resende, M. F. R., and Muñoz, P. R. (2016). AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. *Plant Genome* 9, 1–10. doi: 10.3835/plantgenome2016.01.0009

Amadeu, R. R., Ferrão, L. F. V., Oliveira, I., Benevenuto, J., Endelman, J. B., Munoz, P. R., et al. (2020a). Impact of dominance effects on autotetraploid genomic prediction. *Crop Sci.* 60, 656–665. doi: 10.1002/csc2.20075

Amadeu, R. R., Lara, L. A. C., Munoz, P., and Garcia, A. A. F. (2020b). Estimation of molecular pairwise relatedness in autopolyploid crops. *G3 Genes Genomes Genet.* 10, 4579–4589. doi: 10.1534/g3.120.401669

Bastien, M., Boudhrioua, C., Fortin, G., and Belzile, F. (2018). Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. *Genome* 61, 449–456. doi: 10.1139/gen-2017-0236

Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R., and Munoz, P. (2019). How can a high-quality genome assembly help plant breeders? *Gigascience* 8:giz068. doi: 10.1093/gigascience/giz068

Bernardo, R. (2014). Genomewide selection when major genes are known. *Crop Sci.* 54:68. doi: 10.2135/cropsci2013.05.0315

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., and Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. *Front. Plant Sci.* 9:513. doi: 10.3389/fpls.2018.00513

Butler, D. G., Cullis, B. R., Gilmour, A. R., and Gogel, B. J. (2009). *{ASReml}-R Reference Manual*. Queensland: VSN International.

Cappai, F., Amadeu, R. R., Benevenuto, J., Cullen, R., Garcia, A., Grossman, A., et al. (2020a). High-resolution linkage map and QTL analyses of fruit firmness in autotetraploid blueberry. *Front. Plant Sci.* 11:767. doi: 10.3389/fpls.2020.562171

Cappai, F., Garcia, A., Cullen, R., Davis, M., and Munoz, P. R. (2020b). Advancements in low-chill blueberry *Vaccinium corymbosum* L. tissue culture practices. *Plants* 9:1624. doi: 10.3390/plants9111624

Caruana, B. M., Pembleton, L. W., Constable, F., Rodoni, B., Slater, A. T., and Cogan, N. O. I. (2019). Validation of genotyping by sequencing using transcriptomics for diversity and application of genomic selection in tetraploid potato. *Front. Plant Sci.* 10:670. doi: 10.3389/fpls.2019.00670

Cellon, C., Amadeu, R. R., Olmstead, J. W., Mattia, M. R., Ferrao, L. F. V., and Munoz, P. R. (2018). Estimation of genetic parameters and prediction of breeding values in an autotetraploid blueberry breeding population with extensive pedigree data. *Euphytica* 214, 1–13. doi: 10.1007/s10681-018-2165-8

Clark, L. V., Lipka, A. E., and Sacks, E. J. (2019). polyRAD: genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3 Genes Genomes Genet.* 9, 663–673. doi: 10.1534/g3.118.200913

Colantonio, V., Ferrao, L. F. V., Tieman, D., Bliznyuk, N., Sims, C., Klee, H., et al. (2020). Metabolomic selection for enhanced fruit flavor. *bioRxiv.* 1–24. doi: 10.1101/2020.09.17.302802

Colle, M., Leisner, C. P., Wai, C. M., Ou, S., Bird, K. A., Wang, J., et al. (2019). Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *Gigascience* 8:giz012. doi: 10.1093/gigascience/giz012

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinforma* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

de Bem Oliveira, I., Amadeu, R. R., Ferrão, L. F. V., and Muñoz, P. R. (2020). Optimizing whole-genomic prediction for autotetraploid blueberry breeding. *Heredity* 125, 437–448. doi: 10.1038/s41437-020-00357-x

de Bem Oliveira, I., Resende, M. F. R., Ferrão, L. F. V., Amadeu, R. R., Endelman, J. B., Kirst, M., et al. (2019). Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3 Genes Genomes Genet.* 9, 1189–1198. doi: 10.1534/g3.119.400059

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol.* 23, 40–69. doi: 10.1111/mec.12581

Enciso-Rodriguez, F., Douches, D., Lopez-Cruz, M., Coombs, J., and de Los Campos, G. (2018). Genomic selection for late blight and common scab resistance in tetraploid potato (*Solanum tuberosum*). *G3 Genes Genomes Genet.* 8, 2471–2481. doi: 10.1534/g3.118.200273

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Endelman, J. B., Carley, C. A. S., Bethke, P. C., Coombs, J. J., Clough, M. E., da Silva, W. L., et al. (2018). Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. *Genetics* 209, 77–87. doi: 10.1534/genetics.118.300685

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., et al. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95, 4114–4129. doi: 10.3168/jds.2011-5019

Fang, Y., Nunez, G. H., Silva, M. N., da, Phillips, D. A., and Munoz, P. R. (2020). A review for southern highbush blueberry alternative production systems. *Agronomy* 10:1531. doi: 10.3390/agronomy10101531

FAOSTAT (2021). FAOSTAT. Food and Agriculture Organization of United Nations. Available online at: http://www.fao.org/faostat/en/#data (accessed March 4, 2021)

Ferrão, L. F. V., Benevenuto, J., Oliveira, I. D. B., Cellon, C., Olmstead, J., Kirst, M., et al. (2018). Insights into the genetic basis of blueberry fruit-related traits using diploid and polyploid models in a GWAS context. *Front. Ecol. Evol.* 6:107. doi: 10.3389/fevo.2018.00107

Ferrão, L. F. V., Ferrão, R. G., Ferrão, M. A. G., Fonseca, A., Carbonetto, P., Stephens, M., et al. (2019). Accurate genomic prediction of *Coffea canephora* in multiple environments using whole-genome statistical models. *Heredity* 122, 261–275. doi: 10.1038/s41437-018-0105-y

Ferrão, L. F. V., Johnson, T. S., Benevenuto, J., Edger, P. P., Colquhoun, T. A., and Munoz, P. R. (2020). Genome-wide association of volatiles reveals candidate loci for blueberry flavor. *New Phytol.* 226, 1725–1737. doi: 10.1111/nph.16459

Gallais, A. (2003). *Quantitative Genetics and Breeding Methods in Autopolyploid Plants*. Paris: Quae.

Garcia, A. A. F., Mollinari, M., Marconi, T. G., Serang, O. R., Silva, R. R., Vieira, M. L. C., et al. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Sci. Rep.* 3:3399. doi: 10.1038/srep03399

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*. 1–9.

Gemenet, D. C., da Silva Pereira, G., De Boeck, B., Wood, J. C., Mollinari, M., Olukolu, B. A., et al. (2020). Quantitative trait loci and differential gene expression analyses reveal the genetic basis for negatively associated β-carotene and starch content in hexaploid sweetpotato [*Ipomoea batatas* (L.) Lam.]. *Theor. Appl. Genet.* 133, 23–36. doi: 10.1007/s00122-019-03437-7

Gerard, D., and Ferrão, L. F. V. (2020). Priors for genotyping polyploids. *Bioinformatics* 36, 1795–1800. doi: 10.1101/751784

Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., and Stephens, M. (2018). Genotyping polyploids from messy sequencing data. *Genetics* 210, 789–807. doi: 10.1534/genetics.118.301468

Gezan, S. A., Osorio, L. F., Verma, S., and Whitaker, V. M. (2017). An experimental validation of genomic selection in octoploid strawberry. *Hortic. Res.* 4, 1–9. doi: 10.1038/hortres.2016.70

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

Gilbert, J. L., Guthart, M. J., Gezan, S. A., de Carvalho, M. P., Schwieterman, M. L., Colquhoun, T. A., et al. (2015). Identifying breeding priorities for blueberry flavor using biochemical, sensory, and genotype by environment analyses. *PLoS ONE* 10:e0138494. doi: 10.1371/journal.pone.0138494

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

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

Iezzoni, A. F., McFerson, J., Luby, J., Gasic, K., Whitaker, V., Bassil, N., et al. (2020). RosBREED: bridging the chasm between discovery and application to enable DNA-informed breeding in rosaceous crops. *Hortic. Res.* 7, 1–23. doi: 10.1038/s41438-020-00398-7

Kalt, W., Cassidy, A., Howard, L. R., Krikorian, R., Stull, A. J., Tremblay, F., et al. (2020). Recent research on the health benefits of blueberries and their anthocyanins. *Adv. Nutr.* 11, 224–236. doi: 10.1093/advances/nmz065

Kerr, R. J., Li, L., Tier, B., Dutkowski, G. W., and McRae, T. A. (2012). Use of the numerator relationship matrix in genetic analysis of autopolyploid species. *Theor. Appl. Genet.* 124, 1271–1282. doi: 10.1007/s00122-012-1785-y

Lara, L. A., de, C., Santos, M. F., Jank, L., Chiari, L., Vilela, M. de, M., et al. (2019). Genomic selection with allele dosage in *Panicum maximum* Jacq. *G3 Genes Genomes Genet.* 9, 2463–2475. doi: 10.1534/g3.118.200986

Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P., and Marth, G. T. (2014). MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE* 9:e90581. doi: 10.1371/journal.pone.0090581

Liu, A., Lund, M. S., Boichard, D., Karaman, E., Guldbrandtsen, B., Fritz, S., et al. (2020). Weighted single-step genomic best linear unbiased prediction integrating variants selected from sequencing data by association and bioinformatics analyses. *Genet. Sel. Evol.* 52, 1–17. doi: 10.1186/s12711-020-00568-0

Lyrene, P. M. (2000). "Breeding southern highbush blueberries in Florida," in *VII International Symposium on Vaccinium Culture*, Vol. 574, 149–152.

Lyrene, P. M. (2005). Breeding low-chill blueberries and peaches for subtropical areas. *HortScience* 40, 1947–1949. doi: 10.21273/HORTSCI.40.7.1947

Mackay, I., Piepho, H., and Garcia, A. A. F. (2019). Statistical methods for plant breeding. *Handb. Stat. Genom.* Hoboken, NJ: John Wiley & Sons. 501–520. doi: 10.1002/9781119487845.ch17

Matias, F. I., Alves, F. C., Meireles, K. G. X., Barrios, S. C. L., do Valle, C. B., Endelman, J. B., et al. (2019). On the accuracy of genomic prediction models considering multi-trait and allele dosage in Urochloa spp. interspecific tetraploid hybrids. *Mol. Breed.* 39:100. doi: 10.1007/s11032-019-1002-7

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

Mollinari, M., Olukolu, B. A., Pereira, G. D. S., Khan, A., Gemenet, D., Yencho, G. C., et al. (2020). Unraveling the hexaploid sweetpotato inheritance using ultra-dense multilocus mapping. *G3 Genes Genomes Genet.* 10, 281–292. doi: 10.1534/g3.119.400620

Muñoz, P. R., Resende, M. F. R., Gezan, S. A., Resende, M. D. V., de los Campos, G., Kirst, M., et al. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* 198, 1759–1768. doi: 10.1534/genetics.114.171322

Noman, A., Aqeel, M., and He, S. (2016). CRISPR-Cas9: tool for qualitative and quantitative plant genome editing. *Front. Plant Sci.* 7:1740. doi: 10.3389/fpls.2016.01740

Omori, M., Yamane, H., Osakabe, K., Osakabe, Y., and Tao, R. (2021). Targeted mutagenesis of CENTRORADIALIS using CRISPR/Cas9 system through the improvement of genetic transformation efficiency of tetraploid highbush blueberry. *J. Hortic. Sci. Biotechnol.* 96, 153–161. doi: 10.1080/14620316.2020.1822760

Osorio, L. F., Gezan, S. A., Verma, S., and Whitaker, V. (2020). Independent validation of genomic prediction in strawberry over multiple cycles. *Front. Genet.* 11:1862. doi: 10.3389/fgene.2020.596258

Pereira, G. S., Garcia, A. A. F., and Margarido, G. R. A. (2018). A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. *BMC Bioinform.* 19, 1–10. doi: 10.1186/s12859-018-2433-6

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181. doi: 10.1086/302959

R Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Team.

Ren, D., An, L., Li, B., Qiao, L., and Liu, W. (2021). 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., and Lipka, A. E. (2019). Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in Maize and Sorghum. *Plant Genome* 12:180052. doi: 10.3835/plantgenome2018.07.0052

Rosyara, U. R., De Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome* 9, 1–10. doi: 10.3835/plantgenome2015.08.0073

Sehgal, D., Rosyara, U., Mondal, S., Singh, R., Poland, J., and Dreisigacker, S. (2020). Incorporating genome-wide association mapping results into genomic prediction models for grain yield and yield stability in CIMMYT spring bread wheat. *Front. Plant Sci.* 11:197. doi: 10.3389/fpls.2020.00197

Sharpe, R. H., and Sherman, W. B. (1971). Breeding blueberries for low-chilling requirement. *HortScience* 6, 145–147.

Slater, A. T., Cogan, N. O. I., Forster, J. W., Hayes, B. J., and Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *Plant Genome* 9, 1–15. doi: 10.3835/plantgenome2016.02.0021

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J., et al. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116, 395–408. doi: 10.1038/hdy.2015.113

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

Sverrisdóttir, E., Byrne, S., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., et al. (2017). Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. *Theor. Appl. Genet.* 130, 2091–2108. doi: 10.1007/s00122-017-2944-y,

Uitdewilligen, J. G., Wolters, A.-M. A., Bjorn, B., Borm, T. J. A., Visser, R. G. F., and van Eck, H. J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE* 8:e62355. doi: 10.1371/journal.pone.0062355

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

Williams, J. S. (1962). The evaluation of a selection index. *Biometrics* 18, 375–393. doi: 10.2307/2527479

Zhang, X., Lourenco, D., Aguilar, I., Legarra, A., and Misztal, I. (2016). Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. *Front. Genet.* 7:151. doi: 10.3389/fgene.2016.00151

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* 9:e1003264. doi: 10.1371/journal.pgen.1003264

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., et al. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci.* 11:25. doi: 10.3389/fpls.2020.00025

Conflict of Interest: IB was employed by company Hortifrut North America, Inc.; however, she participated in this research while at University of Florida.

The remaining 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 Plant Science

ORIGINAL RESEARCH
published: 22 June 2021
doi: 10.3389/fpls.2021.658978

OPEN ACCESS

Edited by:
Valentin Wimmer,
KWS Saat, Germany

Reviewed by:
Vincent Garin,
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India
Fikret Isik,
North Carolina State University, United States

***Correspondence:**
Michael Olsen
m.olsen@cgiar.org
Kelly R. Robbins
krr73@cornell.edu

†ORCID
Sikiru Adeniyi Atanda
orcid.org/0000-0001-8758-2063
Michael Olsen
orcid.org/0000-0002-8818-6238
Jose Crossa
orcid.org/0000-0001-9429-5855
Juan Burgueño
orcid.org/0000-0002-1468-4867
Renaud Rincent
orcid.org/0000-0003-0885-0969
Yoseph Beyene
orcid.org/0000-0001-7221-2617
Manje Gowda
orcid.org/0000-0003-4434-6364
Kate Dreher
orcid.org/0000-0003-4652-4398
Prasanna M. Boddupalli
orcid.org/0000-0002-5761-2273
Pangirayi Tongoona
orcid.org/0000-0003-1030-8589
Eric Yirenkyi Danquah
orcid.org/0000-0002-7870-0432
Kelly Robbins
orcid.org/0000-0001-9522-9585

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: 26 January 2021
Accepted: 25 May 2021
Published: 22 June 2021

Scalable Sparse Testing Genomic Selection Strategy for Early Yield Testing Stage

Sikiru Adeniyi Atanda[1,2,3†], Michael Olsen[4*†], Jose Crossa[2†], Juan Burgueño[2†], Renaud Rincent[5†], Daniel Dzidzienyo[1], Yoseph Beyene[4†], Manje Gowda[4†], Kate Dreher[2†], Prasanna M. Boddupalli[4†], Pangirayi Tongoona[1†], Eric Yirenkyi Danquah[1†], Gbadebo Olaoye[6] and Kelly R. Robbins[3*†]

[1] West Africa Center for Crop Improvement (WACCI), University of Ghana, Accra, Ghana, [2] International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, [3] Section of Plant Breeding and Genetics, School of Integrative Plant Sciences, Cornell University, Ithaca, NY, United States, [4] International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya, [5] French National Institute for Agriculture, Food, and Environment (INRAE), Paris, France, [6] Agronomy Department, University of Ilorin, Ilorin, Nigeria

To enable a scalable sparse testing genomic selection (GS) strategy at preliminary yield trials in the CIMMYT maize breeding program, optimal approaches to incorporate genotype by environment interaction (GEI) in genomic prediction models are explored. Two cross-validation schemes were evaluated: CV1, predicting the genetic merit of new bi-parental populations that have been evaluated in some environments and not others, and CV2, predicting the genetic merit of half of a bi-parental population that has been phenotyped in some environments and not others using the coefficient of determination (CDmean) to determine optimized subsets of a full-sib family to be evaluated in each environment. We report similar prediction accuracies in CV1 and CV2, however, CV2 has an intuitive appeal in that all bi-parental populations have representation across environments, allowing efficient use of information across environments. It is also ideal for building robust historical data because all individuals of a full-sib family have phenotypic data, albeit in different environments. Results show that grouping of environments according to similar growing/management conditions improved prediction accuracy and reduced computational requirements, providing a scalable, parsimonious approach to multi-environmental trials and GS in early testing stages. We further demonstrate that complementing the full-sib calibration set with optimized historical data results in improved prediction accuracy for the cross-validation schemes.

Keywords: genomic selection, factor analytic, preliminary yield trials, prediction accuracy, unstructured model, CDmean

INTRODUCTION

Due to climate change threatening crop productivity in sub-Saharan Africa (SSA), breeding for drought tolerance and yield stability across target environments is a high priority for the International Maize and Wheat Improvement Center (CIMMYT) tropical maize breeding program (Beyene et al., 2015, 2019). To achieve genetic gain improvement in alignment with these breeding objectives, the CIMMYT maize breeding programs leverage novel technologies such as doubled haploid (DH) technology, that allows generation of tens of thousands of inbred lines yearly, a low-cost genotyping platform, and genomic selection (GS) that uses whole-genome information to predict the genetic merit of new lines. The CIMMYT maize breeding scheme has five stages of testing. Many hybrid combinations are developed each year and tested in a small number of environments during the early testing phase, in later stages a small number of selected hybrid combinations are tested in many environments. To identify parental lines for the next breeding cycle and develop stress tolerant and high yielding hybrids that meet farmers' needs, hybrids are tested under both well-watered (WW) and water-stress (WS) conditions in the preliminary screening stages. Each stage is characterized by the number of locations and the number of testers. These factors influence selection accuracy in the different testing stages.

At stage 1 or preliminary yield trials, several experimental hybrids are generated by crossing DH lines, or lines developed using the pedigree scheme, to a tester from a complementary heterotic group. The testcross hybrids are evaluated in 3–5 environments, where each environment is a combination of location and management (WS and WW), and the data are used to select the best 10–15 percent of the lines within or across the managements for advancement to stage 2 yield trials (Beyene et al., 2019). Effective selection decisions at stage 1 yield testing are critical for the advancement of lines with the greatest potential to perform in the resource-intensive multi-location, multi-tester testing stages. However, the effectiveness of phenotypic selection (PS) for stage 1 testcross trials is limited by evaluation on one tester and in few environments, which do not adequately represent the target population of environments (Endelman et al., 2014), this is largely due to the number of DH lines for testcross and the number of testcross hybrids for evaluation. Consequently, the CIMMYT Global Maize breeding program is focused on redesigning early-stage yield trials to accelerate genetic gain and reduce the cost of hybrid testing by evolving from a phenotypic based selection to the use of GS to predict the genetic merit of new lines. The efficiency of this method for evaluation of stage 1 candidates has been established (Beyene et al., 2019).

The current GS strategy relies on phenotyping 50 percent of a bi-parental population, observed across WW and WS environments, to predict the genetic merit of un-tested candidates for both WW and WS (Beyene et al., 2015, 2019; Santantonio et al., 2020) in a test-half-predict-half strategy (Atanda et al., 2020). While this strategy results in improved prediction accuracy at lower cost, it is not optimal for reducing breeding cycle time because a subset of the bi-parental population is required for model training (Atanda et al., 2020). The goal of the CIMMYT maize breeding program is to accelerate the early yield testing stage by using information from previously tested genotypes that have been phenotyped and genotyped (historical data) for model training. Based on the predicted genomic estimated breeding value (GEBV), lines will be advanced directly to stage 2 yield trials, the effectiveness of this strategy has been evaluated in our previous study.

Sparse testing represents a promising approach to expand the number of lines tested when GS is used to advance lines directly into stage 2, and for stage 1 screening of lines in cases where the genetic merit of some new lines may not be accurately predicted due to low genetic relationship between new lines and previously evaluated genotypes in the historical dataset. In the case where GEBV of lines cannot be accurately predicted from historical data, sparse testing has been identified as an optimal GS strategy compared to the current CIMMYT GS strategy (test-half-predict-half) that tests half of a full sib family to train genomic prediction models for full sibs that are not tested in stage 1 (Atanda et al., 2020; Santantonio et al., 2020). Given that all populations have phenotypic records in different environments, it is an appealing option for creating a robust historical dataset and allows for borrowing of information across environments resulting in improved prediction accuracy when compared to the test-half-predict-half strategy (Burgueño et al., 2012; Atanda et al., 2020; Santantonio et al., 2020).

To identify a scalable strategy that optimizes the representation of genetic space of the genotypes across environments leading to efficient use of information across the environments at the early yield testing stage, we evaluated two different breeding scenarios: (1) predicting the genetic merit of new bi-parental populations across environments (phenotyping of populations was unbalanced across environments) or, (2) predicting different subsets of a bi-parental population across environments. Here, coefficient of determination (CDmean) was used to split bi-parental populations across environments.

The main objectives of this study were to: (1) determine an effective strategy to implement sparse testing within the CIMMYT tropical maize breeding program and, (2) determine the optimal method to incorporate genotype by environment interaction (GEI) into the GS model for early yield testing stage.

MATERIALS AND METHODS

Plant Materials

The datasets used in this study are described in detail in Atanda et al. (2020). Briefly, the maize datasets consist of 849 and 1,389 DH lines derived from 13 and 45 DH bi-parental populations respectively. The DH lines were unique within each year and were testcrossed to one of three single-cross testers in 2017 and one of two single-cross testers in 2018 respectively. Testcrosses in 2017 and 2018 were grouped into 13 and 34 trials, respectively. The trials were connected by common checks, and each trial was planted in an alpha-lattice incomplete block design with two replications under WW condition in Kiboko and Kakamega, Kenya and WS condition, in Kiboko during the 2017 and 2018 growing seasons. The entries in the trials were planted two-rows per plot, each row was 5 m long, with spacing of 0.75 m between rows and 0.25 m between hills. At planting, two seeds per hill were planted and thinned to one plant per hill 3 weeks after emergence to obtain a final plant population density of 53,333 plants per hectare. Fertilizers were applied at the rate of 60 kg N and 60 kg P_2O_5 per ha, as recommended for the area. Nitrogen was applied in a split dose at planting and 6 weeks after emergence. For the purposes of modeling genotype by environmental interactions

(GEI), several combinations of factors (location, management, and year) were used to classify environments as summarized in **Table 1**.

All DH lines were genotyped using repeat Amplification Sequencing (rAmpSeq) at Cornell Life Science Core Laboratory Center, Ithaca, NY, United States. The genotyping platform takes advantage of knowledge of whole-genome sequences and repetitive sequences to identify DNA sequence polymorphisms using novel bioinformatics tools [for detail see Buckler et al. (2016)]. It provides dominant markers, with the 9,155 sequence tags coded as 0 and 2 based on presence or absence of the dominant marker, respectively. The 6,785 markers with minor allele frequency greater than 0.05 were used for analysis.

Genomic Selection Models

A separate analysis was run for each of the environmental classifications found in **Table 1** using a multi-environment linear mixed model incorporating GEI effect. The covariance structures were defined using the groups in **Table 1** and the model was fit in ASReml using the average information algorithm (Gilmour et al., 1995) as:

$$y = 1_n\mu + X_1b_1 + Z_1u_1 + Z_2u_2 + Z_3u_3 + Z_4u_4 + Z_5u_5 + \varepsilon \tag{1}$$

where y (n × 1) is the vector of phenotypes for each DH lines measured in the environments (1...k), μ is the overall mean and 1_n (n × 1) is a of vector ones, b_1 is a fixed effect of location, $u_1$is the random effect of the interaction between the genomic effect of g-th DH line and v-th environment, u_2 is the random effect of the tester, u_3 is the random effect of the trial, u_4 is the random effect of replication nested within environment, trial and year for the multi-year dataset, u_5 is the random effects of incomplete block nested within replication, trial, location and year for the multi-year dataset. The number of fixed and random effects is represented as n and p, while X_n and Z_p are incidence matrices for fixed and random effects, respectively. The variance of the random effects u_2, u_3, u_4, and $u_5$were assumed to be distributed as:

$$u_p \sim N(0, I_p\sigma^2_{u_p}) \tag{2}$$

where I_p and $\sigma^2_{u_p}$ are the identity matrix and variance of the p-th random effect (u_2- u_5). In Equation 1 all fixed effects and random effects u_2- u_5 are model in the same way for all analyses, while the covariance structure for u_2 and ε varied based on the environmental classifications in **Table 1**.

TABLE 1 | Classification of the environments based on management, location by management, management by year and location by management by year.

Grouping of the environments		Environment
Location by management	Kiboko by WW	LM1
	Kakamega by WW	LM2
	Kiboko by WS	LM3
Management (single year analysis)	WW	M1
	WS	M2
Management by year	WW by 2017	MY1
	WS by 2017	MY2
	WW by 2018	MY3
	WS by 2018	MY4
Management^{++} (multi-year analysis)	WW	M^{+}1
	WS	M^{+}2
Location by management by year	Kiboko by WW by 2017	LMY1
	Kakamega by WW by 2017	LMY2
	Kiboko by WS by 2017	LMY3
	Kiboko by WW by 2018	LMY4
	Kakamega by WW by 2018	LMY5
	Kiboko by WS by 2018	LMY6

M^{+} is the broad classification of management across years as WW and WS.

The random GEI effect u_1 is defined as the Kronecker product ($\bigotimes$) between the g × g genomic relationship matrix (G) and the v × v variance-covariance matrix of the genomic effect of genotypes in and between environments (G_o).

$$u_1 \sim N[0, (G \otimes G_o)] \tag{3}$$

Thus, covariance of the genomic effect of the line (u_1) in multi-environment model, can be represented as:

$$Cov(u_1, u_1^{'}) = G_o \otimes G \tag{4}$$

$$G_o \otimes G = \begin{bmatrix} \sigma^2_{g_1} & \sigma_{g_{12}} & \cdots & \sigma_{g_{1v}} \\ \sigma_{g_{21}} & \sigma^2_{g_2} & \cdots\cdots & \\ & \vdots & & \vdots \\ \sigma_{g_{v1}} & & \ddots & \\ & \vdots & & \sigma^2_{g_v} \end{bmatrix} \otimes G(5) \tag{5}$$

where G_o represents the v × v variance-covariance matrix of the genomic effect of genotypes in the environments. The number of environments v varied based on the environmental classifications in **Table 1**. The diagonal of the G_o matrix is the additive genetic variance $\sigma^2_{g_v}$within the v-th environment. The off-diagonal (σ_{g_1v}) elements represent the genetic covariance between environments.

Fitting the GEI in this way enables examination of the predictive ability of an unstructured model (US) that allows fitting unequal covariance between pairs of environments or managements, in addition to different genetic variances within environment/management. However, the number of parameters to estimate for the US model does not increase linearly with the number of environments, which can result in non-convergence when the number of model parameters is large relative to the number of data points (Smith et al., 2001; Kelly et al., 2007; Oakey et al., 2016). The factor analytic (FA) model has been identified as a more parsimonious approach to fit the complex covariance structure amongst a large number of environments (Piepho, 1998; Smith et al., 2001; Crossa et al., 2004; Oakey et al., 2016; Smith and Cullis, 2018). FA identifies one or few factors underlying the correlation among the k environments by their relationship to unobservable latent variables. Therefore, the GEI is modeled as interaction between the genomic effect of the g-th DH line and one or few factors underlying the environmental/management influences on the genotype (Piepho, 1998; Smith et al., 2001; Crossa et al., 2004; Kelly et al., 2007). FA model for $\mathrm{Cov}(u_g, u_g^{'})$ is expressed as:

$$(\Lambda\Lambda' + \Psi) \otimes G \tag{6}$$

where Λ is a v × m matrix of loading factors, the columns of Λ are associated with the environmental loadings for the m-th latent factor. Ψ is a v × v heterogeneous diagonal matrix with specific environment genetic variances Ψ_v on the diagonal and zero covariance between environments. When the number of environments was less than 4 (as defined in **Table 1**), one multiplicative component was considered (m = 1) and m = 2 as number of environments increased from 4 to 6. We use the extended FA (XFA) model that allows a non-full rank variance matrix for the GEI effects, therefore the mixed model equation is sparser, resulting in reduced computational requirements compared to the standard FA model. Details can be found in Thompson et al. (2003) and Meyer (2009).

The residual variance for the GS model (Equation 1) can be specified as:

$$\varepsilon \sim N(0, R) \tag{7}$$

where R is a heterogeneous diagonal matrix of the residual variances for each environment v:

$$R = \begin{bmatrix} \sigma^2_{\varepsilon_1} * \mathrm{I}_{n_1} & 0 & \cdots 0 \\ 0 & \sigma^2_{\varepsilon_2} * \mathrm{I}_{n_2} & \cdots 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \sigma^2_{\varepsilon_v} * \mathrm{I}_{n_v} \end{bmatrix} \tag{8}$$

where I_{n_v} is a $n_v = n_v$ identity matrix and n_v is the number of observations in environment v. The off-diagonal elements of the R matrix equal zero [$\mathrm{Cov}(\varepsilon, \varepsilon^{'}) = 0$] and diagonal elements represent the residual variance within each of v environments. Generally, the residual variance for multi-environment GS models can take two different forms explaining different model assumptions. For example, a uniform residual variance for all environments ($\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2} \ \ldots = \sigma^2_{\varepsilon_v}$), and a heterogeneous residual variance where each environment has different residual variance ($\sigma^2_{\varepsilon_1} \neq \sigma^2_{\varepsilon_2} \ \ldots \neq \sigma^2_{\varepsilon \ v}$).

The plot level heritability for each environment was calculated from the variance components obtained from the model as:

$$h_v^2 = \frac{\sigma^2_{g_v}}{\sigma^2_{g_v} + \sigma^2_{\varepsilon_v}} \tag{9}$$

where $\sigma^2_{g_v}$ and $\sigma^2_{\varepsilon_v}$ are the genetic and residual variance estimates specific to environment v.

Calibration Set Optimization Criteria

Following Atanda et al. (2020), CDmean and Avg_GRM were used as genetic optimization criteria. Similar to Rincent et al. (2012), CDmean was used to optimize experimental design by determining which individuals were evaluated in each environment. However, in this study, CDmean is the mean of the expected reliability of the predicted genetic values of N-1 individuals in a specific bi-parental population, where N is the size of a given full-sib family with each g-th individual used to predict the reliability of the remaining full-sibs. The expected reliability of the prediction of the different contrasts was expressed as:

$$\mathrm{CD}\,(\mathrm{K}) = \mathrm{diag}\left[\frac{\mathrm{K}^{'}(\mathrm{G} - \lambda(\mathrm{Z}^{'}\mathrm{D}\mathrm{Z} + \lambda \mathrm{G}^{-1})^{-1})\mathrm{K}}{\mathrm{K}^{'}\mathrm{G}\mathrm{K}}\right] \tag{10}$$

where $\mathrm{D} = 1 - \mathrm{X}(\mathrm{X}^{'}\mathrm{X})^{-1}\mathrm{X}^{'}$, G, X, and Z are the same as defined above and K is a matrix of contrast vectors with the sum of each contrast vector equal to zero such that $1'\mathrm{K} = 0$.

In principle $\lambda = \sigma^2_{\varepsilon}/\sigma^2_{g}$, where σ^2_{ε} is the residual error and σ^2_{g} is the genetic variance obtained from Equation 1; however, this cannot be calculated for untested lines. According to Atanda et al. (2020), the efficiency of CDmean is not highly dependent on trait heritability but rather on genomic relationship. Consequently, λ was set to 0.5. In our previous study, when an intermediate value was chosen for (λ = 0.5) the prediction accuracy was close to accuracies achieved using $\lambda = \sigma^2_{\varepsilon}/\sigma^2_{g}$, this was in agreement with Rincent et al. (2012). Therefore, CDmean = mean [diag(CD(K))], each column of the K matrix is a contrast between (N-1) individuals of a full-sib family and the mean of the full-sib family. A contrast using the first individual in the family is set up as:

$$\mathrm{K}_1 = \mathrm{c}\left(\frac{\mathrm{n}-1}{\mathrm{n}}, \frac{-1}{\mathrm{n}}, \frac{-1}{\mathrm{n}}\right) \tag{11}$$

Where n is the number of individuals in the populations. Therefore, one individual of a full-sib in a specific bi-parental population serves as a calibration set to estimate the reliability of predicting the remaining full-sibs. This was repeated N times enabling each g-th individual of a full-sib to serve as calibration set. Consequently, we obtain a CDmean value for each individual in a given bi-parental population and individuals (50 percent of a bi-parental population) with the highest CDmean value represent an optimized calibration set. Theoretically, individuals with high CDmean value maximize the reliability of those with low CDmean value, thus full-sib families where split between environment by keeping high and low CDmean lines together in WW environments, respectively. In the WS environment, a portion of lines from each WW environment were used as the calibration set (**Supplementary Figure 2**). A script to calculate the CDmean is provided in **Supplementary File 1**. This strategy was adopted because it is computationally efficient compared to Rincent et al. (2012) which used an exchange algorithm to randomly exchange one individual between the calibration set (N$^{'}$, – total number of individuals to phenotype) and the un-phenotyped individuals (N- N$^{'}$), where the exchange is accepted if the initial CDmean value improved and rejected otherwise. The process repeated until reaching a plateau. Akdemir et al. (2015) and Heslot and

Feoktistov (2020) also modified Rincent et al. (2012) with improved computational efficiency. The efficacy of these methods was not compared in our study, but results from preliminary analysis show the strategy used in this study improved prediction accuracy compared to Rincent et al. (2012) (results not shown).

The Avg_GRM is a raw estimate of the proportion of the genome shared between a potential training set and all individuals in a specific full-sib family. Based on the results from our previous study (Atanda et al., 2020), CDmean and Avg_GRM genetic optimization criteria have similar efficiency in selecting individuals from historical data closely related with a specific bi-parental population. However, Avg_GRM genetic optimization criterion is computationally more efficient; thus, the Avg_GRM genetic optimization criterion was used to select 300 individuals from the historical data that are closely related to a specific full-sib family. The Avg_GRM can be expressed as:

$$\text{Avg\_GRM}_j = \frac{1}{n}\sum_{g}^{n} G_{gj} \qquad (12)$$

where G_{gj} is the genomic relationship between the g-th individual in a target full-sib family and the j-th line in the historical data and n is the size of target full-sib family.

Cross-Validation Scheme

The predictive ability of two cross-validation schemes was evaluated for possible implementation of a sparse testing GS strategy in the CIMMYT tropical maize breeding program. For even distribution of populations across environments, a bi-parental population with size ≤ 30 was dropped from 2017 dataset and the remaining 12 bi-parental populations were used for the analysis. The first cross-validation scheme (CV1) involved masking six random bi-parental populations of the twelve bi-parental populations in one WW environment with the remaining bi-parental populations masked in the other WW environment. In the WS environment, three random bi-parental populations from each WW environment were masked; this process was repeated 10 times (**Supplementary Figure 1**). Prediction accuracy was calculated as the Pearson correlation of the predicted GEBV obtained from the models and the BLUE estimates of DH testcrosses for each population in each environment. The mean across populations is reported.

In the second cross-validation scheme (CV2), CDmean was used for splitting each bi-parental population equally across WW environments by masking 50 percent of a bi-parental population with lowest CDmean value in one environment and the remaining 50 percent masked in the other WW environment. For the WS environment, half of the individuals unmasked in the WW environments were masked (**Supplementary Table 1** and

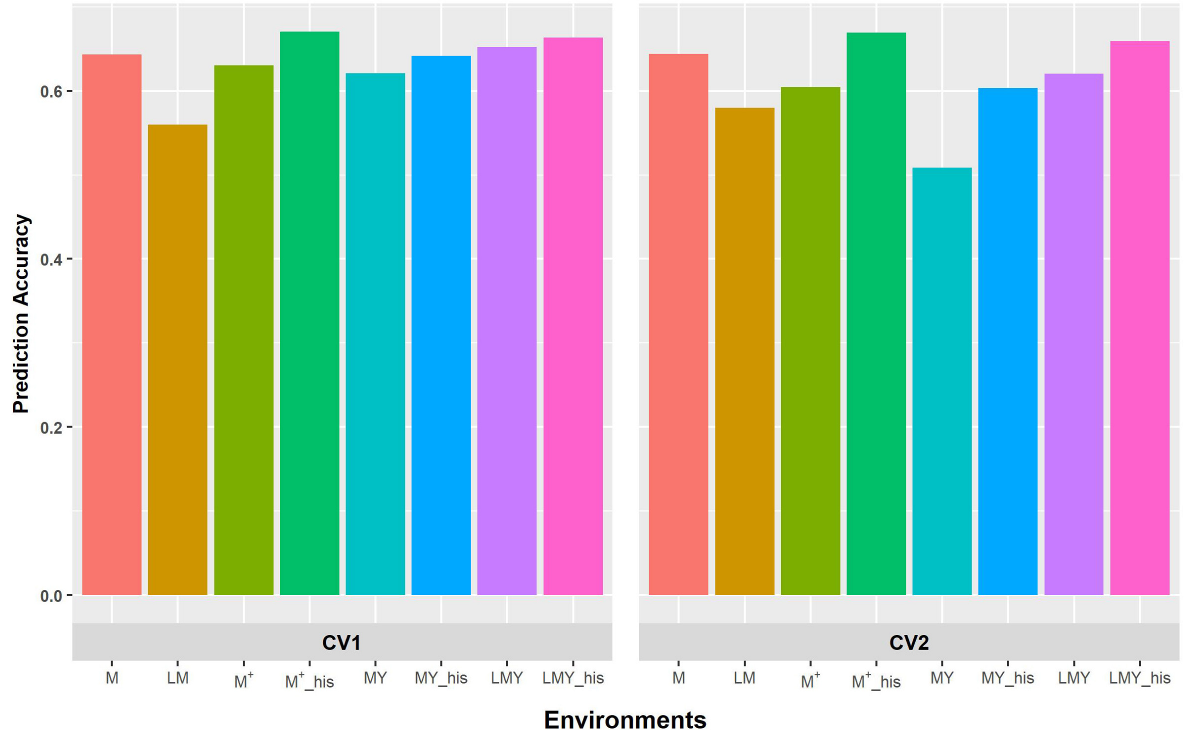


FIGURE 1 | Predictive ability of factor analytic model for the cross-validation schemes (CV1 and CV2) in WS environments/management. LM and M represent prediction accuracy obtained when covariance was modeled across environments and managements, respectively, for within-year prediction. LMY represents classification of environment as location by management by year, MY and M$^+$ represent the broad classification of the management across years as WW and WS, and explicit definition of the management across years as WW 2017 and 2018 and WS 2017 and 2018. LMY, MY and M$^+$ used all available historical data. The suffix "his" represents prediction accuracy obtained with optimized historical data using the Avg_GRM genetic optimization criterion.

Figure 1). Due to the diversity of populations in 2018, the 2018 dataset was chosen to represent "historical" data in this study. Following Atanda et al. (2020), we further assessed the predictive ability of augmenting the training set in both cross-validation schemes with all historical data or with an optimized set of 300 individuals from the historical records closely related to a specific full-sib family using Avg_GRM genetic optimization criterion. In the scenario where full-sib training sets were augmented with historical data, GEI was considered as location by management by year (LMY 1, 2, 3, 4, 5, and 6), management by year (MY 1, 2, 3, and 4) to account for the difference between managements across years in addition to the broad definition of management as WW ($M^{+}1$) and WS ($M^{+}2$). The prediction accuracy was calculated as the Pearson correlation of the predicted GEBV and the BLUE estimates of DH lines in each environment, obtained using the complete dataset for each population, from the combined analysis. The mean across populations is reported.

RESULTS

Residual Variance, Heritability Within Environment/Management, and Correlation Between Pairs of Environments/Managements

Except for when the environment was classified as year by management by location (LMY 1, 2, 3, 4, 5, and 6), where the US model was responsive to the training set and did not consistently converge, the results for FA and US models were equivalent regardless of the cross-validation schemes (Result not shown). Thus, only results from FA model were presented. The genetic correlation between environments (LM 1, 2, and 3) in the CV1 ranges from 0.13 to 0.64 (**Table 2**). A similar trend was observed for CV2 and ranges from 0.22 to 0.363. For CV1, the within environments (LM 1, 2, and 3) plot-level heritability for grain yield ranges from 0.27 to 0.42 and ranges from 0.26 to 0.32 in CV2. When environments were grouped into managements, for CV1, the genetic correlation between WW (M1) and WS (M2) was 0.37 and plot-level heritability within each management was 0.24 and 0.35 respectively. While for CV2, the genetic correlation between M1 and M2 was 0.47, and plot-level heritability within each management was 0.19 and 0.32.

The genetic correlation between environments (LMY 1, 2, 3, 4, 5, and 6) varies across the cross-validation schemes, it ranges from −0.14 to 0.74 for CV1 and −0.02 to 0.79 for CV2. The plot level heritability for each environment across the cross-validation was modest. In analyses where management was defined across years (WW 2017 and 2018 – MY1 and 3, WS 2017 and 2018 – MY2 and 4), the genetic correlation between managements also ranged from negative to moderate correlation for CV1 (**Table 3**). While it ranged from low to moderate in CV2. For the broad definition of management across years as WW ($M^{+}1$) and WS ($M^{+}2$), the genetic correlation was 0.60 and 0.68 for CV1 and CV2, respectively. Generally, the estimates of plot-level heritability for CV1 and CV2 were moderate.

Comparison of Predictive Ability of the Models and the Cross-Validation Schemes

The grouping of the environments into management consistently shows higher prediction accuracy compared to modeling of covariance between environments defined as a combination of location, management and year (**Figures 1**, **2**). Though the prediction accuracy for the cross-validation schemes was similar, the slight difference corroborates the different estimates of heritability and genetic correlation obtained from the cross-validation schemes. The augmentation of the training set with optimized historical information improved prediction accuracy compared to either use of all the historical data plus the full-sib training set or only the full-sib training set. Unsurprisingly, prediction accuracy increases with higher heritability and genetic correlation between environments/managements as observed with prediction accuracy of WW compared to WS. Although prediction accuracy of FA and US models are similar (**Supplementary Table 2**), the US model failed to consistently converge when environment was defined based on the combination of location, management, and year.

TABLE 2 | Plot level heritability (diagonal) and genetic correlations between pairs of managements or environments (upper diagonal) for the two managements (upper half) and three environments (lower half) from the factor analytic model analysis of 2017 dataset.

	Cross-validation scheme					
	CV1			CV2		
M	WW	WS		WW	WS	
WW	0.24 (0.08)	0.37	–	0.19 (0.06)	0.47	–
WS		0.35 (0.06)	–		0.32 (0.09)	–
LM	Kiboko WW	Kakamega WW	Kiboko WS	Kiboko WW	Kakamega WW	Kiboko WS
Kiboko WW	0.27 (0.09)	0.24	0.63	0.26 (0.06)	0.31	0.63
Kakamega WW		0.42 (0.06)	0.15		0.32 (0.10)	0.22
Kiboko WS			0.34 (0.06)			0.32 (0.04)

M represents grouping of locations by management as WW and WS; LM represents the grouping of locations as Kiboko-WW, Kakamega-WW and Kiboko-WS. Plot level heritability estimates within each grouping management (M or LM) are represented in the diagonal. The upper diagonals are genetic correlations between environmental groupings. Standard errors for the heritability estimates are in parentheses.

TABLE 3 | Plot level heritability (diagonal) and genetic correlations between pairs of managements (upper diagonal) for the two managements (upper half) and four managements (lower half) from the factor analytic model analysis of combined 2017 and 2018 dataset.

				CV1		
M^+	WW	WS				
WW	0.31 (0.05)	0.60	–	–	–	–
WS		0.38 (0.03)	–	–	–	–
MY	WW 2017	WS 2017	WW 2018	WS 2018		
WW 2017	0.32 (0.03)	0.31	0.10	0.05	–	–
WS 2017		0.38 (0.03)	−0.11	0.55	–	–
WW 2018			0.27 (0.05)	0.09	–	–
WS 2018				0.20 (0.03)	–	–
LMY	Kiboko WW 2017	Kakamega WW 2017	Kiboko WS 2017	Kiboko WW 2018	KakamegaWW 2018	Kiboko WS 2018
Kiboko WW 2017	0.30 (0.07)	−0.03	0.45	0.04	−0.14	0.19
Kakamega WW 2017		0.46 (0.08)	−0.10	0.29	0.38	0.16
Kiboko WS 2017			0.41 (0.04)	0.23	−0.10	0.32
Kiboko WW 2018				0.49 (0.04)	0.69	0.74
KakamegaWW 2018					0.50 (0.08)	0.33
Kiboko WS 2018						0.38 (0.04)
			CV2			
M^+	WW	WS				
WW	0.35 (0.04)	0.68				
WS		0.39 (0.05)				
MY	WW 2017	WS 2017	WW 2018	WS 2018		
WW 2017	0.35 (0.04)	0.47	0.36	0.20		
WS 2017		0.38 (0.04)	0.30	0.59		
WW 2018			0.15 (0.07)	0.38		
WS 2018				0.20 (0.05)		
LMY	Kiboko WW 2017	Kakamega WW 2017	Kiboko WS 2017	Kiboko WW 2018	KakamegaWW 2018	Kiboko WS 2018
Kiboko WW 2017	0.27 (0.07)	−0.01	0.32	0.12	−0.10	0.19
Kakamega WW 2017		0.38 (0.05)	0.38	0.42	0.54	0.12
Kiboko WS 2017			0.38 (0.06)	0.26	−0.02	0.34
Kiboko WW 2018				0.53 (0.10)	0.73	0.79
KakamegaWW 2018					0.54 (0.05)	0.55
Kiboko WS 2018						0.36 (0.05)

M^+ *represents broad classification of management across years as WW and WS. MY represents the grouping of environments by management (WW and WS) and year (2017 and 2018). LMY groups environments by management (WW and WS), location (Kakamega and Kiboko), and year (2017 and 2018). Plot level heritability estimates for* M^+*, MY, and LMY are represented in the diagonal. The upper diagonals are genetic correlations between environmental groupings. Standard errors for the heritability estimates are in parentheses.*

DISCUSSION

The sparse testing GS strategy in which the genetic merit of new lines is evaluated in different but genetically correlated environments has proven to increase prediction accuracy compared to the test-half-predict-half GS strategy and, provided that all new lines have phenotypic data, it is seemingly robust for developing historical training datasets (Burgueño et al., 2012; Atanda et al., 2020; Santantonio et al., 2020). The evaluation of new genotypes across environments allows the utilization of information across environments using multi-environment models. However, multi-environment models, especially the US model, tend to become non-parsimonious as the number of environments increases resulting in convergence failure (Smith et al., 2001; Kelly et al., 2007; Meyer, 2009). Considering that a small number of environments and genotypes were evaluated in the preliminary yield trials in this study, the use of the US model did not pose any statistical challenge. However, inclusion of historical data in the training set increases the number of environments, which could result in computational challenges for the US approach. Alternatively, the FA model, which is a complexity reduction model for an increased number of environments, requires fewer parameters while accounting for covariance between environments (Smith et al., 2001; Thompson et al., 2003; Crossa et al., 2004; Kelly et al., 2007; Burgueño et al., 2008, 2011, 2012; Smith and Cullis, 2018; Tolhurst et al., 2019), and could be more suitable as historic training datasets increase in size and complexity.

Although the predictive ability of the two cross-validation schemes is comparable, the improved prediction accuracy of CV1 might be due to the close relationship (half-sib relationship) of all the populations. Previous studies (Lehermeier et al., 2014;

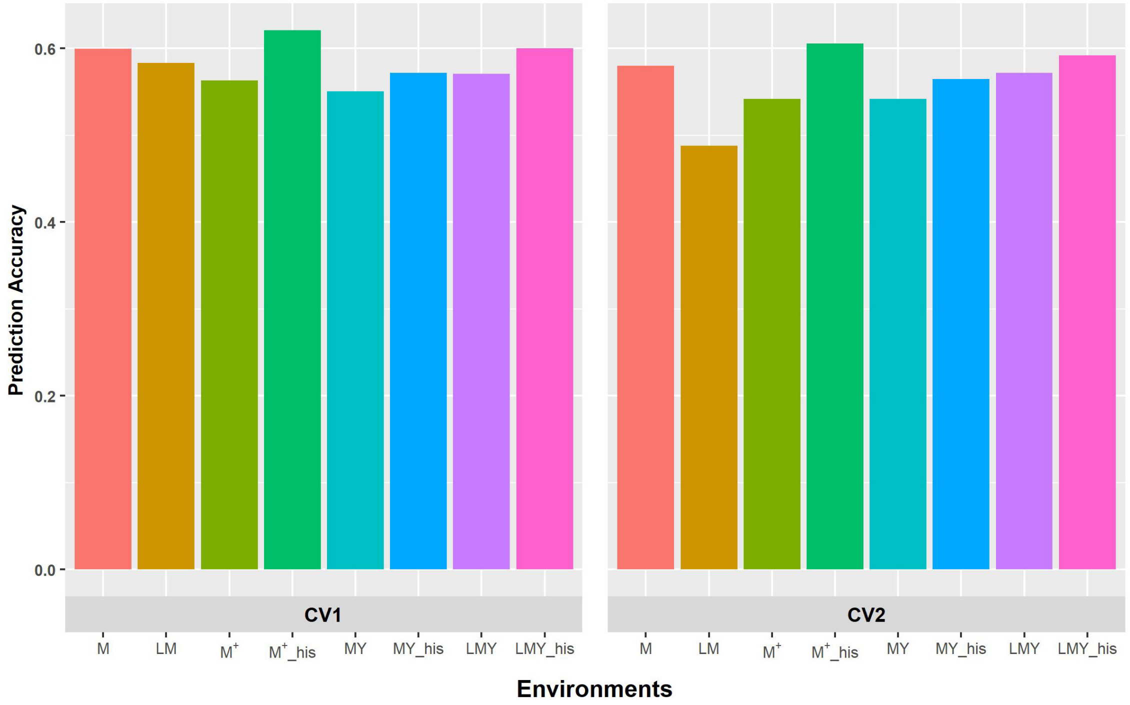


FIGURE 2 | Predictive ability of the factor analytic model for the cross-validation schemes (CV1 and CV2) in WW environments/management. LM and M represent prediction accuracy obtained when covariance was modeled across environments and managements, respectively, for within-year prediction. LMY represents classification of environment as location by management by year, MY and M^+ represent the broad classification of the management across years as WW and WS, and explicit definition of the management across years as WW 2017 and 2018 and WS 2017 and 2018. LMY, MY and M^+ used all available historical data. The suffix "his" represents prediction accuracy obtained with optimized historical data using the Avg_GRM genetic optimization criterion.

Schopp et al., 2017; Atanda et al., 2020) also indicate that use of closely related multiple bi-parental populations as a training set result in improved prediction accuracy. Using diverse populations, one would expect the differences in marker-quantitative trait loci linkage phase across bi-parental populations would result in a lower signal to noise ratio, but that does not appear to be the case in this dataset where several populations share a common parent. The small size of the bi-parental population used in this study might affect the prediction accuracy of CV2. Borrowing of information across environments was the basis for the improved prediction accuracy using sparse testing compared to test-half-predict-half (Atanda et al., 2020), thus, a strategy that optimizes coverage of the genetic space of the genotypes across environments should result in higher predictive ability.

The FA is a parsimonious model for fitting a relatively high number of environments in multi-environment trials utilizing latent factors which give rise to correlations between environments to capture the complexity of covariances among many environments (Burgueño et al., 2012; Oakey et al., 2016; Smith and Cullis, 2018; Tolhurst et al., 2019). However, with few environments and a large dataset to estimate all model parameters, the superiority of the FA model over the US model will likely depend on the ability of the FA model to adequately represent the underlying covariance structure between environments in the dataset (Piepho, 1998; Kelly et al., 2007; Meyer, 2009; So and Edwards, 2009; Ward et al., 2019). While this study looked at relatively few environments, the limitations of the US model became apparent in the multi-year dataset with six environments defined. Under this scenario, US model was sensitive to the training set used and did not consistently converge, suggesting that the utility of US model will diminish rapidly as the number of environments increase. Given reliable convergence and similar performance with a small number of environments, the FA appears to be a more robust approach for modeling sparse testing implementations in the CIMMYT Maize program.

In practice, the CIMMYT tropical maize breeding program advances lines to multi-location, multi-tester yield trials based on relative performance within or across managements (WW and WS), the observed improvement in prediction accuracy when environments were grouped into managements suggests that categorizing the environments into management did not sacrifice information on GEI. Assigning environments/locations into groups using prior information, such as management, as is the case in this study, can serve as a complexity reduction strategy for reducing the number of model parameters, providing a more parsimonious approach for modeling GEI. However, stage 1 yield testing is typified by a small number of environments, which is a limitation to the generalization of the results of this study across different phases of yield testing, in particular with a large number of environments. However, similar to the strategy employed in this study, using multi-environment data, Lado et al. (2016) grouped 35 environments into three mega environments using the additive main and multiplicative interactive (AMMI) model (Zobel et al., 1988), and GS was performed within the mega environments.

Augmenting a given full-sib training set with an optimized set of 300 individuals from historical data using the Avg_GRM genetic optimization algorithm improved prediction accuracy compared to using all available historical records. The similar genetic covariance between managements, heritability, and prediction accuracy obtained when historical data is used to complement the full-sib training set, suggests that an increase in the training set size using historical data results in more stable estimates of model parameters when compared to using only the full-sib records as the training set. The results from this study corroborate our earlier study (Atanda et al., 2020) indicating that the use of genetic optimization criteria to select individuals genetically connected to the breeding population to serve as a training population results in improved prediction accuracy. This further illustrates the importance of genetic relationships between training and breeding populations and indicates that any GS approach carefully consider which historical records are included for training of genomic prediction models. Furthermore, these results suggest that, when genomic information is available breeders should consider utilizing multi-year information for advancement decisions. This could not only improve advancement decisions but could enable earlier recycling of material to reduce generation intervals.

CONCLUSION

Given the similar prediction accuracies obtained in CV1 and CV2, decisions on which sparse testing experimental design will likely depend on cost and ease of implementation. While the prediction accuracy for the cross-validation schemes is equivalent, CV2 has an intuitive appeal in that all bi-parental populations have representation across environments, which would allow efficient use of information across environments and would be ideal for building a robust historical dataset. Further, the CV2 can be extended to resource demanding multi-environment, multi-tester advanced yield testing stages to save resources. In this study, grouping similar environments to model GEI information reduced computational challenges and achieved superior prediction accuracy. In general, including historical information in trial advancement decisions improved prediction accuracy, suggesting that the use of historical information in routine advancement decisions could improve accuracy. Furthermore, selecting historical information based on genetic connectedness with the breeding population proved more effective than including all historical information.

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

MO, KR, and SA conceptualized the study. SA analyzed, interpreted the result, and drafted the manuscript. YB coordinated the field experiments. MG and KD were responsible for phenotypic and genotyping data management. JB, JC, RR, DD, PB, PT, ED, GO, and other authors contributed to the editing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The study was funded by the Bill and Melinda Gates Foundation, and the United States Agency for International Development (USAID) through the Stress Tolerant Maize for Africa (STMA, # OPP1134248), the CGIAR Research Program MAIZE, and the Genomic and Open-source Breeding Informatics Initiative (Grant # OPP1093167). The CGIAR Research Program MAIZE receives W1 and W2 support from the Governments of Australia, Belgium, Canada, China, France, India, Japan, South Korea, Mexico, Netherlands, New Zealand, Norway, Sweden, Switzerland, United Kingdom, United States, and the World Bank. The lead author acknowledges the joint Ph.D. fellowship support received from the CIMMYT Global Maize Program, under CGIAR Research Program MAIZE and the Robbins Lab at Cornell University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.658978/full#supplementary-material

Supplementary Figure 1 | Illustration of cross-validation scheme 1. Each box represents a population; the black color depicts populations masked in an environment and the white color represents populations used for model training to predict the genomic estimated breeding value of masked populations in each environment. Environments (1, 2, and 3) represent Kiboko optimal, Kakamega optimal and Kiboko drought.

Supplementary Figure 2 | Illustration of cross-validation scheme 2. Each box represents a population; the white color depicts individuals within a bi-parental population selected based on their CDmean value to predict the genomic estimated breeding value of masked individuals (black color). Environments (1, 2, and 3) represent Kiboko optimal, Kakamega optimal and Kiboko drought.

Supplementary Table 1 | Masking of subset of a bi-parental population in CV2 across environments.

Supplementary Table 2 | Prediction accuracy for factor analytic models using $m = 1$ and 2 and the unstructured model, depicting model accuracy using either $m = 1$ or 2 as number of environments increase.

Supplementary Table 3 | Eigen analysis of factor analytic matrix, showing variation explained by the latent variables.

REFERENCES

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.

Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., et al. (2020). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theor. Appl. Genet.* 134, 279–294. doi: 10.1007/s00122-020-03696-9

Beyene, Y., Gowda, M., Olsen, M., Robbins, K. R., Pérez-Rodríguez, P., Alvarado, G., et al. (2019). Empirical comparison of tropical maize hybrids selected through genomic and phenotypic selections. *Front. Plant Sci.* 10:1502. doi: 10.3389/fpls.2019.01502

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

Buckler, E. S., Ilut, D. C., Wang, X., Kretzschmar, T., Gore, M., and Mitchell, S. E. (2016). rAmpSeq: using repetitive sequences for robust genotyping. *bioRxiv* [Preprint]. doi: 10.1101/096628

Burgueño, J., Crossa, J., Cornelius, P. L., and Yang, R.-C. (2008). Using factor analytic models for joining environments and genotypes without crossover genotype × environment interaction. *Crop Sci.* 48, 1291–1305. doi: 10.2135/cropsci2007.11.0632

Burgueño, J., Crossa, J., Cotes, J. M., Vicente, F. S., 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 × environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52, 707–719. doi: 10.2135/cropsci2011.06.0299

Crossa, J., Yang, R.-C., and Cornelius, P. L. (2004). Studying crossover genotype × environment interaction using linear-bilinear models and mixed models. *J. Agric. Biol. Environ. Stat.* 9, 362–380. doi: 10.1198/108571104x4423

Endelman, J. B., Atlin, G. N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M. E., et al. (2014). Optimal design of preliminary yield trials with genome-wide markers. *Crop Sci.* 54, 48–59. doi: 10.2135/cropsci2013.03.0154

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51, 1440–1450. doi: 10.2307/2533274

Heslot, N., and Feoktistov, V. (2020). Optimization of selective phenotyping and population design for genomic prediction. *JABES* 25, 579–600. doi: 10.1007/s13253-020-00415-1

Kelly, A. M., Smith, A. B., Eccleston, J. A., and Cullis, B. R. (2007). The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. *Crop Sci.* 47, 1063–1070. doi: 10.2135/cropsci2006.08.0540

Lado, B., Barrios, P. G., Quincke, M., Silva, P., and Gutiérrez, L. (2016). Modeling genotype × environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Sci.* 56, 2165–2179. doi: 10.2135/cropsci2015.04.0207

Lehermeier, C., Krämer, N., Bauer, E., Bauland, C., Camisan, C., Campo, L., et al. (2014). Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. *Genetics* 198, 3–16. doi: 10.1534/genetics.114.161943

Meyer, K. (2009). Factor-analytic models for genotype × environment type problems and structured covariance matrices. *Genet. Sel. Evol.* 41:21. doi: 10.1186/1297-9686-41-21

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

Piepho, H.-P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theor. Appl. Genet.* 97, 195–201. doi: 10.1007/s001220050885

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

Santantonio, N., Atanda, S. A., Beyene, Y., Varshney, R. K., Olsen, M., Jones, E., et al. (2020). Strategies for effective use of genomic information in crop breeding programs serving Africa and South Asia. *Front. Plant Sci.* 11:353. doi: 10.3389/fpls.2020.00353

Schopp, P., Müller, D., Wientjes, Y. C. J., and Melchinger, A. E. (2017). Genomic prediction within and across Biparental families: means and variances of prediction accuracy and usefulness of deterministic equations. *G3 Genes Genomes Genet.* 7, 3571–3586. doi: 10.1534/g3.117.300076

Smith, A., Cullis, B., and 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

Smith, A. B., and Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* 214:143. doi: 10.1007/s10681-018-2220-5

So, Y.-S., and Edwards, J. (2009). A comparison of mixed-model analyses of the iowa crop performance test for corn. *Crop Sci.* 49, 1593–1601. doi: 10.2135/cropsci2008.09.0574

Thompson, R., Cullis, B., Smith, A., and Gilmour, A. (2003). A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. *Aust. N. Z. J. Stat.* 45, 445–459. doi: 10.1111/1467-842x.00297

Tolhurst, D. J., Mathews, K. L., Smith, A. B., and Cullis, B. R. (2019). Genomic selection in multi—environment plant breeding trials using a factor analytic linear mixed model. *J. Anim. Breed. Genet.* 136, 279–300. doi: 10.1111/jbg.12404

Ward, B. P., Brown-Guedira, G., Tyagi, P., Kolb, F. L., Van Sanford, D. A., Sneller, C. H., et al. (2019). Multienvironment and multitrait genomic selection models in unbalanced early-generation wheat yield trials. *Crop Sci.* 59, 491–507. doi: 10.2135/cropsci2018.03.0189

Zobel, R. W., Wright, M. J., and Gauch, H. G. (1988). Statistical analysis of a yield trial. *Agron. J.* 80, 388–393. doi: 10.2134/agronj1988.00021962008000030002x

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 several of the authors JC, YB, MG, JB, and PB.

Citation: Atanda SA, Olsen M, Crossa J, Burgueño J, Rincent R, Dzidzienyo D, Beyene Y, Gowda M, Dreher K, Boddupalli PM, Tongoona P, Danquah EY, Olaoye G and Robbins KR (2021) Scalable Sparse Testing Genomic Selection Strategy for Early Yield Testing Stage. Front. Plant Sci. 12:658978. doi: 10.3389/fpls.2021.658978

frontiers in Plant Science

OPINION
published: 01 July 2021
doi: 10.3389/fpls.2021.674036

Opportunities and Challenges of Predictive Approaches for Harnessing the Potential of Genetic Resources

Johannes W. R. Martini, Terence L. Molnar, José Crossa*, Sarah J. Hearne* and Kevin V. Pixley**

International Maize and Wheat Improvement Center, Texcoco, Mexico

Keywords: genetic resources, germplasm bank collections, predictive breeding, genomic selection, pre-breeding

OPEN ACCESS

Edited by:
Athanassios Molassiotis,
Aristotle University of
Thessaloniki, Greece

Reviewed by:
Alexios Polidoros,
Aristotle University of
Thessaloniki, Greece

****Correspondence:***
Terence L. Molnar
t.molnar@cgiar.org
Sarah J. Hearne
s.hearne@cgiar.org
José Crossa
J.Crossa@cgiar.org
Kevin V. Pixley
k.pixley@cgiar.org

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *28 February 2021*
Accepted: *17 May 2021*
Published: *01 July 2021*

Citation:
Martini JWR, Molnar TL, Crossa J, Hearne SJ and Pixley KV (2021) Opportunities and Challenges of Predictive Approaches for Harnessing the Potential of Genetic Resources. Front. Plant Sci. 12:674036. doi: 10.3389/fpls.2021.674036

INTRODUCTION

Favorable variation from genetic resources is anticipated to play a key role in the adaptation of crops to the increasingly unfavorable production conditions resulting from climate change (FAO, 2015). Weather extremes lead to more frequent occurrences of abiotic stress and facilitate the emergence and spread of diseases. While there is no doubt that alleles and haplotypes offered by accessions from germplasm banks are of enormous value, the integration of beneficial alleles into elite material poses three major challenges:

1. the identification of promising germplasm bank accessions,
2. the separation of beneficial major effect alleles from undesired linkage drag,
3. the repackaging of polygenic variation into elite and adapted materials.

Identifying promising germplasm bank accessions, which may offer single alleles with major effects or beneficial quantitative variation, often resembles looking for a needle in a haystack. In practice, it is almost never possible to phenotype a large portion of the available germplasm due to high costs, challenges with adaptation, restricted facility resources and time pressure. An informed prescreening of the available accessions will be necessary.

Moreover, when accessions with putative alleles for desired traits are identified, the mission is not yet accomplished, since the beneficial variation must be integrated into elite germplasm. In the case of a simple genetic architecture such as an identified major effect gene, the novel allele can be introgressed by marker assisted backcrossing (MABC) or can be approached by gene editing. However, preceding discovery research is required to identify the genetic variation associated with the phenotypic variation. In particular, gene editing requires very precise information on the causative variation. The availability of a trait-associated marker, which may be sufficient for an application in MABC, may be insufficient for a gene editing approach. This research is resource and time consuming and carries the inherent risk of unsuccessful validation experiments due an altered effect of the allele when in combination with the genetic background of elite material.

When dealing with quantitative variation, dedicated mapping experiments are not required. However, it is more difficult to bring quantitative variation into an elite background and have a product acceptable to breeders. Landraces carry many deleterious and inferior alleles which can quickly disrupt the positive linkage blocks painstakingly constructed by breeders over decades. Diminished agronomic performance makes the breeding community reluctant to include such germplasm in their elite breeding programs.

Prediction approaches can help the effective use of genetic resources in two ways. First, predictions can identify the most promising candidate accessions for a certain trait, thus restricting the number of accessions to evaluate in experiments (Yu et al., 2016). Second, predictions can accelerate the pre-breeding (or "germplasm enhancement") process by helping to target the desired alleles for transfer to an elite germplasm background, saving resources and time.

In this commentary, we summarize some activities related to predictive breeding in the context of genetic resources conducted at the International Maize and Wheat Improvement Center (CIMMYT). We then discuss differences between predictive breeding approaches for genetic resources and genomic selection for elite breeding programs. We propose that research on predictive methods for genetic resources should explore approaches which are "enriched" by external information; for example, knowledge of molecular biological mechanisms, or accession "passport" data that provides information on the environmental conditions in which the accession was originally cultivated. Passport data comprising latitude, longitude, and altitude are fundamental initial information for each accession stored in the bank. The inclusion of external information may increase the power of predictive breeding approaches, especially in the context of harnessing genetic resources.

PREDICTIVE BREEDING FOR GENETIC RESOURCES AT CIMMYT

Genotyping of Accessions of CIMMYT's Germplasm Bank

CIMMYT has genotyped most of its maize and wheat collections as part of the Seeds of Discovery Project (SEED). For maize, more than 98% of the CIMMYT and IITA (International Institute of Tropical Agriculture) maize collection have been genotyped. For wheat, 37 and 66%, respectively, of the CIMMYT and ICARDA (International Center for Agricultural Research in the Dry Areas) wheat collection have been genotyped (Sansaloni et al., 2020). The smaller percentages for wheat, compared to maize, are due to the larger size and differing composition of the combined collections. CIMMYT's germplasm bank has ∼28,000 maize, but more than 140,000 wheat accessions. The available genotypic data provides a solid foundation for prediction approaches for screening the collections more systematically.

Genetic Resources for Breeding for Maize Lethal Necrosis Resistance

A recent example of the successful use of germplasm bank material in response to an emerging threat was the development of germplasm tolerant to Maize Lethal Necrosis (MLN). Thirteen out of 1000 screened landraces were identified as showing low susceptibility to Maize Chlorotic Mottle Virus (MCMV), the major causal component of MLN disease (for a review on CIMMYT's activities related to MLN, see Boddupalli et al., 2020). The pre-screening in this study was based on geographical distribution, racial structure, and genomic distance data calculated as described in Franco-Duran et al. (2019). The performance of the developed inbred lines in hybrid combinations is currently tested, in particular under MLN pressure.

Prediction of Wheat Landraces Accessions

For wheat, Crossa et al. (2016) considered genomic prediction on a large set of Mexican (∼8,400) and Iranian (∼2,400) bank accessions for several traits including thousand-kernel weight, grain hardness, grain protein, and plant height. The predictive abilities obtained were mostly between 0.39 and 0.68, when using 20% of the data as training set (Crossa et al., 2016, Table 2). An exception was plant height for the Iranian landraces, which showed a predictive ability of only 0.17. These results indicated that genomic prediction has a potential for (1) fast screening of the whole GB for different traits, and (2) a rapid and efficient pre-breeding method for introgression useful alleles (and haplotypes) into advance breeding lines while not eroding genetic diversity.

Association Studies With Environmental Covariates as Phenotype

A novel approach to use "passport" data of accessions is "environmental genome-wide association studies" (environmental GWAS or EnvGWAS). This approach treats environmental variables of the sites where accessions were collected as phenotypes, and combines this information with genotypic data for the accessions in an association study. The objective is to identify genetic variation which is associated with the adaptation to certain environmental conditions (Lasky et al., 2015; Romero Navarro et al., 2017; Gates et al., 2019). Though this approach conceptually could lead to high false positive rates due spatial distribution impacting phylogeny and environmental variables, this problem can be controlled, as in standard GWAS, by introducing a random polygenetic effect with the genomic relationship as covariance (Yang et al., 2014). Proof of concept work in drought using collection site precipitation data has demonstrated the power of EnvGWAS to detect variants of potential interest in maize landraces (Gates et al., 2019). Validation of the role of these variants in drought response, conducted through independent in silico analysis of transcriptome data and analysis of phenotypic data, has confirmed the value of EnvGWAS for identifying variants and in turn landraces containing variants for further analysis and use in breeding.

DIFFERENCES BETWEEN PREDICTIVE APPROACHES IN THE CONTEXT OF GENETIC RESOURCES AND GENOMIC SELECTION IN AN ELITE GERMPLASM POOL

Although we have witnessed promising results for both maize and wheat, we see conceptual limitations of standard genomic prediction methods when looking for ***novel*** beneficial alleles. Standard prediction approaches predict from a training to a prediction set and can only predict the effect of new combinations of already known segments (Meuwissen et al., 2001). Indeed, this is also the major application of genomic selection in an elite breeding pipeline where most alleles have already been sampled in different combinations. In this situation, one aims at ***recombining*** the positive alleles which have already been observed. This differs fundamentally from a prediction where the objective is to find novel beneficial variation. Therefore, when screening for novel diversity which is not present in the training set, we see the main value of the prediction in its indirect information: a strong accumulation of beneficial alleles that are already present in the training set may be a result of selection pressure in the accession's history. Thus, the probability of finding additional novel alleles for the trait of interest may be increased.

Approaches to Incorporate External Information

To address this conceptual discrepancy between the nature of statistical prediction and the objective of predicting novel diversity, and to go beyond the indirect information provided by a standard genomic selection as described above, we believe different sources of information need to be combined with genotypic data. Examples may be passport data as in EnvGWAS, gene annotation data (Gao et al., 2017), data on biochemical pathways or other data on biological mechanisms, or general (quantitative genetics) knowledge on -for instance- ratios of variances (Hem et al., 2021). Such approaches have already been followed in general genomic prediction literature, but we think that they will especially unfold their potential in the context of genetic resources.

A promising approach to follow for a broader range of traits is the comparison of structure, function and point of action of gene products. Given that some genes involved in the variation of stress resilience are known, bioinformatics tools can identify related genes whose gene products are of similar structure, have a similar predicted function or are relevant in the same biochemical pathways as the known genes. Genomic data can then be used to identify novel variation in the regions around these newly identified genes. Approaches of this kind have been used, for instance as resistance gene enrichment sequencing targeting certain protein motifs to identify resistances to biotic stresses (Jupe et al., 2013; Zhang et al., 2020), and have produced impressive results. However, such a strategy focuses on major gene effects and it remains to be seen whether they can be transferred to a quantitative trait such as yield under abiotic stress.

For the identification of germplasm bank accessions providing beneficial alleles for quantitative traits, we see the accession passport data as central information. This data cannot only be used to identify major effects in an association study, but can also be used in a genomic prediction approach. Here, a genomic relationship matrix of the accessions can be used to predict the environmental variables of the collection sites as "quantitative trait." This "environmental genomic prediction" (EnvGP) then employs the environmental data as a phenotype in the training panel to predict materials of higher value for "hands-on" evaluation. Considering the polygenic nature of many traits of interest, we are currently assessing the potential of EnvGP together with other paradigms such as crop modeling to leverage genetic resources for germplasm development.

As an example addressing the process of repackaging of polygenic variation into elite and adapted materials, we cite Origin Specific Genomic Selection (OSGS; Yang et al., 2020). Here, the additional information used in the prediction is only the knowledge from which parent the alleles are derived. However, this add-on allows a partitioned form of genomic selection which facilitates a more targeted management of the introgression of novel beneficial variation during the introgression process. The genetic value is split into the contribution of the elite parent and the contribution of the "exotic" parent. Having both parts separated, the approach aims at avoiding a systematic selection against exotic alleles due to the higher genetic value of elite material although a certain fraction of exotic alleles may be beneficial. Validation of this approach using simulation and application in existing barley and maize datasets suggests potential for use in polygenic trait introgression in bi- and potentially multi-parental populations.

CONCLUSION

Germplasm bank accessions can be considered as crop "genetic insurance" for the genetic adaptation to increased abiotic and biotic stresses, in particular caused by climate change. As for other fields, "big data," here describing the germplasm bank collections, needs innovative approaches for "data mining," to identify and harness useful variation, and unleash its potential. We see a conceptual key in combining statistical prediction methods with additional data other than genotypes and phenotypes. Approaches of this type have been followed in genomic prediction literature, but we consider them as particularly promising when applied in the context of harnessing genetic resources. The type of data to use, and how to use it provide a large playground for the exploration of creative approaches.

AUTHOR CONTRIBUTIONS

JM wrote the first draft and managed the edits from other authors. All authors discussed and outlined the content of the opinion and approved the published version for publication.

ACKNOWLEDGMENTS

The authors would like to thank funders who have supported work contributing to the article presented. Specifically we wish to thank the Ministry of Agriculture in Mexico-Secretaría de Agricultura y Desarrollo Rural Secretaría de Agricultura y Desarrollo Rural (SADER) through the Sustainable Modernization of Traditional Agriculture (MasAgro) project, the MAIZE CGIAR Research Program (CRP MAIZE), the WHEAT CGIAR Research Program (CRP WHEAT), and the Bill and Melinda Gates Foundation. The MAIZE CRP receives funding from the governments of Australia, Belgium, Canada, China, France, India, Japan, Korea, Mexico, the Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, the United States, and from the World Bank. The WHEAT CRP receives funding from the governments of Australia, Belgium, Canada, France, India, Japan, the Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, the United States, and from the World Bank.

REFERENCES

Boddupalli, P., Suresh, L. M., Mwatuni, F., Beyene, Y., Makumbi, D., Gowda, M., et al. (2020). Maize lethal necrosis (MLN): Efforts toward containing the spread and impact of a devastating transboundary disease in sub-Saharan Africa. *Virus Res.* 282:197943. doi: 10.1016/j.virusres.2020.197943

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* 6, 1819–1834. doi: 10.1534/g3.116.029637

FAO (2015). *Coping With Climate Change – The Roles of Genetic Resources for Food and Agriculture*. Rome.

Franco-Duran, J., Crossa, J., Chen, J., and Hearne, S. J. (2019). The impact of sample selection strategies on genetic diversity and representativeness in germplasm bank collections. *BMC Plant Biol.* 19, 1–17. doi: 10.1186/s12870-019-2142-y

Gao, N., Martini, J. W. R., Zhang, Z., Yuan, X., Zhang, H., Simianer, H., et al. (2017). Incorporating gene annotation into genomic prediction of complex phenotypes. *Genetics* 207, 489–501. doi: 10.1534/genetics.117.300198

Gates, D. J., Runcie, D., Janzen, G. M., Navarro, A. R., Willcox, M., Sonder, K., et al. (2019). Single-gene resolution of locally adaptive genetic variation in Mexican maize. *BioRxiv [Preprint]*. doi: 10.1101/706739

Hem, I. G., Selle, M. L., Gorjanc, G., Fuglstad, G. A., and Riebler, A. (2021). Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge. *Genetics* 217:iyab002. doi: 10.1093/genetics/iyab002

Jupe, F., Witek, K., Verweij, W., Sliwka, J., Pritchard, L., Etherington, G. J., et al. (2013). Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant J.* 76, 530–544. doi: 10.1111/tpj.12307

Lasky, J. R., Upadhyaya, H. D., Ramu, P., Deshpande, S., Hash, C. T., Bonnette, J., et al. (2015). Genome-environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.* 1:e1400218. doi: 10.1126/sciadv.1400218

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

Romero Navarro, J. A., Willcox, M., Burgueño, J., Romay, C., Swarts, K., Trachsel, S., et al. (2017). A study of allelic diversity underlying flowering-time. *Nat. Genet.* 49:476–80. doi: 10.1038/ng.3784

Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., et al. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-020-18404-w

Yang, C. J., Sharma, R., Gorjanc, G., Hearne, S., Powell, W., and Mackay, I. (2020). Origin specific genomic selection: a simple process to optimize the favorable contribution of parents to progeny. *G3* 10, 2445–2455. doi: 10.1534/g3.120.401132

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet*. 46, 100–106. doi: 10.1038/ng.2876

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, 1–7. doi: 10.1038/nplants.2016.150

Zhang, J., Zhang, P., Dodds, P., and Lagudah, E. (2020). How Target-sequence Enrichment and Sequencing (TEnSeq) pipelines have catalysed resistance gene cloning in the wheat-rust pathosystem. *Front. Plant Sci.* 11:678. doi: 10.3389/fpls.2020.00678

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.

SYSTEMATIC REVIEW
published: 01 July 2021
doi: 10.3389/fpls.2021.658267

Optimizing Genomic-Enabled Prediction in Small-Scale Maize Hybrid Breeding Programs: A Roadmap Review

Roberto Fritsche-Neto [1]*, *Giovanni Galli* [1], *Karina Lima Reis Borges* [1], *Germano Costa-Neto* [1], *Filipe Couto Alves* [2], *Felipe Sabadin* [1], *Danilo Hottis Lyra* [3], *Pedro Patric Pinho Morais* [4], *Luciano Rogério Braatz de Andrade* [5], *Italo Granato* [6] and *Jose Crossa* [7,8]

[1] Laboratory of Allogamous Plant Breeding, Genetics Department, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, Brazil, [2] Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, United States, [3] Department of Computational and Analytical Sciences, Rothamsted Research, Harpenden, United Kingdom, [4] Department of Agronomy, Federal University of Viçosa, Viçosa, Brazil, [5] Brazilian Agricultural Research Corporation (EMBRAPA), Cassava and Fruits, Cruz das Almas, Brazil, [6] Laboratoire d'Ecophysiologie des Plantes sous Stress Environnementaux (LEPSE), Institut National de la Recherche Agronomique (INRA), Univ. Montpellier, SupAgro, Montpellier, France, [7] Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Carretera México - Veracruz, Texcoco, Mexico, [8] Colegio de Posgraduado, Montecillo, Mexico

OPEN ACCESS

Edited by:
Valentin Wimmer,
KWS Saat, Germany

Reviewed by:
Bernardo Ordas,
Misión Biológica de Galicia, Spanish National Research Council, Spain
Nicholas Santantonio,
Virginia Tech, United States

***Correspondence:**
Roberto Fritsche-Neto
roberto.neto@usp.br

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *25 January 2021*
Accepted: *10 May 2021*
Published: *01 July 2021*

Citation:
Fritsche-Neto R, Galli G, Borges KLR, Costa-Neto G, Alves FC, Sabadin F, Lyra DH, Morais PPP, Braatz de Andrade LR, Granato I and Crossa J (2021) Optimizing Genomic-Enabled Prediction in Small-Scale Maize Hybrid Breeding Programs: A Roadmap Review. Front. Plant Sci. 12:658267. doi: 10.3389/fpls.2021.658267

The usefulness of genomic prediction (GP) for many animal and plant breeding programs has been highlighted for many studies in the last 20 years. In maize breeding programs, mostly dedicated to delivering more highly adapted and productive hybrids, this approach has been proved successful for both large- and small-scale breeding programs worldwide. Here, we present some of the strategies developed to improve the accuracy of GP in tropical maize, focusing on its use under low budget and small-scale conditions achieved for most of the hybrid breeding programs in developing countries. We highlight the most important outcomes obtained by the University of São Paulo (USP, Brazil) and how they can improve the accuracy of prediction in tropical maize hybrids. Our roadmap starts with the efforts for germplasm characterization, moving on to the practices for mating design, and the selection of the genotypes that are used to compose the training population in field phenotyping trials. Factors including population structure and the importance of non-additive effects (dominance and epistasis) controlling the desired trait are also outlined. Finally, we explain how the source of the molecular markers, environmental, and the modeling of genotype–environment interaction can affect the accuracy of GP. Results of 7 years of research in a public maize hybrid breeding program under tropical conditions are discussed, and with the great advances that have been made, we find that what is yet to come is exciting. The use of open-source software for the quality control of molecular markers, implementing GP, and envirotyping pipelines may reduce costs in an efficient computational manner. We conclude that exploring new models/tools using high-throughput phenotyping data along with large-scale envirotyping may bring more resolution and realism when predicting genotype performances. Despite the initial costs, mostly for genotyping, the GP platforms in combination with these other data sources can be a cost-effective approach for predicting the performance of maize hybrids for a large set of growing conditions.

Keywords: accuracy, quantitative genomics, R packages, genomic selection, breeding schemes

INTRODUCTION

Hybrid breeding programs are usually based on pureline methods, including the development of inbreeding lines by self-pollination or double-haploids, followed by progeny evaluation across heterotic pools (Hallauer et al., 2010). The great challenge of this approach is to adequately test the performance in all possible combinations of lines in crosses (Bernardo, 1994). In this context, we have conducted several studies indicating the usefulness of genomic prediction (GP, Meuwissen et al., 2001). Since the first studies of GP in maize (Bernard and Yu, 2007), several applications have been made to improve different steps of maize breeding, such as the selection under diverse breeding populations (Lorenzana and Bernardo, 2009; Lehermeier et al., 2014), the rapid cycle improvement of parental inbreeds (Zhang et al., 2017; Cui et al., 2020; Das et al., 2020), the prediction of double-haploid lines (e.g., Cooper et al., 2016; Messina et al., 2018), and the prediction of the performance of single-crosses for single or multi-environment conditions (Windhausen et al., 2012; Dias et al., 2018; Alves et al., 2019; Millet et al., 2019; Costa-Neto et al., 2020; Rogers et al., 2021).

Here we focused our review efforts on the GP of maize hybrids, particularly in the single-crosses of F_1. From the last 10 years of research in this field, several research groups pointed to affect the main factors that drastically affect the accuracy of GP for hybrid prediction, such as (1) the genetic design and the genotypes used to form the training population; (2) the presence of a population structure; (3) the importance of non-additive effects controlling the desired characteristic; (4) the source of molecular markers used; and (5) the genotype × environment (G × E) interaction over contrasting environments. Therefore, this review aims to describe the most important outcomes in this field and report our research experience in a small-scale low budget breeding program under tropical growing conditions.

ROADMAP FOR IMPLEMENTING GP IN HYBRID BREEDING PROGRAMS

Here, we highlighted the most important outcomes obtained by the Allogamous Breeding Laboratory of the University of São Paulo (USP, Brazil) and some other groups in testing GP for predicting maize hybrids. We present our review as a roadmap for small-scale and low-budget breeding programs due to the fact that most of our research is focused on optimizing GP in order to find the best training sets (TS), to select the best genotyping pipelines, and to choose the best multi-environment structures to predict scenarios of genotype × environment interaction. Our roadmap began with the efforts for germplasm characterization, which involves both molecular and phenotypic characterization. Before this step, it is necessary to develop the inbred lines during successive cycles of self-crossing. For most breeding programs, this step may involve the use of double haploid technology. After seed replication, field trials must be well-conducted, following certain management practices, which may evolve, for example, the use of optimum vs. nitrogen-limited conditions. A good statistical analysis and phenotype correction are important steps that impact further genomic analysis (Galli et al., 2018).

Then, after the characterization of lines, we focused on maize hybrid predictions. The second step of the roadmap considers schemes for mating design and choosing the genotypes used to compose the training population in field phenotyping trials. Factors including population structure and the importance of non-additive effects (dominance and epistasis) controlling the desired trait are also outlined. Finally, we present how the source of the molecular markers, environment, and the modeling of genotype × environment interaction can affect the accuracy of GP. We also point out that the use of dominance effects in GP is crucial to deliver accurate predictions of maize hybrids. Results of 7 years of research in our public maize hybrid breeding program under tropical conditions are discussed, and with the great advances that have been made, we find that what is yet to come is exciting. In the end, we revised some fields of work and the lessons we learned from both our experience and the results from other groups.

GERMPLASM CHARACTERIZATION

Tropical Germplasm of USP, Brazil

The very first step on our scientific road was to carry out germplasm characterization on the newly acquired inbred lines (Sant'Ana et al., 2020). Genomic diversity and population structure of germplasm (e.g., heterotic groups) are widely known to accelerate genetic gains in breeding programs. This structure and diversity are allocated to two major groups, such as temperate and tropical germplasm in tropical maize. While tropical maize germplasm has a greater genetic diversity, the temperate one has more pronounced heterotic patterns (Mir et al., 2013). Moreover, tropical maize germplasm lacks information on its genetic diversity regarding low-nitrogen (N) stress (De Andrade et al., 2016; Torres et al., 2018). In this context, in order to analyze the population structure of tropical maize accessions and identify genomic regions related to low-N tolerance, an initial set of 64 inbred lines was evaluated under ideal and low N availability conditions. The lines were genotyped using 417, 112 Single Nucleotide Polymorphism (SNP) markers from the Affymetrix platform described above. The grouping, based on the

Nitrogen Acquisition Efficiency (NAE) values, classified the lines into two phenotypic groups, the first of which was composed of genotypes with high NAE (called H_NAE group) and the second of genotypes with low NAE (called L_NAE group). The groups H_NAE and L_NAE presented mean NAE values of 3,304 and 1,644, respectively (Sant'Ana et al., 2020). The population structure analysis revealed a weak relationship between genetic and phenotypic diversities. Simultaneously, line pairs having a high NAE and a considerable genetic distance were identified.

In greater detail, we noticed that a set of 29 single nucleotide polymorphism (SNP) markers displayed a significant difference in the allelic frequencies (Fst > 0.2) between groups H_NAE and L_NAE. Pearson's correlation between NAE and the favorable alleles in this set of SNPs was 0.69. These SNPs can be useful for the marker-assisted selection (MAS) for low-N tolerance in maize breeding programs. The results of this study can assist maize breeders when identifying genotypes to be used in the development of low-N tolerance cultivars.

Using this information, we have chosen 49 lines to compose the genitor bank of our breeding programs. We carried out the first complete diallel, which outlined the heterotic groups and the first GP training population, both used in the studies described below.

Finding Population Structure in Hybrid Breeding Populations

Due to the considerable diversity, we then tried to identify whether the population structure within the dataset should be considered (Lyra et al., 2018). Population structure arises mainly due to geographical isolation and natural/artificial selections. Individuals are distributed into a few to several distinct subgroups that display different allele frequencies (**Figure 1A**). In a genome-wide association study (GWAS), individuals within the diversity panel present a specific phenotype of one or more lines that may generate misleading estimates on the linkage imbalance. As a result, whenever a phenotype is correlated with a subpopulation, this phenotype will probably show spurious associations. Although these associations are a major concern for GWAS, the use of highly structured subgroups in hybrid prediction could influence the achievement of reliable estimates of genomic estimated breeding values (GEBVs) for quantitative traits (Larièpe et al., 2017; Werner et al., 2020).

There are many ways to account for population structure in GP. Traditionally, the use of only a genomic relationship matrix is enough to predict phenotypes within breeding populations. However, when there is a strong structure (e.g., diverse panels), one strategy is to incorporate autovectors and admixture coefficients as covariates (fixed effects) in genomic models (**Figure 1A**). The use of principal components (PCs) in the genomic best linear unbiased prediction (GBLUP) method might result in a poorly positioned model because PCs enter both as fixed effects and implicitly, *via* the random effect (de Los Campos and Sorensen, 2014). Another option is to consider population structure in the cross-validation scheme, ensuring that each subpopulation is equally represented in the training and validation sets, consequently maximizing relatedness (Atanda et al., 2021). A third approach essentially divides the population into homogeneous (putative unstructured) subgroups (**Figure 1B**). When predictions are limited to specific subpopulations, the predictive ability is generally greater than predicting between subgroups or correcting for PS covariables (Guo et al., 2014). On the other hand, despite efforts to control the heterogeneity of marker effects among subpopulations (e.g., MG-GBLUP model, Lehermeier et al., 2015), dividing the population into subgroups may lead to a reduction in population size and a loss of diversity, thus reducing the predictive ability.

Tropical and subtropical maize genotypes are not as organized as temperate ones, which mean that more than two heterotic pools can be used in crosses. Equivalently, a diverse population of inbred lines can be crossed with testers representing different genetic origins. Thus, although only the effect of alleles and their interactions make up the genetic structures of hybrids, it is essential to find the structure patterns and understand how this information affects the predictions. In this sense, we investigated the effect of population structure in the GPs of simple crossbreeding considering two scenarios: (1) applying the traditional GBLUP and four methods of adjusting population structure in the whole group and (2) using homogeneous (A-GBLUP), within-group analysis (W-GBLUP), multi-group analysis (MG-GBLUP), and inter-group analysis (AC-GBLUP) in stratified groups (Lyra et al., 2018).

No advantages were found in the addition of population structure covariables to the prediction model based on the predictive ability. Thus, one explanation could be that the genomic relationship matrix has implicitly captured the genetic variation of population structure and hybrid mixing; another reason could be the similarity in the average performance of the characteristics in the subpopulation. Our second strategy was to divide the population into stratified groups. From our results, the predictive ability was significantly higher in A-GB and MG-GBLUP than W-GB for both characteristics, suggesting that considering the heterogeneity of the marker effects among subpopulations may be a promising strategy.

Our results suggest that the population structure problem for the GP can be efficient for highly structured (defined) populations but not for single hybrids. These results provided further knowledge about our germplasm and reassuring ways to perform GP.

DESIGN OF TRAINING POPULATIONS FOR GENOMIC PREDICTION

Finding the Best Mating Design for Training Populations

Post-hoc but relevant information about creating a training population is included in our realm of projects. We realized that the literature concerning GP in maize was quite vast, yet there was a significant shortage of studies on the best genetic design to build the training population.

Therefore, we handled a study to verify genomic selection accuracy to predict the performance of maize hybrids under different genetic designs (Fristche-Neto et al., 2018). Several

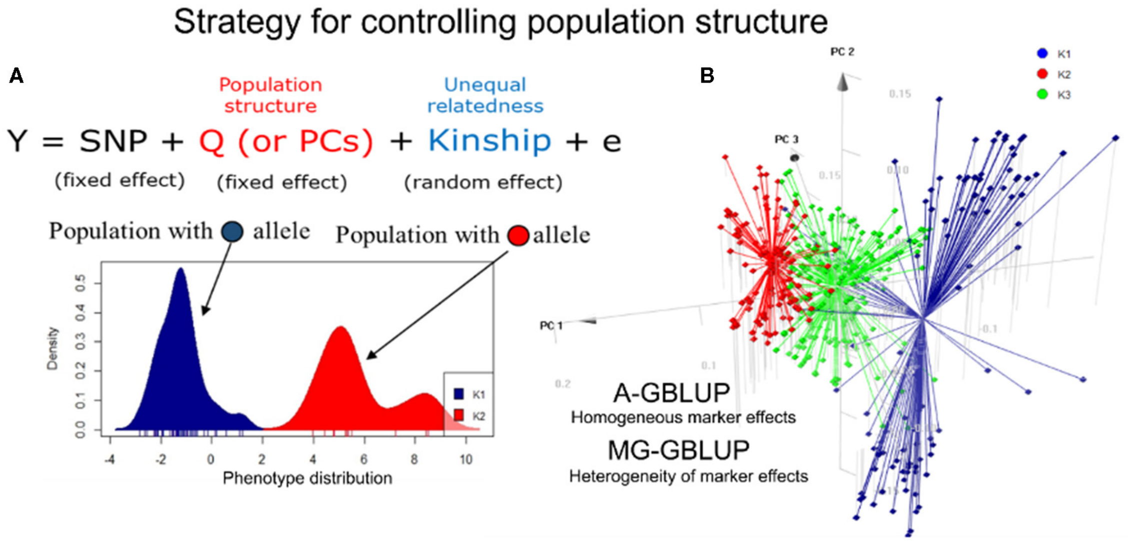


FIGURE 1 | Approaches to control the maize population structure. **(A)** A mixed linear model accounts for the covariates of population structure (fixed effect) and the genomic relationship matrix (kinship). An example of an allele-specific the population is shown in the graph. **(B)** 3D graph for the first three major components (PCs) using 452 simple tropical maize hybrids. Two stratification methods for the prediction of hybrids are shown in the panel. The first is a homogeneous group approach (A-GBLUP), which assumes constant marker effects between groups. The second is a multivariate approach (MG-GBLUP) that uses data from several groups and considers heterogeneity, with population-specific marker effects that can be correlated between subpopulations.

mating designs, such as Griffing's methods, partial diallel, North Carolina Design II (NCII), and test crossing (Hallauer et al., 2010) have been proposed. These methods have the following four main goals: (i) to provide information on the genetic control of the trait under investigation; (ii) to generate populations to be used as a basis for the selection and development of cultivars; (iii) to provide estimates of genetic gain; (iv) to obtain information to evaluate the genitors used in the breeding program, based on the general and combination-specific capabilities (GCA and SCA), respectively. Although many articles have been published on GP in maize (Lorenzana and Bernardo, 2009; Windhausen et al., 2012; Lehermeier et al., 2014; Cooper et al., 2016; Zhang et al., 2017; Dias et al., 2018; Messina et al., 2018; Alves et al., 2019; Millet et al., 2019; Costa-Neto et al., 2020; Cui et al., 2020; Das et al., 2020; Wang et al., 2020; Rogers et al., 2021), no studies on the best genetic design to build the training population have yet been conducted. This population should maximize the accuracy and contemplate practical restrictions, such as the costs and logistics of crosses to be made. Thus, in this study, we aimed (i) to empirically evaluate the effect of genetic designs when used as a GP training population of single maize hybrids obtained through full diallel (FD) or *via* NCII, and (ii) to identify the possibility of reducing the number of crosses and genitors to compose these TSs (Fristche-Neto et al., 2018).

In addition to the standard genetic designs, we also evaluated the possibility of using optimized training populations (OTS) aiming to reduce the number of individuals for training genomic prediction without reducing accuracy. For this purpose, we used the algorithm proposed by Akdemir et al. (2015) with predefined population size. Therefore, to predict the FD, we used the NCII, the testcross (TC), and OTS as the TS with sizes of 32 (with the same size of the TC data set), 152, 272, and 393 hybrids (with the same size as the NCII data set). Following the same idea of aiming to predict NCII, we used the TC and OTS with 32, 152, and 272 hybrids.

Our results suggest that TC is the worst genetic design to be used as a TS to predict simple maize crosses that must be obtained through FD or NCII. On the other hand, NCII is the best TS for the prediction of hybrids taken from FD. In addition, combinations from FD or NCII can be well predicted using OTS, thus reducing the total number of crosses to be made. However, the number of parents and crosses per parent in the ST should be maximized.

Training Populations Using Public Databases—An Alternative

Due to the scarcity of resources in the initial phases, we addressed the possibility of incorporating public databases in the composition of our training populations (Morais et al., 2020). Small-scale public and private programs with limited budgets often lack the financial ability to genotyping a considerable number of individuals to apply GP efficiently. In this regard, Morais et al. (2020) have evaluated the usefulness of incorporating public database panels to compose tropical GP training populations. In this context, the following public databases were used: (a) ASSO—Nested Association Mapping Population (NAM) combined with the Maize Association Panel 282 (166 + 282 endogamic lines, respectively); (b) NCRPIS—United States Department of Agriculture—Agricultural Research Service (USDA-ARS), North Center Regional Plant Introduction Station (2.046 endogamic lines); (c) USP—tropical endogamic lines of the University of São Paulo (64 endogamic lines).

These databases contained phenotypic information regarding plant height (PH, in cm), ear height (EH, in cm), and the SNP markers data. A total of 29 training populations (TPs) were defined and divided into four scenarios to determine the best strategy to apply public databases to predict lines.

The best predictions were achieved with the strategy of the TP composed by candidates selected with an optimization algorithm from all the public database and private lines, even at the smallest TP sizes evaluated (81 and 281 TP sizes). On the other hand, the lowest predictive abilities were achieved using only the Tropical USP database as training and validation populations (VP), due to its lack of genetic variability and reduced population size, hindering prediction. The results of all four scenarios of TP formation showed that the predictive ability increased with the increase of TP size, the relationship rate between TP and VP, and genetic variability. (Rife et al., 2018) revealed a similar potential of GP to predict wheat traits using historical data across several public breeding programs, reinforcing the possibility of using external data for model training.

The optimization of the training population proposed by Akdemir et al. (2015) showed promising results, even when the training population size was reduced. For example, small groups of individuals (250) selected in public panels are enough to achieve predictive abilities of over $r = 0.44$ and $r = 0.53$, for PH and EH, respectively. Optimizing the TP can increase the representation of the subpopulation, allowing for an efficient and controlled updating of the training population over the years (Akdemir et al., 2015).

Nevertheless, what is the real reason to use public databases, and how does it fit into a breeding framework? The use of public data aims to an early-start GP with reduced costs and over the years, to setup a more complex GP training population. The number of individuals from the program genotyped and phenotyped will increase as time goes on, reducing the participation of public databases in the training population and thus paying off the costs of genotyping the population in training over the years. For example, the total cost of the training population could be divided over 5 years, with the public database replacing 20% per year of the training population with individuals from the program.

Considering a training population that is 10 times bigger than the VP, this strategy should be conducted as follows: in the first year, (a) genotyping and phenotyping of the germplasm program, composing 10% of TP, along with external individuals selected by optimization procedures (90% of TP), (b) out of 10%, established as the VP (new progeny with no phenotyping data), (c) validation and prediction of GEBV. In the second year, (a) genotyping and phenotyping of individuals from the germplasm program (10% of the TP), (b) once again, new individuals are to make up the VP (10% of TP), while the remaining individuals from the germplasm program are to be a part of TP, with the TP composed by 70% of external individuals selected from optimization procedures and 30% of internal individuals genotyped previously, (c) validation and prediction of GEBV. As genotyping will be performed annually, after 6 years, the TP would be composed exclusively of individuals from the program. In the sixth year, the best performer could optimize the training population with internal individuals, maintaining a good prediction ability index. This procedure optimizes the technical, operational, and financial balance, considering the resources available over time and each harvest.

TABLE 1 | Reports on the comparison between GBS and array regarding genomic studies.

Compared platforms	Species	Method	Overall result	References
GBS and array	Wheat	GP	GBS comparable to or better than an array	Elbasyoni et al., 2018
GBS and array	Barley	GWAS	Broadly similar conclusions	Darrier et al., 2019
SSR, GBS, and array	Wheat	GP and diversity	Array underestimates diversity measures; similar predictive abilities	Chu et al., 2020
GBS and array	Maize	GWAS	Platforms were complementary for detecting QTL	Negro et al., 2019
GBS and array	Maize	GP	Similar results depending on the prediction model	Sabadin and Fritsche-Neto, 2020

QTL, quantitative trait loci.

SEARCHING FOR NEW SOURCES OF MARKERS AND REFERENCE GENOMES

Impact of the Genotyping Platform in GP

Nowadays, SNPs are the most widely used molecular markers in genomic studies, as they are abundant and evenly distributed in the genome. In addition, genotyping platforms that provide many markers have quickly, accurately, and cost-effectively allowed for the use of molecular tools, including GP. High-performance genotyping platforms, such as SNP-array and next-generation sequencing (NGS) provide thousands of markers for hundreds of samples, making them very suitable (Rasheed et al., 2017) for this purpose. Since there are different technologies to be detected, SNP-type markers can be different and located in distinct points of the genome so that later genomic studies can be affected by them. Recent studies have suggested comparable GWAS results, genetic diversity, and GP using different genotyping platforms in several species, including maize (Elbasyoni et al., 2018; Darrier et al., 2019; Negro et al., 2019; Chu et al., 2020) (**Table 1**).

In this context, we studied how SNP markers obtained from two genotyping platforms (616K SNP-array and GBS) affect the GP in our germplasm (Sabadin, 2020). We also attempted to verify the effect of the use of different reference genomes in SNP calls *via* GBS (i) using the most common reference genome, line B73 (GBS-B73), (ii) using a simulated reference genome built with GBS data, considering all inbred lines (GBS-Mock-All), and (iii) using a simulated reference genome built with GBS data from a single line, our heterotic pool tester L56 (GBS-Mock-L56). For this purpose, we used the USP data set mentioned above (see section above "Training populations using public databases"). To build the simulated genome, we used a pipeline developed by Melo et al. (2016), which captures the polymorphism regardless of an external genome. Finally, for each set of SNP marker data obtained from different platforms and approaches, we performed the GPs considering both the additive (GBLUP additive) and the additive-dominance (GBLUP additive-dominance) models.

Density and Distribution of SNPs

The density and distribution of SNP markers varied according to the genotyping platform chosen. In our study, the SNP markers discovered by SNP-array and GBS-B73 had the same reference genome, which allowed us to compare them regarding marker distribution on chromosomes and detect coincident SNP as well. Despite the difference in the number of SNP markers (62,409 for SNP-array and 5,594 for GBS-B73), both platforms had similar distributions along the genome. However, only 300 SNP markers coincided, suggesting that they detected polymorphisms in different regions. Although this is an important result, these differences were only consistent for some GP models.

The GBLUP model is based on the genomic relationship between genotypes to estimate the genetic values of non-phenotyped individuals. Therefore, assessing the genomic relationship is more important than the polymorphism resolution, which was confirmed when we evaluated the additive genomic relationship matrix (Ga) and the genomic dominance matrix (Gd). For the Ga matrices, high correlations were observed between the SNP-array, GBS-B73, and GBS-Mock-All SNP data sets ($r = 0.88$), revealing that these approaches estimate the additive genomic relationship between hybrids in a similar way. However, for the Gd matrices, lower correlations were observed among all SNP data sets, which show that the polymorphism captured by these platforms estimated the dominance effects differently. GBS-Mock-L56 displayed low correlations with other SNP data sets and had a low performance for all downstream analyses, proving that it is an erroneous alternative to sample polymorphism within the population, since only polymorphisms between L56 and other individuals were identified. This information is crucial when the aim is to predict the genetic values of hybrids, although the architecture of the feature can influence the performance of GP models.

Similarly, when considering the variance captured by the additive effects and the dominance deviations, these proportions also vary depending on the genotyping platform and the genetic architecture of the characteristic (**Figure 2**), which can be explained by the reduction in the number of markers, which consequently inflates the effective size of these markers. On the other hand, the SNP-array captured higher proportions of total variance and dominance, yet it was close to zero in the GBS-Mock-L56, considering all characteristics. In addition, the differences for grain yield (GY) were more significant than for simple characteristics (plant and ear heights).

As far as predictive abilities are concerned (**Figure 3**), genotyping platforms and reference genomes do not affect the additive model, except for GBS-Mock-L56. Furthermore, the use of a reference genome historically unrelated to the evaluated germplasm, such as the B73 genome (temperate maize), seems to be enough to capture the additive relationship of the genotypes within the population.

This situation can change greatly when we consider the effects of dominance to estimate genetic values. In our study, except for GBS-Mock-L56, small differences in predictive capabilities were observed among SNP data sets, when we performed the GBLUP additive-dominance model. Furthermore, the differences were more remarkable for GY, supporting the fact that the inclusion of the dominance effects of GP models is more relevant for complex traits. The coefficients of determination between GEBV estimates remained high (the lowest was for GY, $R^2 = 0.88$) but below that when obtained with the additive model.

Finally, for GP purposes, the most common genotyping platforms (SNP-array and GBS) offer very similar predictive abilities when using only additive effects in GP models. However, when we add dominance effects, their performance may change, especially when estimating hybrid performance. Dominance effects are critical to hybrid GP, and therefore, the choice of a genotyping platform may affect the estimates of genetic values. However, the differences appear to be small and acceptable in some cases. Furthermore, the use of a reference genome historically unrelated to the evaluated germplasm does not seem to be a decisive factor for GP since it can sample the haplotype variability among genotypes within the population. Another highlight uses a simulated reference genome to discover SNP since it does not depend on an external genome to detect polymorphisms. This strategy may be a valid alternative when conducting GP studies with reliable estimates, especially for orphan crops, where a reference genome is not yet available. Somehow, sampling polymorphisms consistently, using all genotypes within the population, is recommended to build the simulated genome.

GENETIC ARCHITECTURE AND FURTHER GENOMIC PREDICTION MODELING

Connecting Phenotypic and Genomic Variation

Once optimal germplasm characterization, population structure, training population mating design and composition, and genotyping methodology were defined, there was interest in further improving predictive abilities through modeling (Alves et al., 2019, Galli et al., 2020). The ability of the GP to connect phenotype and genotype has been proven to have a strong relationship with the genetic architecture of the trait. In this sense, tools such as GWAS have been applied, and the results have suggested the existence of a wide range of genetic control patterns in agronomic traits. Thus, many GP methods have been proposed to address the domain of genetic architectures. However, for open pollination species, such as maize, while the identification of variants and architectures by GWAS is usually performed in inbred lines, the GP is mainly directed at selecting hybrids. In this sense, the usefulness of *a priori* GWAS in lines to predict its hybrid offspring has been explored by Galli et al. (2020).

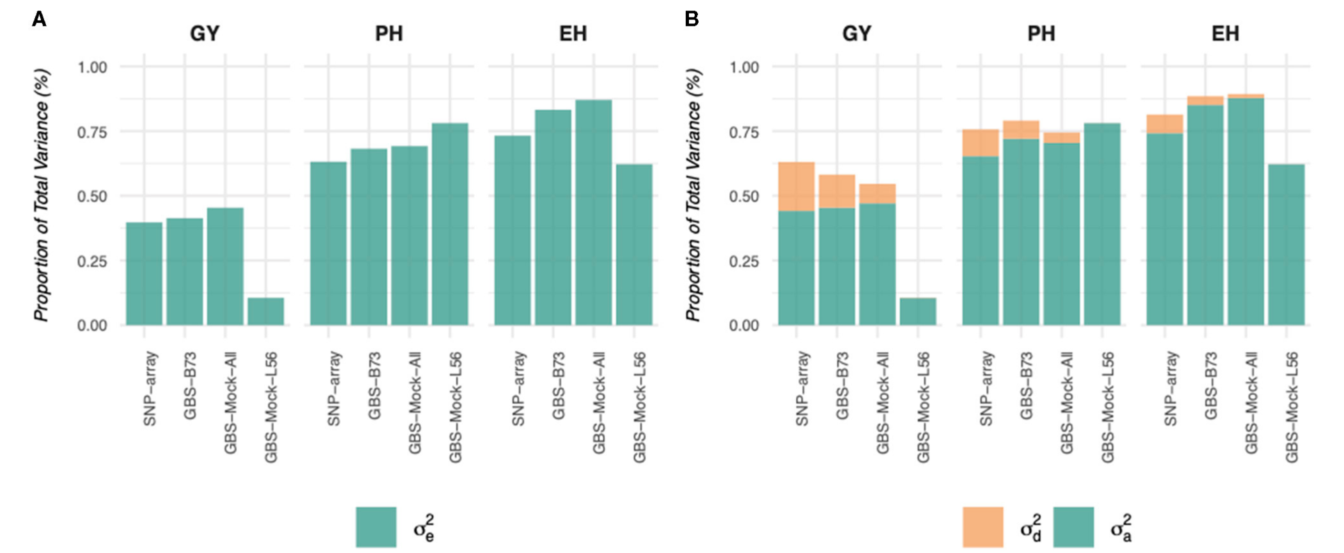


FIGURE 2 | Proportion of the phenotypic variance explained by the estimated components of variance in the different traits (EH, ear height; PH, plant height; GY, grain yield), models and scenarios studied.

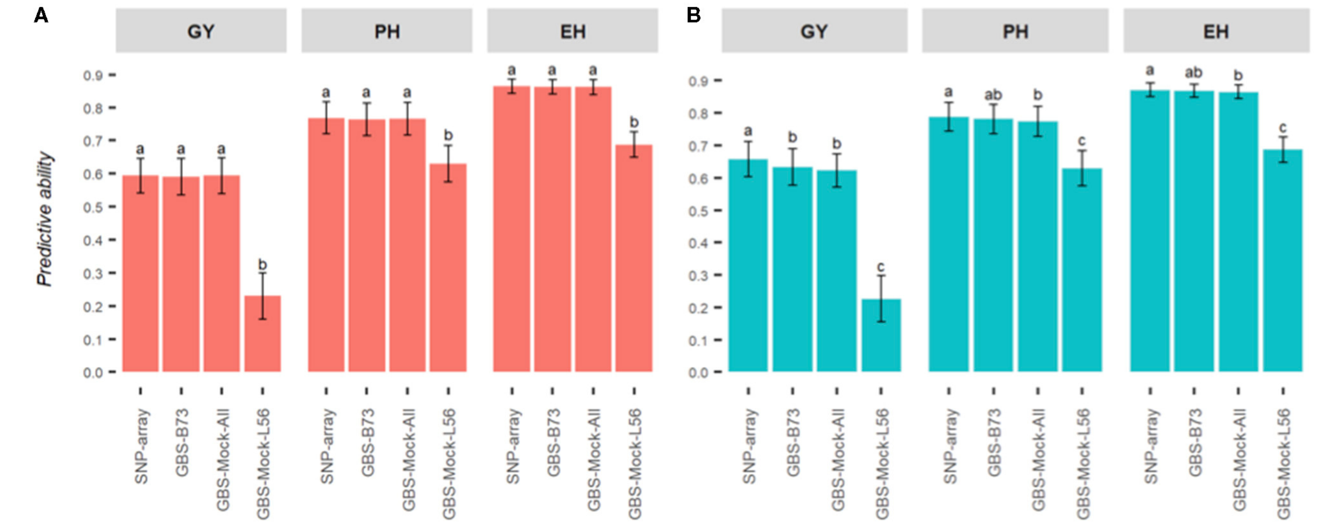


FIGURE 3 | Summary of the predictive abilities for each combination of model and genotyping scheme studied for three agronomic traits in maize (EH, ear height; PH, plant height; GY, grain yield).

The trait used in the case study was the low-nitrogen tolerance index (LNTI).

In previous GWAS (Morosini et al., 2017), four significant trait marker associations were identified in the parental population. The influence of these associations was verified for MAS, GP, and the MAS + GP of hybrids (**Figure 4**). The GP was performed with all molecular markers, except when associated with the MAS. For MAS + GP, the significant markers were removed before calculating the genomic relationship matrices. Three GP methods, namely BayesB, GBLUP, and RKHS (**Figure 4A**). Finally, GWAS was performed considering the additive, dominance, and additive and dominance in hybrids to verify the coincidence of associations with the parental lines (**Figure 4B**). The predictive ability of LNTI was observed to be low, ranging from −0.019 to 0.107 (**Figure 4A**). It was also shown that (i) the MAS of hybrids with markers identified in inbred lines had the lowest predictive abilities; (ii) adding *a priori* information from inbred lines of GWAS decreased the predictive ability of GP (MAS + GP); (iii) GP alone produced the best results.

To date, many studies have found that GP accuracy can be enhanced using *a priori* information, especially from GWAS

(Zhang et al., 2014; Spindel et al., 2016). However, the results are conditioned by factors, such as trait heritability and the variation explained by the main genes (Bernardo, 2014). Furthermore, the results obtained by Galli et al. (2020) corroborate the long-standing hypothesis of the lack of connection between inbred lines and the performance of their hybrid offspring. In addition, the GWAS of hybrids produced different marker-trait associations to those found for the parental lines published in 2017. The differences observed were both the nature of intralocus interaction and the location of markers, suggesting that the most important genes driving phenotypes in inbred lines and hybrids might be different.

Understanding the Impact of Heterosis in GP

According to Sprague and Tatum (1942), hybrid performance can be divided into two components, namely general combining ability (GCA) and specific combining ability (SCA). The GCA component can be explained by the differences between the average performance of parental lines in crosses and the average of the overall population. In this sense, the GCA of a line depends on the substitution effects of the allele and involves additive and non-additive genetic effects (Reif et al., 2007). The SCA, on the other hand, represents the deviation of hybrid performance from parental averages. This component is often attributable to deviations from additivity due to dominance and epistasis (Reif et al., 2007), and it is one of the most critical components of hybrid performance. Thus, the additive and non-additive effects of markers must be estimated to consider all the genetic variance present in a population.

The modeling of non-additive effects in genomic studies can provide several advantages (Technow et al., 2012; Varona et al., 2018), such as (1) increasing the accuracy of prediction of genomic selection methods, (2) allowing for the allocation of crossover and consequently, and (3) a better exploration of heterosis (Kadam et al., 2016). However, one of the barriers is that additive and non-additive effects are often not mutually orthogonal. For this reason, the parameters of variance that enter genomic models (for example, the additive and the dominance variances) cannot be used directly to break down total genetic variance into GCA and SCA components. As presented by Alves et al. (2019), due to their flexibility, Bayesian models can be used to estimate these important parameters, especially when the genetic design does not allow an orthogonal decomposition of genetic variance in these components.

In this context, Alves et al. (2019) presented a method to decompose genetic variance into GCA and SCA using Bayesian genomic models that account for additive and non-additive effects (dominance and epistasis).

The proposed method can be applied not only to single hybrids but also to double and triple hybrids. As proof of concept, the proposed approach was applied to the data set described above (USP, see section Germplasm Characterization). The results showed that non-additive effects play a crucial role in expressing quantitative characters under stress conditions (especially GY, **Figure 5**). This study also showed that the accuracy of the prediction models that account for the additive and non-additive effects depends on interest characteristics. It was also found that selecting 30% of the best single-crosses during the pre-selection phase in the field, based on GP with additive and non-additive effects, leads to a subset of hybrids that contained 85–95, 70–80, and 75–85 of the 5% higher hybrids for ear height, plant height, and GY (**Figure 5**), respectively.

MODELING GENOTYPE × ENVIRONMENT INTERACTION (G × E) IN GP

Finding Novel Kernel Methods and Modeling Structures for G × E

The G × E is a multiplicative non-additive effect due to the non-parallel trait-specific phenotypic responses, a function of genotype diversity and environmental variation. Since 2012, when the marker by environment interaction approach was developed (Burgueño et al., 2012), the analysis and modeling of G × E have evolved from the genotype to the gene or genomic level (Crossa, 2012). However, multi-environment modeling to predict maize hybrids started with Dias et al. (2018) (**Table 2**). Since then, several efforts have been made to extend those modeling approaches when considering different kernel methods and structures. For example, different G × E approaches to include genomics and large-scale environmental data (enviromics) (Bandeira e Sousa et al., 2017; Costa-Neto et al., 2020; Rogers et al., 2021) using explicit covariates for modeling reaction-norms (Millet et al., 2019) or implicit covariates derived from multivariate structures (e.g., Dias et al., 2018; Krause et al., 2020).

Our research group aimed to understand how environmental characterization (envirotyping) and non-linear kernels could improve prediction models, including G × E (Bandeira e Sousa et al., 2017; Costa-Neto et al., 2020). Below, we detail a case study using our tropical maize germplasm from USP, in which we were able to test novel G × E structures and kernel methods to model genomic × environment effects.

We conducted an extensive study on G × E over three agronomic traits in tropical maize (GY, PH, and EH) for two different sets in Brazil. Bandeira e Sousa et al. (2017) tested two kernel methods, a linear (GBLUP, hereafter abbreviated as GB) and non-linear (Gaussian Kernel, GK) kernel and four modeling structures for G × E using (i) single-environment (SE) model, using the average values of the genotypes for all environments; (ii) multi-environment, main genotypic effects model (MM); (iii) multi-environment, single variance G × E deviation model (MDs), and (iv) multi-environment, environment-specific variance G × E deviation model (MDe). Models without G × E structures (SM and MM) were less accurate than those including G × E effects (MDs and MDe). For the MM, MDs, and MDe models, the increase in the prediction accuracy of GK over GB ranged from 9 to 49%. As expected, GY was the less predictable trait due to its polygenic nature, and because of that, this trait became the main target for

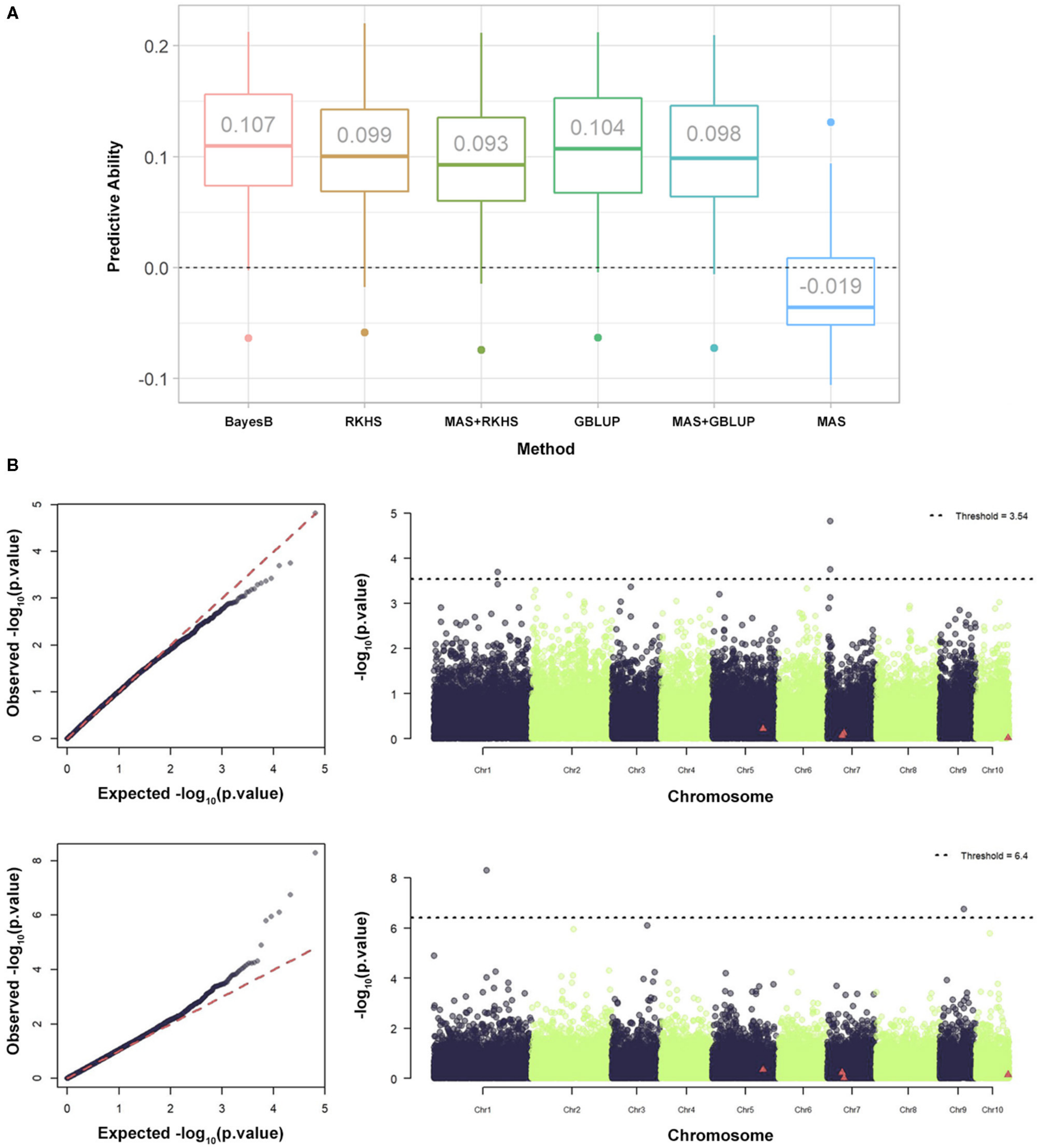


FIGURE 4 | Performance of different statistical models and GWAS-based strategies for genomic prediction of maize hybrids. **(A)** Summary of the predictive capabilities of the Low Nitrogen Tolerance Index (LNTI) in maize hybrids using BayesB, RKHS, MAS + RKHS, GBLUP, MAS + GBLUP, and MAS additive. **(B)** Summary of GWAS, QQ, and Manhattan graphs for LNTI. The graphs represent additive GWAS (upper) and dominance (lower). The MAS was based on statistically significant associations identified for LNTI by Morosini et al. (2017).

further studies. For all traits, few differences were observed between the MDs and MDe models. Gaussian Kernel was observed to outperform all GB-based models in accuracy for all models, with an average accuracy gain from 34 to 70%. However, for EH and PH, the gains using GK were smaller than using GB.

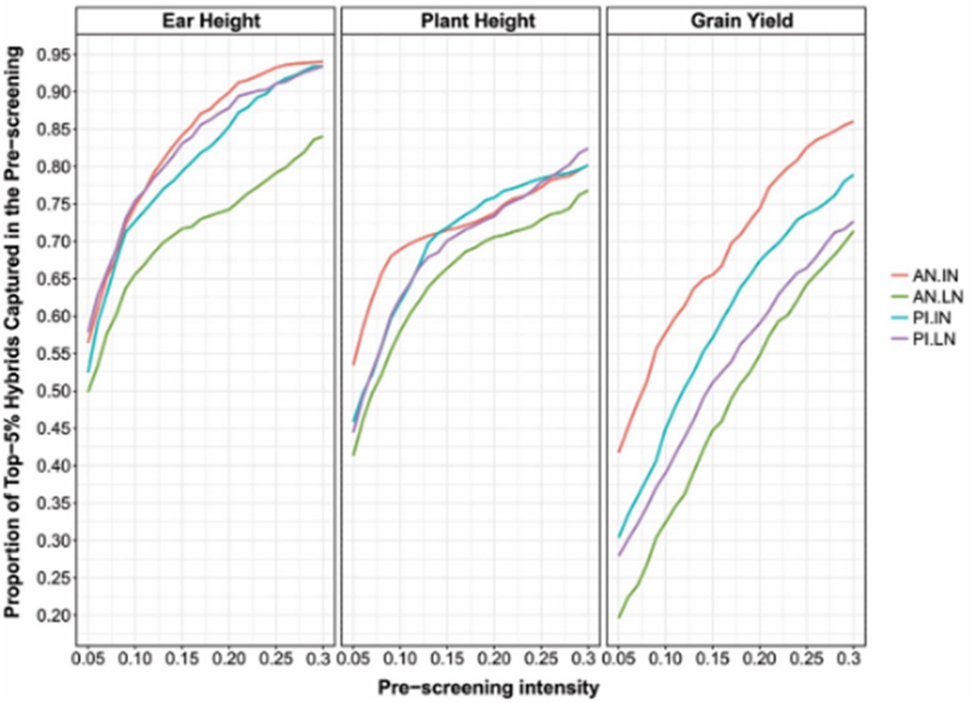


FIGURE 5 | According to phenotypic classification, the proportion of 5% higher hybrids was identified by pre-screening based on cross-validation via GP using the additive + dominance model at a certain selection intensity (x-axis). Each panel corresponds to one evaluated character. The lines within a graph represent different environments (AN: Anhembi; PI: Piracicaba; LN: Low nitrogen; IN: Ideal nitrogen).

Understanding the Contribution of Non-additive Effects for G × E

Since 2017, some studies have pointed that the use of additive (A) plus non-additive effects (e.g., dominance, D; epistasis, A × A) might drastically improve the accuracy of GP for maize hybrids (Acosta-Pech et al., 2017; Dias et al., 2018; Alves et al., 2019, 2021; Costa-Neto et al., 2020; Ferrão et al., 2020; Ramstein et al., 2020; Rogers et al., 2021), especially with G × E under multi-environment conditions. It seems that the main dominance effect (D) plus dominance by the environment interaction (D × E) corresponds to about 50% of the observed phenotypic variation for complex traits, such as GY in hybrid maize. This is an important issue because the usage or non-usage of non-additive effects only depends on the computational effort expected, that is, from raw molecular marker data, it is feasible and easy, nowadays, to compute both additive or non-additive effects and their relatedness-based matrices to implement GBLUP and kernel models (Alves et al., 2019). The use of algebra resources to remove the complexity of the variance–covariance matrices, such as the singular decomposition value (Costa-Neto et al., 2020; Cuevas et al., 2020) and factor analytic structuration (Dias et al., 2018; Rogers et al., 2021) is a computationally smart way to translate model complexity into accuracy gains. Here, we detail the results we found as an extension of the study of Bandeira e Sousa et al. (2017), related to the first option resource previously mentioned.

We investigated different models involving additive (A) and additive-dominance (AD) main effects (MM model, but using A + D), along with the interactions (MDs models) including reaction-norm for A and D effects to predict GY (Costa-Neto et al., 2020). After the use of GB and GK, a third kernel method was also tested, the so-called deep kernel (DK), which takes advantage of the arcsine kernel that thought the available phenotypic data could mimic different hidden layers an in-depth learning approach. Thus, DK is also a non-linear kernel, but unlike GK, it approaches the genomic relatedness into an empirical relatedness of the individuals across a diverse set of environments. Our results suggest that DK outperforms GB and GK when exploring dominance effects in hybrid prediction. In terms of explaining the phenotypic variation across multi-environment, the DK and GK models better captured the genomic and enviromic sources and reduced the residual variance of the models. Then, we tested three scenarios, namely CV1, novel genotypes in known environments; CV2, sparse MET conditions, some genotypes at some environments, and CV0, novel environments.

In addition, our results indicated that GK and DK explore the G × E variation better (in this case, G × E = A × E + D × E) in a less computationally expensive way than GB. The GB kernel was the worst kernel method for exploring D effects to predict GY in maize hybrids. For all prediction scenarios (CV1, CV2, and CV0), we observed that accuracy gains could only be achieved

TABLE 2 | Strategies and main results for multi-environment genomic prediction of grain yield, the main agronomic trait in hybrid maize breeding since 2017.

Germplasm	Core ideas and importance	References
Tropical hybrids	The first use of GP for modeling G × E and predicting maize hybrids	Acosta-Pech et al., 2017
	Differences of several variance–covariance structures and Gaussian kernel in the prediction of G × E	Bandeira e Sousa et al., 2017
	Contribution of dominance effects and factor analytic structures for G × E	Alves et al., 2019
Temperate DH lines	The use of crop models with genomic prediction (CGM-WGP) is better than GBLUP	Cooper et al., 2016
	Update of CGM-WGP and application in predicting phenotypic landscapes	Messina et al., 2018
Temperate hybrids	Use of factorial regression to find covariates that explain genomic-enabled reaction norms	Millet et al., 2019
Tropical hybrids	Deep kernels accounting for genomic and near-infrared relatedness kernels	Cuevas et al., 2019
	The importance of additive (A), dominance (D), and AA, DD, and AD covariances under Bayesian prediction approaches	Alves et al., 2019
	The use of deep kernel and Gaussian kernel for modeling additive and dominance G × E effects with reaction norm	Costa-Neto et al., 2020
	Multivariate GBLUP using factor analytic structures	Krause et al., 2020
Temperate hybrids	The use of dominance and functional enrichments to increase GP	Ramstein et al., 2020
	The use of difference variance–covariance structures to model dominance and reaction-norm	Rogers et al., 2021
Tropical hybrids	Contribution of non-additive effects and mega-environment grouping in prediction accuracy	Alves et al., 2021

for GB-based models when including some envirotyping data as the main effect (W) or as reaction-norm (G × W = A × W + D × W). The non-linear kernels were also more efficient at using the phenotypic records in training models for CV1, CV2, and mostly for CV0. For CV0, the combination of DK and more straightforward reaction-norm models (including only A + D + W effects) achieved almost the same accuracy as more complex structures (A + D + W + A × W + D × W). This suggests that to predict future scenarios using actual TSs, the use of enviromic sources combined with additive and dominance genomic data, both modeled with non-linear kernels, is the best way to achieve higher mathematical accuracy biologically that better represents novel G × E conditions.

Finding Novel Enviromic Approaches to Deal With G × E

Combined with phenotypic and genotypic data, the use of envirotypic data sources can leverage the molecular breeding strategies addressing the prediction of tested and untested environments, such as climate change scenarios (Millet et al., 2016, 2019; Messina et al., 2018; Bustos-Korts et al., 2019; de los Campos et al., 2020; Guo et al., 2020). These data have been incorporated into GP in the last ten years to better model the G × E interaction according to the reaction norm (Heslot et al., 2014; Jarquín et al., 2014; Gillberg et al., 2019; Costa-Neto et al., 2020; Rogers et al., 2021). However, it is difficult for most breeders to deal with this interaction between environmental models, ecophysiology, and genetics (Costa-Neto et al., 2021), in which we need to (i) implement a cost-effective and intuitive pipeline to integrate envirotyping data in GP and (ii) find novel enviromic approaches, more capable of describing phenotype-envirotype covariances and translate it into accuracy gains. Below, we briefly present the results by Costa-Neto et al. (2021), who implemented an envirotyping pipeline and then review some of the main applications of enviromic data achieved for other groups.

Costa-Neto et al. (2021) presented two novel approaches to modeling the environmental similarity from enviromic data. Using a proof-of-concept data set, we tested the importance of (i) EC-specific kernels for main environmental factors and (ii) the envirotyping level at each key development stage of crop development. For the latter, we proved accuracy gains of the reaction-norm models using a specific environmental relatedness, built using ECs for each development stage, concerning the benchmark environmental relatedness (single-environmental kernel using all ECs at all development stages). This approach enabled a better understanding of which development stage impacts the relatedness of individuals across MET. We tested a CV1 scheme to predict GY using a drastically reduced phenotyping level (only 20% of the phenotypes were used as TS). We showed that a model without enviromic data has a minimal prediction accuracy ($r = 0.101$), and the inclusion of envirotyping data boosted the prediction up to $r = 0.504$ (enviromic by development stage) and $r = 0.485$ (enviromic for all crop development stages).

An alternative approach for the use of environmental relatedness kernels is the adoption of single-covariate regressions (Ly et al., 2018) or the first step of screening in which the ECs that best explain the trait variation are used to fit a simpler but more accurate linear reaction-norm structure (Millet et al., 2019). These ECs can be collected from in-field sensors or public databases (for more details, see the next section) and also consider stress-covariates derived from crop growth models (CGM) (Heslot et al., 2014; Rincent et al., 2017). For the latter, a more robust single-step approach relies on the integrated use of GP with CGM, which was successful in predicting the performance of DH maize lines on water-stressed environments (Cooper et al., 2016) and across a large target region of the breeding program in the United States (Messina et al., 2018). For low-budget breeding programs that are unable to invest in large phenotyping for ecophysiology traits (e.g., biomass accumulation during crop life) need to improve accuracy in training CGM. An alternative can be in the exploring of the environmental relatedness or EC-specific regressions, which increases the accuracy of GP in hybrid prediction more simply (Costa-Neto et al., 2020; Rogers et al., 2021) with a satisfactory ability to predict cultivar responses (de los Campos et al., 2020) and explain the reaction-norm for both complex quantitative traits (Ly et al., 2018; Millet et al., 2019) and less complex traits (Guo et al., 2020; Jarquin et al., 2020).

OPEN-SOURCE R PACKAGES TO FACILITATE THE ADOPTION OF GENOMIC PREDICTION

Since the first work on GP, published approximately 20 years ago (Meuwissen et al., 2001), a wide number of computational solutions have been developed to process data and run prediction models, such as *BGLR* (Pérez and de los Campos, 2014), *rrBLUP* (Endelman, 2011), and *sommer* (Covarrubias-Pazaran, 2016). For plant breeding, most of these solutions were implemented in R, an open statistical-computational environment. Nowadays, these software solutions can offer the processing of genotyping data (Granato et al., 2018b), fit marker regressions or genomic wide association analysis (Endelman, 2011), run GP accounting for several multi-trait multi-environment approaches (Pérez and de los Campos, 2014; Covarrubias-Pazaran, 2016; de los Campos and Gr?neberg, 2016; Granato et al., 2018a; Montesinos-López et al., 2019), and integrate envirotyping sources in the reaction-norm modeling of G × E (Costa-Neto et al., 2021). Here, we briefly discuss three software developed by the Allogamous Plant Breeding Laboratory of the University of São Paulo as part of our experience in the field of genomic-enabled prediction of maize hybrids.

To deal with genotyping data, we developed the package *snpReady* (Granato et al., 2018b), which helps the user with quality control and the recoding of markers. In addition, it helps obtain some parameters of population genomics. This package implements a pipeline of conversion, imputation of missing data, and preparation of genotyping data for genomic analysis, outputting matrices in appropriate formats for different software. These applications are simple and enough to be integrated into the breeding pipelines or coupled with other environments, such as shiny (Matias et al., 2019).

After that, we realized the need to implement a computationally efficient approach that facilitates the use of multi-environment prediction structures accounting for G × E. To fill this gap, we developed the package, Bayesian Genotype plus Genotype Environment (BGGE, Granato et al., 2018a), which considers a wide number of genomic environmental structures and two kernel methods (linear GBLUP and non-linear Gaussian kernel) in a processing time of five times faster than Bayesian Generalized Linear Regressions *(BGLR)*. Furthermore, it uses algebra resources resulting in a significant gain in processing speed, especially for large data sets (Granato et al., 2018a), such as near-infrared data (Cuevas et al., 2019), historical yield trial data (Cuevas et al., 2020), and enviromics (Costa-Neto et al., 2020).

For the latter, since the first work involving the use of environmental information in GP (Heslot et al., 2014; Jarquín et al., 2014), there is a need to fine-tune the methodologies of collection, processing, and the use of this data in GP. Generally, the collection, organization, and processing of environmental data are steps that require the installation of equipment in the field. In turn, such equipment may be expensive or difficult to access for some research groups in specific regions or countries. Therefore, we have decided to enter a routine of climate data collection through NASA's Prediction of Worldwide Energy Resources (NASA-POWER, Sparks, 2018), which can access information daily, anywhere in the world. Thus, the computational development of these routines evolved to the development of the first open-source envirotyping pipeline, named *EnvRtype* (Costa-Neto et al., 2021). Three modules of envirotyping are offered in this package, namely (i) the collection of raw environmental information from public platforms, requiring only the geographic and temporal coordinates of the experiments and processing data set, (ii) environmental characterization based on the use of the processed environmental covariables to describe the typology of the environments, and (iii) the implementation of GP models enriched with ecophysiological parameters, considering three different structures of reaction-norm, and subsequently incorporate them into the prediction models under a Bayesian framework in the same way as in *BGGE*.

FINAL REMARKS

This work aimed to present a review of our results, which shows that it is possible to increase the accuracy in the prediction of hybrids. This requires the use of optimized training populations, the inclusion of non-additive genetic effects in the prediction models, and environmental information to compose the matrices of G × E covariance and non-linear kernels of genomic relationship. On the other hand, there are no significant gains in the accuracy using GWAS information in parental lines, population structure, or using markers from new generation sequencing. Below, we conclude our work by describing some lessons we learned, both from our studies and other groups.

GWAS Might Be Useful to Discover the Architecture of G × E for Further GP Modeling

Going back to our experience with GWAS described in this review, we found in our road map that the use of GWAS for further prediction modeling might be more successful, especially to understand genomic-environment sources of G × E in our tropical germplasm. For example, Vidotti et al. (2019) used GWAS to establish a relation between the genetic control of the maize responsiveness and *Azospirillum brasilense*, a plant growth-promoting bacteria (PGPB) common in tropical soils and related to maize nitrogen fixation. The GWAS outcomes helped understand how heterosis is important for improving the quality of crop systems by increasing the nitrogen use efficiency (NUE) of maize. Another promising approach is presented by Millet et al. (2016), which involves the use of GWAS to find genomic regions associated with the reaction norm for key environmental factors expected in future scenarios.

A similar approach uses only the phenotypic data to model parameters of adaptability and stability, as in the work by Gage et al. (2017). Combining GWAS and such parameters that reflect the effect of G × E for specific genotypes, these authors were able to explore the genomic-related sources that explain the drivers of phenotypic plasticity and how the artificial selection shaped these patterns on the temperate maize germplasm in the United States. Finally, another good example is given by Ramstein et al. (2020). These authors used GWAS to find quantitative trait loci (QTLs) related to the phenotypic variation of some important traits in maize. Then, using gene annotation, it was possible to explore the functional contribution of those QLTs to express the phenotypes and the increasing accuracy of GP. This functional enrichment in further GP models contributed to the increase in the accuracy of a hybrid panel of temperate maize cost-effectively. It can also be useful for the tropical germplasm, which still demands the development of a higher panel of inbred lines to address the test of those hypotheses.

How to Deal With the Complexity and Diversity of Big Data?

In the last 20 years of genomic selection research, the plant breeding community is still learning how to connect a wide number of data sources related to the "Central Dogma of Molecular Biology" with the observed phenotypic variation of traits in field trials, which began as a regression of phenotypes over molecular markers evolved by the integration of different data sources and modeling structures. Computational research in GP must develop to capture other data sources in a computationally smart way and find which structure is better to integrate each type of data. For example, Costa-Neto et al. (2020) suggest that the use of Deep Kernels (DK) is a faster and more accurate way to model both genomic and enviromic relatedness than benchmark GBLUP approaches, which is similar to results by Cuevas et al. (2019) who used near-infrared data. However, it seems that the paradigm of "less means more" when dealing with some sources of data, such as enviromics, in which we still have a long pathway in optimizing approaches capable of capturing gene × envirotype interactions across crop fields. In addition, in our studies, we observed that a good enviromic kernel (W) added in the GP models as the main effect is sometimes better than modeling a full-rank reaction-norm model accounting for the genomic environment and genomic enviromics. On the other hand, works by authors, such as Cuevas et al. (2020) and de los Campos et al. (2020) show that big historical data can be implemented by different computational approaches and have a satisfactory accuracy to support the selection decisions. Thus, methodological approaches must be developed to capture exploitable patterns in big data and computational tools to implement them, the latter preferably as open-source software.

Deep learning approaches accounting for this data source can be a more parsimonious approach to taking advantage of big data without over-fitting prediction models. Finally, we find that using multi-trait multi-environment data might help design better field phenotyping trials for training GP models. As the modern computational tools attempt better to explore G × E and G × G within a multi-environment multi-trait context, the opposite path might be taken by using historical data to design future trials (Rincent et al., 2017) and scenarios (Millet et al., 2016; Bustos-Korts et al., 2019), but also to predict cultivars at novel growing conditions (Gillberg et al., 2019; Millet et al., 2019; de los Campos et al., 2020).

Are Prediction-Based Tools Cost-Effective Approaches?

Prediction-based tools are cost-effective approaches. Plant breeding is based on selecting the best-evaluated genotypes in target environments, demanding many field-testing resources (physical and financial). Therefore, GP has proven to be useful to enlarge the spectrum of individuals evaluated *in silico* but with a limited accuracy in multiple environmental conditions due to the non-additive effects related to G × E and G × G interactions. Recently the emerging new ways to include environmental data and CGM in the GP are considered good strategies to correct this deficiency in predicting G × E interaction deviations (Messina et al., 2018). In addition, these new applications allow genotype screening at reduced phenotyping costs considering virtual scenarios.

Despite the great advances that have been made, what is to come is exciting for hybrid maize breeding. New tools and models, such as the integrated use of high throughput phenotyping, CGM, and optimized tools for simulation of improvement methods can bring more resolution, realism, and depth to the predictions. With HTP, we will be able to evaluate the same plant several times over the crop cycle and increase the effective size of training populations. Additionally, even before running HTP studies in the field, it is possible to validate some protocols *in silico* for phenotyping traits, such as PH (Galli et al., 2021). On the other hand, both pathways of enviromics and CGM will allow us to build virtual improvement scenarios and predict the deviations of G × E interaction more accurately. Finally, with the simulations, we will be able to test a series of scenarios cheaply and easily, helping outline the best improvement strategies and resource allocations.

Finding Research Partnerships to Expand the Field-Testing Network

Most of the applications described in the last section consider datasets with at least four environments and almost one thousand entries (lines, DH, and hybrids), which represent the reality for at least a small-scale breeding program. As discussed in the previous sections, with the increase in the availability of data, the computational demand and the power of cutting-edge testing hypotheses in maize breeding also increase (Rogers et al., 2021). We envisage that maize hybrid breeding programs can take advantage of historical multi-environment testing data (Dawson et al., 2013) to explore the environmental impacts on the plasticity of germplasm, collecting during this process data from enviromics, and other sources of data useful to train accurate models. During this step, it is possible to integrate some simulation platform capable of generating reliable environmental scenarios (Millet et al., 2016) or phenotypic landscapes (Bustos-Korts et al., 2019), such as CGM. The use of public databases to test hypotheses, train models, or import datasets for your own purposes that might reduce costs and provide a guideline to follow. However, as we have pointed out in section Germplasm Characterization, the implementation of a well-conducted field trial for phenotypic, genotypic, and envirotypic characterization of the so-called "Modern Plant Breeding Triangle" (Crossa et al., 2021), is crucial for providing good quality data to test a wide number of hypotheses. Another interesting option is to establish partnerships with other small-scale breeding programs and public institutions in order to create a large network of field data, such as the successful partnership of public institutions in the United States—*The Genome to Field Project* (McFarland et al., 2020). In Brazil, the first steps of this approach were led by the Allogamous Plant Breeding Laboratory from USP. We tried to share every genomics database, enviromics, and high-throughput phenotyping (available in https://data.mendeley.com/datasets/5gvznd2b3n).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://data.mendeley.com/research-data/?page=0&search=%22Roberto%20Fritsche%20Neto%22.

AUTHOR CONTRIBUTIONS

RF-N conceived and designed all studies. GG, FA, FS, DL, PM, LB, GC-N, and IG generated the dataset and performed the data analysis. RF-N wrote the manuscript. GG, KB, GC-N, and JC revised the text, which all the authors finally edited. All authors contributed to the article and approved the submitted version.

FUNDING

The following public agencies supported these studies: Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, CAG-APQ-00559-12, and CAG-APQ-00555-12), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, 2013/24135-2 and2017/24327-0), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, 400029/2016-2). The private companies, namely GDM Seeds, Dupont-Pioneer, and Helix Seeds also financed many of these studies. We also acknowledge the financial support provided by the Foundation for Research Levy on Agricultural Products (FFL) and the Agricultural Agreement Research Fund (JA) in Norway through NFR grant 267806. We are also thankful for the financial support provided by the CIMMYT CRP (maize and wheat), the Bill and Melinda Gates Foundation, and the USAID projects (Cornell University and Kansas State University).

ACKNOWLEDGMENTS

The authors thank the Allogamous Plant Breeding Laboratory (University of São Paulo), all the funding agencies and companies cited above, and the Biometrics and Statistics Unit (CIMMYT) for the technical and scientific support.

REFERENCES

Acosta-Pech, R., Crossa, J., de los Campos, G., Teyss?dre, S., Claustres, B., P?rez-Elizalde, S., et al. (2017). Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. *Theor. Appl. Genet.* 130, 1431-1440. doi: 10.1007/s00122-017-2898-0

Akdemir, D., Sanchez, J. I., and Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel. Evol.* 47. doi: 10.1186/s12711-015-0116-6

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. doi: 10.1007/s10681-021-02779-y

Alves, F. C., Granato, Í. S. C., Galli, G., Lyra, D. H., Fritsche-Neto, R., and de los Campos, G. (2019). Bayesian analysis and prediction of hybrid performance. *Plant Methods* 15:14. doi: 10.1186/s13007-019-0388-x

Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., et al. (2021). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theor. Appl. Genet.* 134, 279–294. doi: 10.1007/s00122-020-03696-9

Bandeira e Sousa, M., Cuevas, J., de Oliveira Couto, E. G., Pérez-Rodríguez, P., Jarquín, D., Fritsche-Neto, R., et al. (2017). Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. *G3 (Bethesda)* 7, 1995–2014. doi: 10.1534/g3.117.042341

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

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. (2014). Genome wide selection when major genes are known. *Crop Sci.* 54, 68–75. doi: 10.2135/cropsci2013.05.0315

Burgueño, J., Campos, G., de los, 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

Bustos-Korts, D., Malosetti, M., Chenu, K., Chapman, S., Boer, M. P., Zheng, B., et al. (2019). From QTLs to adaptation landscapes: using genotype-to-phenotype models to characterize G×E over time. *Front. Plant Sci.* 10:1540. doi: 10.3389/fpls.2019.01540

Chu, J., Zhao, Y., Beier, S., Schulthess, A. W., Stein, N., Philipp, N., et al. (2020). Suitability of single-nucleotide polymorphism arrays versus genotyping-by-sequencing for genebank genomics in wheat. *Front. Plant Sci.* 11:42. doi: 10.3389/fpls.2020.00042

Cooper, M., Technow, F., Messina, C., Gho, C., and Radu Totir, L. (2016). Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial. *Crop Sci.* 56, 2141–2156. doi: 10.2135/cropsci2015.08.0512

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2020). Non-linear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity (Edinb).* 126, 92–106. doi: 10.1038/s41437-020-00353-1

Costa-Neto, G., Galli, G., Carvalho, H. F., Crossa, J., and Fritsche-Neto, R. (2021). EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. *G3 (Bethesda)* 11:jkab040. doi: 10.1093/g3journal/jkab040

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

Crossa, J. (2012). From genotype × environment interaction to gene × environment interaction. *Curr. Genomics* 13, 225–244. doi: 10.2174/138920212800543066

Crossa, J., Fritsche-Neto, R., Montesinos-lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-lopez, A., et al. (2021). The modern plant breeding triangle : optimizing the use of genomics, phenomics, and enviromics data. *Front. Plant Sci.* 12:651480. doi: 10.3389/fpls.2021.651480

Cuevas, J., Montesinos-López, O., Juliana, P., Guzmán, C., Pérez-Rodríguez, P., González-Bucio, J., et al. (2019). Deep kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3 (Bethesda)* 9, 2913–2924. doi: 10.1534/g3.119.400493

Cuevas, J., Montesinos-López, O. A., Martini, J. W. R., Pérez-Rodríguez, P., Lillemo, M., and Crossa, J. (2020). Approximate genome-based kernel models for large data sets including main effects and interactions. *Front. Genet.* 11:567757. doi: 10.3389/fgene.2020.567757

Cui, Z., Dong, H., Zhang, A., Ruan, Y., He, Y., and Zhang, Z. (2020). Assessment of the potential for genomic selection to improve husk traits in maize. *G3 (Bethesda)* 10, 3741–3749. doi: 10.1534/g3.120.401600

Darrier, B., Russell, J., Milner, S. G., Hedley, P. E., Shaw, P. D., Macaulay, M., et al. (2019). A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. *Front. Plant Sci.* 10:544. doi: 10.3389/fpls.2019.00544

Das, R. R., Vinayan, M. T., Patel, M. B., Phagna, R. K., Singh, S. B., Shahi, J. P., et al. (2020). Genetic gains with rapid-cycle genomic selection for combined drought and waterlogging tolerance in tropical maize (*Zea mays* L.). *Plant Genome* 13, 1–15. doi: 10.1002/tpg2.20035

Dawson, J. C., Endelman, J. B., Heslot, N., Crossa, J., Poland, J., Dreisigacker, S., et al. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Res.* 154, 12–22. doi: 10.1016/j.fcr.2013.07.020

De Andrade, B., Neto, R. F., and Roge, L. (2016). Genetic vulnerability and the relationship of commercial germplasms of maize in brazil with the nested association mapping parents. *PLoS ONE* 11:e0163739. doi: 10.1371/journal.pone.0163739

de los Campos, G., and Gr?neberg, A. (2016). *MTM (Multiple-Trait Model) Package.* Available online at: http://quantgen.github.io/MTM/vignette.html

de los Campos, G., Pérez-Rodríguez, P., Bogard, M., Gouache, D., and Crossa, J. (2020). A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. *Nat. Commun.* 11:4876. doi: 10.1038/s41467-020-18480-y

de Los Campos, G., and Sorensen, D. (2014). On the genomic analysis of data from structured populations. *J. Anim. Breed. Genet.* 131, 163–164. doi: 10.1111/jbg.12091

Dias, K. O. D. G., Gezan, S. A., Guimar?es, C. T., Nazarian, A., da Costa e 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

Elbasyoni, I. S., Lorenz, A. J., Guttieri, M., Frels, K., Baenziger, P. S., Poland, J., et al. (2018). A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Sci.* 270, 123–130. doi: 10.1016/j.plantsci.2018.02.019

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome J.* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Ferrão, L. F. V, Marinho, C. D., Munoz, P. R., and Resende, M. F. R. (2020). Improvement of predictive ability in maize hybrids by including dominance effects and marker × environment models. *Crop Sci.* 60, 666–677. doi: 10.1002/csc2.20096

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

Gage, J. L., Jarquin, D., Romay, C., Lorenz, A., Buckler, E. S., Kaeppler, S., et al. (2017). The effect of artificial selection on phenotypic plasticity in maize. *Nat. Commun.* 8:1348. doi: 10.1038/s41467-017-01450-2

Galli, G., Alves, F. C., Morosini, J. S., and Fritsche-Neto, R. (2020). On the usefulness of parental lines GWAS for predicting low heritability traits in tropical maize hybrids. *PLoS ONE* 15:e0228724. doi: 10.1371/journal.pone.0228724

Galli, G., Lyra, D. H., Alves, F. C., Granato, Í. S. C., e Sousa, M. B., and Fritsche-Neto, R. (2018). Impact of phenotypic correction method and missing phenotypic data on genomic prediction of maize hybrids. *Crop Sci.* 58, 1481–1491. doi: 10.2135/cropsci2017.07.0459

Galli, G., Sabadin, F., Costa-Neto, G. M. F., and Fritsche-Neto, R. (2021). A novel way to validate UAS-based high-throughput phenotyping protocols using *in silico* experiments for plant breeding purposes. *Theor. Appl. Genet.* 134, 715–730. doi: 10.1007/s00122-020-03726-6

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

Granato, I., Cuevas, J., Luna-Vázquez, F., Crossa, J., Montesinos-López, O., Burgueño, J., et al. (2018a). BGGE: a new package for genomic-enabled prediction incorporating genotype × environment interaction models. *G3 (Bethesda)* 8, 3039–3047. doi: 10.1534/g3.118.200435

Granato, I. S. C., Galli, G., de Oliveira Couto, E. G., e Souza, M. B., Mendonça, L. F., and Fritsche-Neto, R. (2018b). snpReady: a tool to assist breeders in genomic analysis. *Mol. Breed.* 38:102. doi: 10.1007/s11032-018-0844-8

Guo, T., Mu, Q., Wang, J., Vanous, A. E., Onogi, A., Iwata, H., et al. (2020). Dynamic effects of interacting genes underlying rice flowering-time phenotypic plasticity and global adaptation. *Genome Res.* 30, 673–683. doi: 10.1101/gr.255703.119

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

Hallauer, A. R., Filho, J. B. M., and Carena, M. J. (2010). "Breeding plants," in *Quantitative Genetics in Maize Breeding,* eds A. R. Hallauer, M. J. Carena, and J. B. Miranda Filho (New York, NY: Springer New York). doi: 10.1007/978-1-4419-0766-0_12

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

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., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., et al. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3 Genes Genomes Genet.* 10:2725. doi: 10.1534/g3.120.401349

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 (Bethesda)* 6, 3443–3453. doi: 10.1534/g3.116.031286

Krause, M. D., Dias, K. O., das, G., Pedroso Rigal dos Santos, J., de Oliveira, A. A., Guimarães, L. J. M., et al. (2020). Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models. *Crop Sci.* 60, 3049–3065. doi: 10.1002/csc2.20253

Larièpe, A., Moreau, L., Laborde, J., Bauland, C., Mezmouk, S., Décousset, L., et al. (2017). General and specific combining abilities in a maize (*Zea mays* L.) test-cross hybrid panel: relative importance of population structure and genetic divergence between parents. *Theor. Appl. Genet.* 130, 403–417. doi: 10.1007/s00122-016-2822-z

Lehermeier, C., Krämer, N., Bauer, E., Bauland, C., Camisan, C., Campo, L., et al. (2014). Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. *Genetics* 198, 3–16. doi: 10.1534/genetics.114.161943

Lehermeier, C., Schön, C.-C., and de los Campos, G. (2015). Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics* 201, 323–337. doi: 10.1534/genetics.115.177394

Lorenzana, R. E., and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor. Appl. Genet.* 120, 151–161. doi: 10.1007/s00122-009-1166-3

Ly, D., Huet, S., Gauffreteau, A., Rincent, R., Touzy, G., Mini, A., et al. (2018). Whole-genome prediction of reaction norms to environmental stress in bread wheat (Triticum aestivum L.) by genomic random regression. *F. Crop. Res.* 216, 32–41. doi: 10.1016/j.fcr.2017.08.020

Lyra, D. H., Granato, Í. S. C., Morais, P. P. P., Alves, F. C., dos Santos, A. R. M., Yu, X., et al. (2018). Controlling population structure in the genomic prediction of tropical maize hybrids. *Mol. Breed.* 38:126. doi: 10.1007/s11032-018-0882-2

Matias, F. I., Morosini, J. S., Espolador, F. G., and Fritsche-Neto, R. (2019). Be-Breeder 2.0: a web application for genetic analyses in a plant breeding context. *Crop Sci.* 59, 1371–1373. doi: 10.2135/cropsci2018.10.0621le

McFarland, B. A., Alkhalifah, N., Bohn, M., Bubert, J., Buckler, E. S., Ciampitti, I., et al. (2020). Maize genomes to fields (G2F): 2014-2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. *BMC Res. Notes* 13:71. doi: 10.1186/s13104-020-4922-8

Melo, A. T. O., Bartaula, R., and Hale, I. (2016). GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics* 17:29. doi: 10.1186/s12859-016-0879-y

Messina, C. D., Technow, F., Tang, T., Totir, R., Gho, C., and Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). *Eur. J. Agron.* 100, 151–162. doi: 10.1016/j.eja.2018.01.007

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

Millet, E. J., Welcker, C., Kruijer, W., Negro, S., Coupel-Ledru, A., Nicolas, S. D., et al. (2016). Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios. *Plant Physiol.* 172, 749–764. doi: 10.1104/pp.16.00621

Mir, C., Zerjal, T., Combes, V., Dumas, F., Madur, D., Bedoya, C., et al. (2013). Out of America: tracing the genetic footprints of the global diffusion of maize. *Theor. Appl. Genet.* 126, 2671–2682. doi: 10.1007/s00122-013-2164-z

Montesinos-López, O. A., Montesinos-López, A., Tuberosa, R., Maccaferri, M., Sciara, G., Ammar, K., et al. (2019). Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. *Front. Plant Sci.* 10:1311. doi: 10.3389/fpls.2019.01311

Morais, P. P. P., Akdemir, D., Braatz de Andrade, L. R., Jannink, J., Fritsche-Neto, R., Borém, A., et al. (2020). Using public databases for genomic prediction of tropical maize lines. *Plant Breed.* 139, 697–707. doi: 10.1111/pbr.12827

Morosini, J. S., Mendonça, L., Lyra, D. H., Galli, G., Vidotti, M. S., and Fritsche-Neto, R. (2017). Association mapping for traits related to nitrogen use efficiency in tropical maize lines under field conditions. *Plant Soil* 421, 1–11. doi: 10.1007/s11104-017-3479-3

Negro, S. S., Millet, E. J., Madur, D., Bauland, C., Combes, V., Welcker, C., et al. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biol.* 19:318. doi: 10.1186/s12870-019-1926-4

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

Ramstein, G. P., Larsson, S. J., Cook, J. P., Edwards, J. W., Ersoz, E. S., Flint-Garcia, S., et al. (2020). Dominance effects and functional enrichments improve prediction of agronomic traits in hybrid maize. *Genetics* 215, 215–230. doi: 10.1534/genetics.120.303025

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., et al. (2017). Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol. Plant* 10, 1047–1064. doi: 10.1016/j.molp.2017.06.008

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

Rife, T. W., Graybosch, R. A., and Poland, J. A. (2018). Genomic analysis and prediction within a US public collaborative winter wheat regional testing nursery. *Plant Genome.* 11. doi: 10.3835/plantgenome2018.01.0004

Rincent, R., Kukn, E., Monod, H., Oury, F.-X., Rousset, M., Allard, V. et al. (2017). Optimization of multi-environment trials for genomic selection based on crop models. *Theor. Appl. Genet.* 130, 1735–1752. doi: 10.1007/s00122-017-2922-4

Rogers, A. R., Dunne, J. C., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I. A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3 (Bethesda)* 11:jkaa050. doi: 10.1093/g3journal/jkaa050

Sabadin, J. F. G. (2020). Haploid *Maize Seeds Prediction Using Deep Learning and Using Mock Reference Genomes for Genomic Predicion of Hybrids*. Thesis, College of Agriculture Luiz de Queiroz.

Sabadin, J. F. G., and Fritsche-Neto, R. (2020). Genome mock to predict single-crosses. *Mendeley Data V1*. doi: 10.17632/4nccgtcpgn.1

Sant'Ana, G. C., Espolador, F. G., Granato, Í. S. C., Mendonça, L. F., Fritsche-Neto, R., and Borém, A. (2020). Population structure analysis and identification of genomic regions under selection associated with low-nitrogen tolerance in tropical maize lines. *PLoS ONE* 15:e0239900. doi: 10.1371/journal.pone.0239900

Sparks, A. (2018). nasapower: a NASA POWER global meteorology, surface solar energy and climatology data client for R. *J. Open Source Softw.* 3:1035. doi: 10.21105/joss.01035

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.-L., et al. (2016). Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity (Edinb).* 116, 395–408. doi: 10.1038/hdy.2015.113

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

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

Torres, L. G., Rodrigues, M. C., Lima, N. L., Trindade, T. F. H., Silva, F. F., and Azevedo, C. F. (2018). Multi-trait multi-environment Bayesian model reveals G x E interaction for nitrogen use efficiency components in tropical maize. *PLoS ONE* 13:e0199492. doi: 10.1371/journal.pone.0199492

Varona, L., Legarra, A., Herring, W., and Vitezica, Z. G. (2018). Genomic selection models for directional dominance: an example for litter size in pigs. *Genet. Sel. Evol.* 50:1. doi: 10.1186/s12711-018-0374-1

Vidotti, M. S., Matias, F. I., Alves, F. C., Rodríguez, P. P., Beltran, G. A., Burguen,õ, J., et al. (2019). Maize responsiveness to *Azospirillum brasilense*: insights into genetic control, heterosis and genomic prediction. *PLoS ONE* 14:e0217571. doi: 10.1371/journal.pone.0217571

Wang, N., Wang, H., Zhang, A., Liu, Y., Yu, D., Hao, Z., et al. (2020). Genomic prediction across years in a maize doubled haploid breeding program to accelerate early-stage testcross testing. *Theor. Appl. Genet.* 133, 2869–2879. doi: 10.1007/s00122-020-03638-5

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., et al. (2020). How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. *Front. Plant Sci.* 11:592977. doi: 10.3389/fpls.2020.592977

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 (Bethesda)* 2, 1427–1436. doi: 10.1534/g3.112.003699

Zhang, X., Pérez-Rodríguez, P., Burgueño, J., Olsen, M., Buckler, E., Atlin, G., et al. (2017). Rapid cycling genomic selection in a multiparental tropical maize population. *G3 (Bethesda)* 7, 2315–2326. doi: 10.1534/g3.117.043141

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., et al. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS ONE* 9:e93017. doi: 10.1371/journal.pone.0093017

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 one of the authors JC.

frontiers in Plant Science

OPINION
published: 13 July 2021
doi: 10.3389/fpls.2021.681624

Strengthening Public Breeding Pipelines by Emphasizing Quantitative Genetics Principles and Open Source Data Management

Giovanny Covarrubias-Pazaran[1]*, *Johannes W. R. Martini*[2], *Michael Quinn*[1] and *Gary Atlin*[3]

[1] Excellence in Breeding Platform, Consultative Group for International Agricultural Research, Texcoco, Mexico, [2] Genetic Resources Program, International Maize and Wheat Improvement Center, Texcoco, Mexico, [3] Bill and Melinda Gates Foundation, Seattle, WA, United States

Keywords: breeding scheme optimization, quantitative genetics and breeding, genomic selection, open source, developing and transition countries

OPEN ACCESS

Edited by:
Sergio J. Ochatt,
INRA UMR1347 Agroécologie, France

Reviewed by:
Anil Adhikari,
Bayer Crop Science, United States
Brigid Meints,
Oregon State University,
United States

****Correspondence:***
Giovanny Covarrubias-Pazaran
cova_ruber@live.com.mx;
g.covarrubias@cgiar.org

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *16 March 2021*
Accepted: *18 June 2021*
Published: *13 July 2021*

Citation:
Covarrubias-Pazaran G, Martini JWR, Quinn M and Atlin G (2021) Strengthening Public Breeding Pipelines by Emphasizing Quantitative Genetics Principles and Open Source Data Management. Front. Plant Sci. 12:681624. doi: 10.3389/fpls.2021.681624

INTRODUCTION

The strategic goals of the "Consultative Group on International Agricultural Research" (CGIAR), which serves small-scale agricultural producers in the developing world, include the increase of nutrition and food security, the reduction of poverty, and the reduction of the "environmental footprint" of agricultural production systems (https://www.cgiar.org/how-we-work/strategy/). For each of these goals, progress can be made by breeding new crop varieties with increased productivity, stress resilience, nutritional value, and reduced requirement for fertilizer or agrochemicals.

Despite the great success of CGIAR breeding in the last decades, we posit that quantitative genetics principles must be more strongly emphasized in breeding strategies to keep pace with the accelerated demand and with changes in production conditions resulting in a growing demand for food, climate change and newly introduced breeding objectives -such as nutritional quality.

Traditionally, molecular breeding approaches focused on the identification of major genes, often for disease resistance, and the introgression of these alleles into elite material. This has been a fruitful strategy to prevent or mitigate production losses since disease resistances are essential traits for most target populations of environments (TPEs). However, the focus on major genes for disease resistances may also have slowed down genetic gain for yield in some programs. We advocate the redesign of breeding pipelines with a stronger orientation on quantitative genetics principles, optimizing the components of the "breeder's equation" to deliver a high selection response for quantitative traits like yield. Moreover, to improve the basis on which selection decisions are made, we propose an open-source breeding approach in which individual public and private institutions collaborate, align their activities, and share data to enhance efficiency for all participants.

We will briefly present the breeder's equation and highlight the terms that can be manipulated to increase genetic gain per time and per dollar invested. We will also present some guidelines recommended by the Excellence in Breeding (EiB) platform to optimize the selection response in a classical breeding scheme. We then discuss how genome-assisted prediction methods (genomic selection, GS) can be used for further optimization.

THE BREEDER'S EQUATION

In its simplest form, the breeder's equation for one trait, or even a composite of traits integrated in a selection index, states that the ***genetic gain*** per unit of time, expressed as the difference of the means of the (additive) genetic values before and after selection ($\Delta\mu$) divided by time t, is given by:

$$\frac{\Delta\mu}{t} = \frac{(h^2\ i\ \sigma_p)}{t}$$

Here, h^2 is the narrow sense *heritability*, i is the standardized selection differential, that is the *difference in phenotypic standard deviations* between the mean of the selected fraction and the mean of the initial population, and σ_p is the *phenotypic standard deviation* of the population before selection (modified from Lynch and Walsh, 1998). The cycle length t in years describes the time needed for one breeding "cycle" including recombination, evaluation, and selection of parents for the new set of crosses. The breeder's equation highlights the parameters that can be optimized to increase genetic gain per unit of time. We can increase the accuracy of our selection, h^2 for instance, by improving trial quality or increasing replication. We can increase the selection intensity i by selecting fewer individuals, or from a greater number of candidates (or both). Finally, we could reduce the cycle time t by shortening the time from cross to evaluation and to crosses of selected progeny (rapid generation advance).

RECOMMENDATIONS FOR THE DESIGN OF PROGRESSIVE BREEDING PIPELINES

The Excellence in Breeding (EiB) platform (https://excellenceinbreeding.org/) provides guidance to the CGIAR system and its national partners on the successful implementation of genomic prediction methods. EiB has proposed, as a first step, to optimize classical programs by addressing resource allocation in the light of the breeder's equation, which may sometimes require a radical redesign of the pipeline. The routine use of genomic selection to select parents is then implemented in a second iteration. EiB has modeled many of the breeding pipelines of CGIAR centers in detail and has evaluated a range of approaches to crossing, evaluation, and selection decisions in simulations. Some general recommendations are summarized below:

1) Formalize the breeding objective by defining market segments and corresponding product profiles describing the "ideal" product.

Point (1) guarantees that we clearly define in which direction we would like to breed. Moreover, market intelligence from a wide range of sources can be brought to bear on variety design (Cobb et al., 2019). We do not advocate for a particular methodology but emphasize the importance of investing resources in de- and refining the breeding goal. Client and market intelligence can be assembled from participatory plant breeding approaches (Witcombe et al., 1996; Ashby, 2009; Ragot et al., 2018) for subsistence-oriented systems, but product design for market-oriented cropping systems requires formal engagement with farmers, processors, and marketers to ensure that breeding objectives result in products that are both producible and marketable.

2) Form the crossing blocks out of small elite populations of 20–30[1] parents (avoid closely related individuals) and keep the crossing block as a mostly closed system. Use diversity measures and the variance of the traits defined in the product profile to monitor the diversity in the population over time.

Point (2) allows concentration on the "most elite" material (i.e., material with high breeding or genetic value) for our breeding objectives, which increases selection intensity (i). Moreover, a smaller effective population size avoids unnecessary crossing and testing, which saves resources. Experimental populations, theory and simulations show that a small number of elite individuals contain enough variance to avoid genetic bottlenecks in short and medium-term breeding time-horizons (Moose et al., 2004; Gaynor et al., 2017). This recommendation is linked to the breeder's equation by effectively managing the genetic variance and optimizing selection intensity.

3) The rate of new "diversity" injected into the pipeline each cycle should be low rather than high, which means parents of a cycle should be mainly chosen from the progeny of the previous cycle (recurrent selection strategy). New diversity (e.g., alleles conferring disease resistance) should be mainly injected in the form of donors of elite background with high-value haplotypes that do not currently exist in the population. This diversity must be carefully introduced to minimize linkage drag associated with new resistance alleles.

The restriction of the input of new diversity in point (3) is critical to the success of methods such as pedigree BLUP (Best Linear Unbiased Predictor) or genomic BLUP to improve the accuracy of selection of parents for the subsequent cycle. A certain degree of relatedness is required for these methods to be accurate. In addition, introgressing too many new parents can reduce the accuracy of quantitative genetics methods (Lynch and Walsh, 1998; Walsh and Lynch, 2018). When a recurrent selection strategy is used properly, almost any introgression would be a step backwards in terms of general performance and breeding value, and should only be used for special trait introgression or if genetic variance has been exhausted (Allier et al., 2020). This recommendation is linked to effectively managing the genetic variance in breeder's equation.

4) Formalize the crossing, evaluation, and selection decisions as variables in a process that is comprised of different stages (e.g., crossing blocks, nursery, early testing, late testing, etc.).

The formalization described in point (4) is required to apply selection criteria consistently and to characterize the breeding scheme more easily for simulations (point 5) and continuous improvement processes.

5) Changes in crossing, evaluation or selection procedures and resource allocations should be supported by simulations

[1] The number of elite parents is suggested for 30-year breeding time horizon of a classical program that takes between 3–5 years to recycle parents. In addition, the number of elite parents in the crossing block must be increased when adopting an aggressive GS scheme (recycling F1s) because the number of effective cohorts decreases drastically.

or experiments measuring the effect of the change on genetic gain while considering other influencing parameters, including costs.

It is critical that all the steps and processes used in the breeding pipeline be accurately costed, permitting simulation and modeling to be used to allocate resources to maximize the rate of genetic gain delivered per year and per dollar spent.

6) Use and document selection indices or independent culling to formalize the selection decisions when breeding for several quantitative traits simultaneously. The goal should be to make parent selection as objective and "data-driven" as possible, such that anyone having the underlying data can understand how the selection decision was made. All traits which are included in the selection decision should also be formally included in the recorded data and in the description of the selection criteria.

Selection indices allow application of selection criteria more consistently, can increase the selection intensity for several traits simultaneously and make use of genetic correlations between traits, if they are approximately known (Lynch and Walsh, 1998).

7) Use a data management and analytical system as a high priority to enable the analytical pipelines.

Adoption of analytical methods such as state of the art experimental design, spatial modeling to increase accuracy, and use of BLUP are all critical to acceleration of genetic gains. Organized, digitized data collection and storage and querying systems linking phenotypic, pedigree, and genotypic data are required to provide predictions routinely and rapidly. This recommendation is linked to all terms of the breeder's equation since better data management and analytics lead to more accurate selections, better management of diversity and in general to more accurate decisions.

8) According to the number of plots available and the breeding time-horizon, optimize the number of crosses and progeny per cross to maximize variation among and within families that can be selected.

The trade-off between allocating resources between number of families and family size will depend on factors like the number of traits included in the product profile, their genetic correlations and the time that we expect our breeding program to operate (longer periods benefit of putting more resources in the number of families and shorter breeding periods benefit of putting more resources in bigger families). We recommend the use of simulations to approach this question. This recommendation is linked to the breeder's equation by optimizing the selection intensity.

9) Parents for recycling should be selected from the first one or two testing stages of phenotyping yield (early recycling) to reduce cycle length. Also, breeders should avoid using the same parent repeatedly for several years in new crosses, which substantially lengthens the breeding cycle. Indeed, with emphasis on a short cycle time, selected progeny from a parent should always be preferred to the parent itself.

Shortening the breeding cycle while maintaining confidence in the selection of parents will often require reallocation of resources to improve data quality and quantity of the first and second testing stages of phenotyping.

10) Multiplication time (e.g., line generation, clonal propagation) should be reduced to the minimum possible (aiming for an overall cycle time as short as biology allows. For example, in seed crops that might be 2–3 years), leveraging new methodologies such as speed breeding, semi-autotrophic hydroponics, among others.

A successful example of renewing a traditional breeding pipeline at the International Rice Research Institute (IRRI) has been described at by Collard et al. (2019).

An overview, as well as a more detailed description of the different simulations supporting the recommendations above, can be found in the toolbox of EiB (https://excellenceinbreeding.org/toolbox). Once an aggressive classical breeding program with most of the features described above has been implemented, the adoption of genome-assisted prediction methods is recommended for parent selection. Implementation may follow the approach suggested below.

INCORPORATION OF GENOMIC SELECTION IN THE BREEDING STRATEGY

Much plant breeding literature on genomic selection (GS) focuses on predictive ability, especially the prediction of the performance of a selection candidate in the absence of any phenotypic data. Predicting the commercial performance of material that has not been phenotyped, which would mean that we substitute experiments with predictions, is an important application of GS, but it is not necessarily the most impactful one, especially not for small programs. The most important application of GS is the inference of the individual's genomic estimated breeding values (GEBV) from the phenotypes of its available relatives, for the purpose of selecting parents of the next cycle. In the context of population improvement, with the objective of maximizing genetic gain per year, we are not primarily interested in the phenotype of a selection candidate itself, but rather would like to know which candidates we should select as parents of new crosses to achieve the highest improvement in the new generation. The breeding value aims at capturing the improvement of the new generation when randomly crossing the line under consideration with other lines of the population (Mrode, 2014).

The first simple step in applying GS is therefore increasing accuracy by the use of the GEBV as the selection criterion, instead of EBV or phenotypes in isolation. This application can be incorporated into any breeding pipeline, usually at the agronomic testing stage, provided that genotypic data is available. Moreover, the resulting increase in accuracy can also give more freedom to reduce cycle time t, for instance by allowing parents to be selected from the first stage of agronomic testing (see point 9 above) due to the increased accuracy of surrogates of trait genetic merit compared to pure phenotypic information.

A second step could be the use of GS for sparse phenotyping. Sparse phenotyping means that not each genotype is tested in each environment, but some genotypes are tested at only a subset of locations. Such an approach can increase the accuracy of the estimation of genetic values by sampling from more environments, which again reduces the error resulting from genotype-by-environment (GxE) interaction. Moreover, sparse phenotyping can be used to increase the number of candidates tested which increases selection intensity. Both, increasing the number of environments and increasing the number of tested candidates can be approached by sparse testing subjected to fixed costs. GS helps to keep the data quality when reducing the data points and models including genotype-by-environment (GxE) effects can be of additional advantage (Jarquin et al., 2020).

A third application of GS is to recycle selection candidates as early as possible (e.g. nursery stage) based on their GEBVs, or genomic estimated genetic value (GEGV; additive plus non-additive effects). The training population should be formed by phenotypes generated from related candidates from the same program (not exotic diversity panels) phenotyped in previous seasons.

The fourth step in applying GS is to use genomic marker information to predict the crossing process, e.g., not only the expected performance of genotypes coming from a certain cross, but also the variability within a family of siblings. This can be used to optimize family sizes for different crosses, and to use predicted within family variance to maximize long-term gain.

For these points see for instance Cobb et al. (2019), Clark et al. (2013), Lehermeier et al. (2017), Gorjanc and Hickey (2018) and Henryon et al. (2019), Werner et al. (2020). Any of these steps can be incorporated independently, but the order proposed reflects an increasing level of complexity of the related logistics, and therefore may lead to a more successful implementation of GS.

OPEN SOURCE BREEDING

CGIAR centers together with NARs breeding centers form networks that phenotype and disseminate breeding materials that primarily originate from CGIAR centers. We envision an “open-source” breeding model that combines resources from different public and/or private partners for the benefit of all participants (intellectual property questions would need to be addressed to make a participation attractive for private partners). GS would permit the pooling of experimental data from different institutions that work within the same TPE. This would enable a better coverage of the TPE through a stronger testing network that shares (highly) related material. This way, CG centers, NARs and local companies could “borrow strength” from each other by sharing data on a central platform (Atlin and Jannink, 2010). A similar approach is currently used in dairy breeding, where the data are centrally processed and managed. In the context of public plant breeding, this would mean that the data from participating programs is jointly used to generate a stronger,

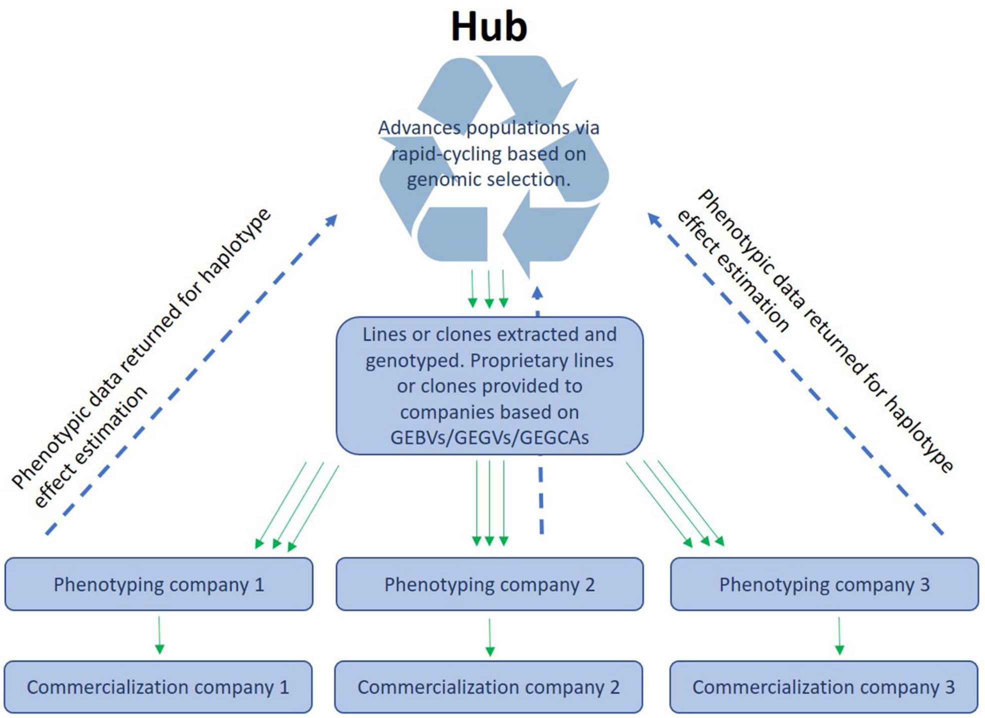

Hub
Advances populations via rapid-cycling based on genomic selection.
Phenotypic data returned for haplotype effect estimation
Phenotypic data returned for haplotype effect estimation
Lines or clones extracted and genotyped. Proprietary lines or clones provided to companies based on GEBVs/GEGVs/GEGCAs
Phenotyping company 1
Phenotyping company 2
Phenotyping company 3
Commercialization company 1
Commercialization company 2
Commercialization company 3

FIGURE 1 | Organization of open-source breeding: A hub receives the data (phenotypes and genotypes) from different programs or companies and makes the data available as training sets to enable the different ways of using genomic prediction (see main text).

more accurate prediction model than any single program could generate independently. A hub could then manage a source population and deliver lines or clones to local partners who could utilize the lines in a product development pipeline and give the experimental results back to the central data management unit (see **Figure 1**). Moreover, they could also use the lines as parents in their own pipelines. No CGIAR breeding networks have yet been formally constituted as open-source GS networks, but several have begun generating GEBVs for all new selection candidates and are therefore ready to implement the model with their national partners. The open-source GS network model has many advantages, including allowing breeding programs serving small-scale producers in the developing world to make selections and advance populations even when trials are lost to biotic or abiotic stress, or when disruptions such as a human pandemic hamper or prevent field testing, as happened in many breeding programs in 2020–2021. The open-source GS model will also permit highly efficient, two-stage rapid-cycle recurrent GS methods (Gaynor et al., 2017) that can reduce the breeding cycle to the biological limit imposed by the juvenility period of the species (time interval needed to move from seed to seed in seed crops may be 1 year or less but in tree species may be a couple of years) to be applied in the service of small-scale producers in Africa.

CONCLUSION

An efficient implementation of genomic prediction methods in CGIAR-NARs breeding programs (and maybe other publicly funded programs) depends on forming structured programs that follow certain design rules. Such programs must be outcome-oriented, with well-defined targets expressed in formal product profiles that guide selection decisions. We suggest that the first step in this process is to implement a classical breeding pipeline optimized based on quantitative genetics principles (reducing cycle time to the biological limit while increasing the accuracy of early testing and managing the genetic diversity at the proper program size). From there, the adoption of GS methods will be a natural extension guided by the breeder's equation. A first step would then be the use of GEBVs as selection criteria instead of phenotypic data in isolation. The breeding populations should be (almost) closed, using a relatively small number of elite parents. In the next steps, GS should be used to reduce evaluation costs while increasing the coverage of the TPE using sparse testing supported by marker data. Moreover, GS should be used to reduce the breeding cycles down to 1 year in a stepwise fashion for most crops if the phenotyping and selection methods are up to the challenge (data for all traits and use of indices is a pre-requisite for the most extreme use of GS). Simultaneously, it should be explored how "open source" breeding structures could be implemented in CGIAR-NARs networks, allowing small breeding programs to borrow strength from each other by incorporating the data generated by other programs working in the same crop but different regions with highly related material.

AUTHOR CONTRIBUTIONS

GC-P, JM, MQ, and GA wrote the manuscript and conceived the ideas of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the Bill and Melinda Gates Foundation (grant OPP1177070) and USAID.

ACKNOWLEDGMENTS

We thank to the Bill and Melinda Gates Foundation who funds the Excellence in Breeding Platform. We thank all the breeders across the CGIAR that have contributed to mature the ideas presented in this opinion by working one to one with the EiB platform.

REFERENCES

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. doi: 10.1186/s12864-020-6756-0

Ashby, J. A. (2009). "The impact of participatory plant breeding," in *Plant Breeding Farmer Participation*, eds S. Ceccarelli, E. P. Guimaraes, and E. Weltzein (Rome: FAO), 649–671.

Atlin, G. N., and Jannink, J. L. (2010). "Genomic selection breeding plans for maize hybrid development that use the haplotype as the selection unit," in *ASA-CSSA-SSSA International Annual Meetings* (Long Beach, CA).

Clark, S. A., Kinghorn, B. P., Hickey, J. M., and van der Werf, J. H. (2013). The effect of genomic information on optimal contribution selection in livestock breeding programs. *Genetics Sel. Evolut.* 45:44. doi: 10.1186/1297-9686-45-44

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., et al. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theor. Appl. Genet.* 132, 627–645. doi: 10.1007/s00122-019-03317-0

Collard, B. C., Gregorio, G. B., Thomson, M. J., Islam, M. R., Vergara, G. V., Laborte, A. G., et al. (2019). Transforming rice breeding: re-designing the irrigated breeding pipeline at the international rice research institute (IRRI). *Crop Breed. Genet. Genom.* 1, 1–19. doi: 10.20900/cbgg20190008

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

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

Henryon, M., Liu, H., Berg, P., Su, G., Nielsen, H. M., Gebregiwergis, G. T., et al. (2019). Pedigree relationships to control inbreeding in optimum-contribution selection realise more genetic gain than genomic relationships. *Genet. Sel. Evolut.* 51:39. doi: 10.1186/s12711-019-0475-5

Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W., et al. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3* 10, 2725–2739. doi: 10.1534/g3.120.401349

Lehermeier, C., Teyssèdre, S., and Schön, C. C. (2017). Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. *Genetics* 207, 1651–1661. doi: 10.1534/genetics.117.300403

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

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

Mrode, R. A. (2014). *Linear Models for the Prediction of Animal Breeding Values*. Oxfordshire: Cabi. doi: 10.1079/9781780643915.0000

Ragot, M., Bonierbale, M., and Weltzien, E. (2018). "From market demand to breeding decisions: a framework. Lima (Peru). CGIAR Gender and Breeding Initiative (No. 2)," in *GBIWorking Paper* (Lima).

Walsh, B., and Lynch, M. (2018). *Evolution and Selection of Quantitative Traits*. Sunderland, MA: Oxford University Press. doi: 10.1093/oso/9780198830870.001.0001

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

Witcombe, J. R., Joshi, A., Joshi, K. D., and Sthapit, B. R. (1996). Farmer participatory crop improvement. I. Varietal selection and breeding methods and their impact on biodiversity. *Exp. Agric.* 32, 445–460. doi: 10.1017/S001447970000380X

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 Plant Science

ORIGINAL RESEARCH
published: 15 July 2021
doi: 10.3389/fpls.2021.690059

Genetic Dissection of Grain Yield of Maize and Yield-Related Traits Through Association Mapping and Genomic Prediction

Juan Ma and Yanyong Cao*

Institute of Cereal Crops, Henan Academy of Agricultural Sciences, Zhengzhou, China

High yield is the primary objective of maize breeding. Genomic dissection of grain yield and yield-related traits contribute to understanding the yield formation and improving the yield of maize. In this study, two genome-wide association study (GWAS) methods and genomic prediction were made on an association panel of 309 inbred lines. GWAS analyses revealed 22 significant trait–marker associations for grain yield per plant (GYP) and yield-related traits. Genomic prediction analyses showed that reproducing kernel Hilbert space (RKHS) outperformed the other four models based on GWAS-derived markers for GYP, ear weight, kernel number per ear and row, ear length, and ear diameter, whereas genomic best linear unbiased prediction (GBLUP) showed a slight superiority over other modes in most subsets of the trait-associated marker (TAM) for thousand kernel weight and kernel row number. The prediction accuracy could be improved when significant single-nucleotide polymorphisms were fitted as the fixed effects. Integrating information on population structure into the fixed model did not improve the prediction performance. For GYP, the prediction accuracy of TAMs derived from fixed and random model Circulating Probability Unification (FarmCPU) was comparable to that of the compressed mixed linear model (CMLM). For yield-related traits, CMLM-derived markers provided better accuracies than FarmCPU-derived markers in most scenarios. Compared with all markers, TAMs could effectively improve the prediction accuracies for GYP and yield-related traits. For eight traits, moderate- and high-prediction accuracies were achieved using TAMs. Taken together, genomic prediction incorporating prior information detected by GWAS could be a promising strategy to improve the grain yield of maize.

Keywords: grain yield, genome-wide association study, trait-associated markers, prediction accuracy, fixed model

Edited by:
Valentin Wimmer,
KWS Saat, Germany

Reviewed by:
Armin Christoph Hölker,
KWS Saat, Germany
Eduardo Beche,
University of Missouri, United States

****Correspondence:***
Juan Ma
majuanjuan85@126.com

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *02 April 2021*
Accepted: *14 June 2021*
Published: *15 July 2021*

Citation:
Ma J and Cao Y (2021) Genetic Dissection of Grain Yield of Maize and Yield-Related Traits Through Association Mapping and Genomic Prediction.
Front. Plant Sci. 12:690059.
doi: 10.3389/fpls.2021.690059

INTRODUCTION

Maize serves as an important cereal and forage crop and plays an important role in sustaining global food security. Improvement of grain yield is a major and longstanding breeding goal for maize. Kernel number per ear (KNE) and thousand kernel weight (HKW) are the major components of grain yield per plant (GYP). Kernel number per row (KNR) and kernel row number (KRN) are the important components of the KNE. Ear length (EL) and ear diameter (ED) affect GYP in different degrees.

In general, compared to GYP, yield components and related traits are less affected by environments and have higher heritability, and therefore, can be directly used to facilitate the final yield of maize (Shi et al., 2017). Identifying loci associated with GYP and yield-related traits will contribute to understanding their basis and the correlations between them at a molecular level. In addition, the identification of important loci and genes involved will provide useful information for whole-genome selection of high-yield potential.

Using linkage mapping and genome-wide association study (GWAS), a large number of quantitative trait loci (QTLs) or single-nucleotide polymorphisms (SNPs) have been identified among different populations. For instance, under drought and heat environments, Millet et al. (2016) detected a large number of significant SNPs for the grain yield and the grain number using single-environment and multi-environment GWAS methods. Zhang et al. (2017) identified 23 QTLs and 25 significant SNPs for HKW, KRN, and KNR in recombinant inbred lines and an association panel of 240 maize inbred lines, and a stable locus (*PKS2*) influencing KRN, HKW, and kernel shapes was identified. Using an intermated B73 × Mo17 Syn10 doubled haploid population and a natural population, Zhang et al. (2020) detected 100 QTLs and 138 SNPs for GYP and yield-related traits and found that eight significant SNPs were co-located within intervals of seven QTLs. These studies enforce the complex of GYP and yield-related traits, which are governed by a mixture of many large-effect and small-effect genomic components.

Traditional marker-assisted selection (MAS) and marker-assisted recurrent selection (MARS) use only a few large-effect QTLs or markers, where efficient selections are made in maize breeding programs. Genomic selection (GS) uses whole genome-wide molecular markers to predict the breeding values of individuals. Therefore, it can capture both major and minor effect markers and is efficient for complex traits, especially for grain yield. GS has been shown to outperform MAS for grain yield and physiological traits in maize doubled haploid populations (Cerrudo et al., 2018), and for days to silking/anthesis and anthesis–silking interval in a nested association mapping population (Guo et al., 2021). Annual gain from GS outperformed that from MAS by 2-fold for winter wheat and approximately 3-fold for maize at a moderate accuracy (Heffner et al., 2010). Genetic gains of maize stover index and yield + stover index were 14–50% larger with GS than with MARS (Massman et al., 2013), which is consistent with the simulation results that GS produced up to 43% greater genetic gains than MARS for polygenic traits with low heritability (Bernardo and Yu, 2007). The primary advantages of GS over phenotypic selection are reflected in its low cost per cycle and the time for variety development. In maize advanced test-cross yield trials, GS reduced the cost by 32% over phenotype-based selection with similar selection gains (Beyene et al., 2019). With respect to cost reduction in maize breeding, breeders can test-cross half of all available lines, evaluate them in first-stage multi-environment trials, and then utilize the phenotypic data to predict the remaining half through GS (Crossa et al., 2017).

In GS, prediction models are established using prior phenotypic and marker data in a training population. The genomic estimated breeding value (GEBV) is predicted based on the marker effects estimated from the training population in a test population with genotypic data and no phenotypic data (Meuwissen et al., 2001). Many parametric methods such as GBLUP and Bayesian (Bayes) methods including Bayes A, Bayes B, Bayes C, and Bayes least absolute shrinkage and selection operator, semi-parametric models such as RKHS, and nonparametric methods have been developed to fit marker effects and predict phenotypes (Meuwissen et al., 2001; Gianola et al., 2006, 2011; Parmley et al., 2019; Sun et al., 2020). Multivariate models were developed to simultaneously consider information from multi-environment trials or multi-trait data (Burgueño et al., 2012; Montesinos-López et al., 2016; Schulthess et al., 2018). Previous studies showed that no single GS model had better performance compared with other models in all cases due to different backgrounds of training and testing populations, different traits, and different experimental designs (Pérez-Rodríguez et al., 2012; Ali et al., 2020). In maize, practical applications of GS have been widely demonstrated in many aspects including inbred line prediction (Zhao et al., 2012; Liu et al., 2019), hybrid performance prediction (Guo et al., 2019; Schrag et al., 2019; Li et al., 2020), and combining ability prediction (Riedelsheimer et al., 2012). These findings demonstrate the potential of GS helping in the selection of elite parents and hybrid combinations.

Both GWAS and GS use the same input datasets, including a phenotype dataset and a genotype dataset; thus, only additional analyses are required (Spindel et al., 2016). Several studies have discussed the advantages of combining GWAS and GS models that incorporate trait-associated markers (TAMs) detected by GWAS as random or fixed effects in GS models (Spindel et al., 2016; Bian and Holland, 2017; Herter et al., 2019; Liu et al., 2019; Rice and Lipka, 2019). However, the effects of TAM derived from different GWAS methods on prediction accuracy have rarely been reported. In this study, an association panel of 309 inbred lines was genotyped with 58,129 markers using genotyping-by-sequencing (GBS), and the performance of GYP, ear weight (EW), HKW, KNE, KNR, KRN, EL, and ED was evaluated in multi-environment trials. The main objectives of this study were to (1) identify significant SNPs for eight traits using two GWAS methods, (2) compare the prediction accuracies of different GS models, (3) investigate the prediction accuracy by treating significant SNPs and population structure as the fixed effects, and (4) evaluate the effects of TAMs derived from different GWAS methods on prediction accuracy.

MATERIALS AND METHODS

Plant Materials and Trial Designs

The panel consisted of 16 new selected inbred lines, 128 core germplasms of China, and 165 expired U.S. plant variety protection inbred lines, as previously reported (Ma et al., 2021). The panel was evaluated at four sites: Dancheng (33.646° N, 115.257° E), Yuanyang (35.012° N, 113.704° E), Yucheng (34.411° N, 116.274° E), and Sanya (18.381° N, 109.183° E) in 2017, and at one site (Yuanyang) in 2019. The field trial had a randomized complete block design with three replicates per genotype and environment. Entries were planted in two-row plots that were 3.75 m in length, 0.60 m spacing between rows, and 0.33 m spacing between plants.

Phenotyping and Analyses

Grain yield per plant, EW, HKW, KRN, KNR, EL, and ED were measured manually in three ears with good self-pollination for each genotype. KNE was calculated from KRN and KNR. Heritability at the per mean level and multi-environment ANOVA were calculated using QTL IciMapping v4.0 software (Meng et al., 2015). Pearson's correlation coefficient was calculated using the R package Performance Analytics. Best linear unbiased estimate (BLUE) values of each trait were calculated using QTL IciMapping v4.0 and were used as phenotypes for GWAS and GS analyses.

Association Mapping Analysis

The GBS genotypic data of the panel have been described in a previous study (Ma et al., 2021). Markers with minor allele frequencies (MAF) less than 5%, missing rates greater than 10%, and heterozygous rates greater than 10% were removed. Finally, 58,129 SNPs were adopted for GWAS. The kinship matrix was calculated using the Centered_IBS method in TASSEL v5.2.60 (Bradbury et al., 2007). The subgroups (*K*) were estimated using the Bayesian Markov chain Monte Carlo method in Structure v2.3.4 (Pritchard et al., 2000). The Q matrix of two subgroups ($K = 2$) was used to control the population structure as previously described (Ma et al., 2021). To reduce false associations, a single-locus method, namely, compressed mixed linear model (CMLM) (Zhang et al., 2010), and one multi-locus method, namely, fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al., 2016), were carried out using the GAPIT package (Lipka et al., 2012). The Q and K matrices were incorporated into both GWAS methods. A multiple testing correction is not required in multi-locus methods because all loci are estimated and tested simultaneously (Zhang et al., 2019b). Therefore, a less stringent *p*-value threshold of $1/58,129 = 1.72E-05$ was used to identify significant SNPs in the two GWAS methods. Other parameters were set default based on the GAPIT manual. Linear regression was used to calculate the phenotypic variation explained (PVE) of FarmCPU, whereas the PVE of CMLM was calculated using GAPIT. Candidate genes were scanned from 50 kb upstream to downstream of each significant locus using ANNOVAR (Wang et al., 2010).

Genomic Prediction

The prediction was done using GBLUP, Bayes A, Bayes B, Bayes C, and RKHS. Kernel averaging was used in the RKHS, and bandwidth parameters were set at 1/5M, 1/M, and 5/M, where M is the median squared Euclidean distance. Seven subset sizes of TAMs, that is, 100, 500, 1,000, 5,000, 10,000, 20,000, and 40,000 were selected according to the ranks of $-\log_{10}(p \text{ value})$ calculated by FarmCPU and CMLM based on BLUE values. The prediction accuracy of seven subsets was compared to that of all markers (58,129). For the eight traits, TAMs were all treated as the random effects (random model) in all GS models. For traits where significant SNPs ($p < 1.72E-05$) were detected, the significant SNPs were treated as the fixed effects and other remaining markers were treated as the random effects (fixed model). In the fixed model, one Q matrix (Q1) calculated using Structure was added into GBLUP and RKHS models as the fixed effects to evaluate the impact of population structure on the prediction accuracy. In addition, significant SNPs were all fitted as the random effects in RKHS to evaluate their potential application.

Randomized imputation was adopted for missing makers, according to the known genotype frequency. For each marker, individuals were coded as 2 (homozygous minor allele), 0 (homozygous major allele), and 1 (heterozygous). Recoding and imputation were carried out using the R software. Five GS models, TAMs, fixed model, random model, and fixed effects of Q matrix were performed using the R package, BGLR (Pérez and de los Campos, 2014). For all models, the length of the Gibbs chain was 12,000 iterations, with the first 3,000 samples discarded as burn-in. A 5-fold cross-validation scheme with 100 replicates was used to divide the association panel into training and testing sets. The mean correlation coefficient between GEBVs and BLUE values in the testing sets was used to estimate the accuracies of different GS models and different SNP densities.

RESULTS

Phenotypic Descriptions and Correlations

Descriptive statistics revealed that extensive phenotypic variations were observed in GYP and seven yield-related traits in the panel under different environments (**Supplementary Table 1**). The heritability for eight traits ranged from 0.59 (EW and KRN) to 0.77 (EL) (**Supplementary Table 2**). Significant and positive pairwise correlations were observed between different traits. GYP had high correlations with EW, KNE, KNR, and ED, moderate correlations with KRN and EL, and low correlations with HKW (**Supplementary Figure 1**). ANOVA across environments showed that the effects of genotype, environment, and genotype × environment interactions were significant ($p < 0.001$) for all traits (**Supplementary Table 2**). This showed that the association panel was highly affected by environments. Therefore, the BLUE values were used for GWAS and GS analyses.

Significant Trait Marker Associations and Their Prediction Accuracies

In total, 58,129 high-quality SNPs were used to perform GWAS for eight traits using BLUE values. FarmCPU and CMLM were used to control false associations for all traits. A total of 22 significant SNPs were identified with a *p*-value threshold of 1.72E−05, and the average PVE of all significant signals was 4.20% (**Table 1**). FarmCPU detected 17 association signals, which was higher than CMLM (7) (**Table 1**). One significant SNP each was found for GYP, EW, and HKW. Eight, eight, and four significant SNPs were detected for KRN, ED, and EL, respectively. One pleiotropic SNP (S3_62750920) was found between EW and ED. A SNP for ED, namely, S7_174915679, was detected using the two GWAS methods. The prediction accuracy of the significant SNPs was ranged from 0.26 to 0.45 using RKHS (**Supplementary Figure 2**).

TABLE 1 | Significant SNPs and candidate genes for grain yield and yield-related traits using two GWAS methods.

SNP name*	Trait§	*p* value	PVE†	Method#	Candidate gene
S3_53872814	GYP	1.68E−05	5.92	FarmCPU	Zm00001d040612
S3_62750920	EW	1.02E−05	5.98	FarmCPU	Zm00001d040748, Zm00001d040751
S1_47210783	HKW	1.56E−05	6.16	CMLM	Zm00001d028812
S1_10685412	KRN	1.43E−05	1.99	FarmCPU	Zm00001d027671
S1_179199207	KRN	3.38E−06	4.80	FarmCPU	Zm00001d031137, Zm00001d031138
S3_134708533	KRN	2.45E−06	1.44	FarmCPU	Zm00001d041715, Zm00001d041716
S4_135839291	KRN	2.79E−06	1.32	FarmCPU	Zm00001d050992
S4_234082607	KRN	1.54E−07	2.29	FarmCPU	Zm00001d053559
S4_86484873	KRN	1.08E−07	1.74	FarmCPU	Zm00001d050406, Zm00001d050409
S7_105588532	KRN	5.13E−08	7.38	FarmCPU	Zm00001d020310, Zm00001d020311
S8_145121832	KRN	2.46E−06	0	FarmCPU	Zm00001d011266
S1_69620597	EL	5.84E−07	1.35	FarmCPU	Zm00001d029416
S3_174651102	EL	2.11E−08	7.21	FarmCPU	Zm00001d042631, Zm00001d042632
S4_117775505	EL	7.78E−06	0.70	FarmCPU	Zm00001d050712, Zm00001d050714
S4_174433366	EL	4.36E−06	4.80	FarmCPU	Zm00001d051912
S1_233432714	ED	7.53E−06	10.26	FarmCPU	Zm00001d032659, Zm00001d032661
S2_118387989	ED	1.47E−05	5.43	CMLM	Zm00001d004568, Zm00001d004571
S2_118390724	ED	1.59E−05	5.39	CMLM	Zm00001d004568, Zm00001d004571
S2_118625688	ED	1.46E−05	5.43	CMLM	Zm00001d004572, Zm00001d004573
S2_118744667	ED	1.21E−05	5.54	CMLM	Zm00001d004573, Zm00001d004574
S3_62750920	ED	1.01E−05	5.64	CMLM	Zm00001d040748, Zm00001d040751
S7_13345176	ED	4.01E−06	3.77	FarmCPU	Zm00001d019027, Zm00001d019028
S7_174915679	ED	1.22E−05	5.54	CMLM	Zm00001d022310
S7_174915679	ED	3.94E−06	0.76	FarmCPU	Zm00001d022310

*Numbers before and after "_" represent chromosome and position, respectively.

§GYP, EW, HKW, KRN, EL, and ED are abbreviations of grain yield per plant, ear weight, thousand kernel weight, kernel row number, ear length, and ear diameter, respectively.

†PVE, phenotypic variation explained.

#CMLM, compressed mixed linear model; FarmCPU, fixed and random model Circulating Probability Unification.

Prediction Accuracy of Different Prediction Models

Five GS models were evaluated using seven subsets of TAMs derived from FarmCPU and CMLM. The prediction accuracies ranged from 0.10 to 0.84 and differed among prediction models and traits. Regardless of the marker effects, the prediction accuracy of RKHS using TAMs was the highest, followed by GBLUP, and Bayes B was the least for GYP, EW, and KNE (**Tables 2** and **3**, **Supplementary Table 3**). The prediction accuracies of the RKHS exceeded those of the other models by 3.85–68% for GYP and by 1.52–33.33% for KNE (**Table 2**, **Supplementary Table 3**). For EW, the percentage increase in accuracy of RKHS over the other four models using CMLM-derived TAMs ranged from 1.85 to 64%, whereas that of RKHS over the other models using FarmCPU-derived TAMs was large, with the percentage increase ranging from 26.09 to 210% (**Table 3**). Slight increases in the prediction accuracies of RKHS over the other models were also demonstrated in most subsets for KNR, EL, and ED (**Supplementary Tables 4–6**). For HKW, GBLUP was slightly superior to RKHS, Bayes A, Bayes B, and Bayes C (**Table 4**). In most of the marker sets, a small advantage of GBLUP over other models was also observed in KRN (**Table 5**).

Impact of Using Significant SNPs and Population Structure as Fixed Effects on Prediction Accuracy

The prediction accuracies of using significant SNPs and population structure as the fixed effects were evaluated in traits where significant SNPs were detected. In most of the TAM subsets, using 4–8 significant SNPs as the fixed effects improved the prediction accuracy by 1.43–40% and 1.37–22.41% for KRN and EL, respectively, when compared with the random model in all five models (**Table 5**, **Supplementary Table 5**). For GYP, EW, and HKW, the prediction accuracy did not change (or slightly decreased) when treating one significant SNP as a fixed effect compared to fitting all markers as the random effects in GBLUP and RKHS. However, the accuracy of the fixed model slightly increased or was similar to that of the random model in the three Bayes prediction models. For ED, the fixed model based on FarmCPU-derived markers improved the accuracy by 1.35—16%, whereas that of CMLM-derived markers had similar prediction performance as the random model in most cases. In general, the prediction accuracy could be improved when significant SNPs were fitted as the fixed effects.

To evaluate the effect of population structure on prediction accuracies, the Q matrix calculated using Structure was

TABLE 2 | Prediction accuracy of random model, fixed model, and population structure model based on trait-associated markers in five prediction models for grain yield per plant.

Model[*]	Scenario[§]	Prediction accuracy[#]							
		100[†]	500	1,000	5,000	10,000	20,000	40,000	58,129
Bayes A	CMLM-RAN	0.51 (0.08)	0.56 (0.07)	0.56 (0.08)	0.56 (0.08)	0.53 (0.08)	0.47 (0.09)	0.29 (0.11)	0.09 (0.12)
	FarmCPU-RAN	0.51 (0.08)	0.56 (0.07)	0.56 (0.07)	0.56 (0.08)	0.53 (0.08)	0.46 (0.09)	0.29 (0.11)	
	FarmCPU-FIX	0.52 (0.08)	0.56 (0.08)	0.56 (0.08)	0.57 (0.08)	0.54 (0.09)	0.47 (0.10)	0.33 (0.11)	
Bayes B	CMLM-RAN	0.48 (0.09)	0.53 (0.08)	0.53 (0.08)	0.54 (0.08)	0.51 (0.09)	0.44 (0.09)	0.26 (0.11)	0.08 (0.12)
	FarmCPU-RAN	0.48 (0.09)	0.54 (0.08)	0.53 (0.08)	0.54 (0.08)	0.51 (0.09)	0.44 (0.09)	0.25 (0.11)	
	FarmCPU-FIX	0.49 (0.09)	0.54 (0.08)	0.54 (0.08)	0.56 (0.08)	0.53 (0.09)	0.45 (0.10)	0.32 (0.12)	
Bayes C	CMLM-RAN	0.50 (0.09)	0.55 (0.07)	0.55 (0.08)	0.56 (0.08)	0.53 (0.08)	0.46 (0.09)	0.28 (0.12)	0.09 (0.12)
	FarmCPU-RAN	0.50 (0.09)	0.55 (0.07)	0.55 (0.07)	0.56 (0.08)	0.53 (0.08)	0.46 (0.09)	0.28 (0.11)	
	FarmCPU-FIX	0.51 (0.09)	0.56 (0.08)	0.57 (0.08)	0.57 (0.08)	0.53 (0.09)	0.46 (0.10)	0.33 (0.11)	
GBLUP	CMLM-RAN	0.52 (0.08)	0.57 (0.07)	0.57 (0.08)	0.59 (0.08)	0.56 (0.09)	0.49 (0.09)	0.30 (0.11)	0.10 (0.12)
	FarmCPU-RAN	0.52 (0.08)	0.57 (0.07)	0.57 (0.07)	0.59 (0.08)	0.55 (0.09)	0.48 (0.09)	0.30 (0.11)	
	FarmCPU-FIX	0.52 (0.08)	0.57 (0.07)	0.57 (0.08)	0.57 (0.08)	0.53 (0.09)	0.46 (0.10)	0.33 (0.12)	
	FarmCPU-FIX-PS	0.52 (0.08)	0.57 (0.08)	0.57 (0.08)	0.57 (0.08)	0.53 (0.09)	0.46 (0.10)	0.32 (0.12)	
RKHS	CMLM-RAN	0.54 (0.09)	0.62 (0.07)	0.61 (0.08)	0.62 (0.08)	0.59 (0.09)	0.54 (0.10)	0.42 (0.12)	0.32 (0.14)
	FarmCPU-RAN	0.54 (0.09)	0.62 (0.07)	0.61 (0.08)	0.62 (0.08)	0.59 (0.09)	0.54 (0.10)	0.42 (0.12)	
	FarmCPU-FIX	0.54 (0.08)	0.61 (0.08)	0.61 (0.08)	0.61 (0.08)	0.57 (0.09)	0.52 (0.10)	0.42 (0.11)	
	FarmCPU-FIX-PS	0.54 (0.09)	0.61 (0.08)	0.61 (0.08)	0.61 (0.08)	0.57 (0.09)	0.52 (0.10)	0.42 (0.11)	

[*] *GBLUP, genomic best linear unbiased prediction; RKHS, reproducing kernel Hilbert space.*

[§] *CMLM-RAN and FarmCPU-RAN, traits-associated markers from compressed mixed linear model (CMLM) and fixed and random model Circulating Probability Unification (FarmCPU) are treated as random effects; FarmCPU-FIX, significant SNPs ($p < 1.72E-05$) are treated as the fixed effects and other remaining markers are treated as the random effects (fixed model); FarmCPU-FIX-PS, the Q matrix is treated as fixed effect in the fixed model.*

[†] *100–40,000, the number of trait-associated markers.*

[#] *Prediction accuracy is represented by mean and standard deviation in brackets.*

incorporated into the fixed model in GBLUP and RKHS. For GYP, EW, and KRN, the accuracy did not change when the Q matrix was included as a fixed effect in most cases of RKHS and GBLUP (**Tables 2**, **3**, **5**). For HKW, the population structure had no effect on accuracies in RKHS, whereas the accuracy decreased by 0.01–0.05 when the Q matrix was used in the GBLUP model. For EL, the accuracy reduced by 0.02–0.07 at 500–20,000 TAMs when population structure was added into the GBLUP fixed model. For ED, the accuracy improved by 0.02 at 100 and 40,000 CMLM-derived TAMs and decreased by 0.05 at 40,000 FarmCPU-derived TAMs when the Q matrix was added in GBLUP, and the accuracy was same or slightly decreased in the remaining scenarios.

Effect of Different GWAS Methods on Prediction Accuracy

For GYP, the prediction accuracies of TAMs derived from CMLM and FarmCPU were compared in the five models (**Table 2**), regardless of the random or fixed models. For EL and ED, the prediction accuracy of 100 TAMs by FarmCPU was 2.74–5.97% higher than that by CMLM in the five models. For the other subsets, the prediction accuracies of CMLM-derived markers were 8.22–42.11% higher than those of FarmCPU-derived markers in EL and ED (**Supplementary Tables 5**, **6**). For the other five traits, the prediction accuracies of CMLM-TAMs were consistently superior to those of FarmCPU-TAMs across all subsets in the five models. In particular, the increase

TABLE 3 | Prediction accuracy of random model, fixed model, and population structure model based on trait-associated markers in five prediction models for ear weight.

Model[*]	**Scenario**[§]	**Prediction accuracy**[#]							
		100[†]	**500**	**1,000**	**5,000**	**10,000**	**20,000**	**40,000**	**58,129**
Bayes A	CMLM-RAN	0.54 (0.09)	0.58 (0.08)	0.57 (0.08)	0.55 (0.09)	0.50 (0.09)	0.44 (0.10)	0.27 (0.12)	
	FarmCPU-RAN	0.20 (0.11)	0.12 (0.11)	0.15 (0.10)	0.19 (0.11)	0.18 (0.11)	0.14 (0.12)	0.10 (0.12)	0.09 (0.12)
	FarmCPU-FIX	0.19 (0.12)	0.14 (0.11)	0.18 (0.11)	0.20 (0.11)	0.19 (0.11)	0.16 (0.11)	0.12 (0.12)	
Bayes B	CMLM-RAN	0.51 (0.09)	0.55 (0.08)	0.54 (0.08)	0.53 (0.09)	0.48 (0.09)	0.41 (0.10)	0.25 (0.12)	
	FarmCPU-RAN	0.16 (0.11)	0.13 (0.11)	0.16 (0.11)	0.18 (0.11)	0.17 (0.12)	0.14 (0.12)	0.11 (0.12)	0.09 (0.12)
	FarmCPU-FIX	0.15 (0.11)	0.13 (0.11)	0.18 (0.11)	0.20 (0.12)	0.19 (0.12)	0.16 (0.12)	0.12 (0.12)	
Bayes C	CMLM-RAN	0.53 (0.09)	0.57 (0.08)	0.57 (0.08)	0.55 (0.09)	0.50 (0.09)	0.43 (0.10)	0.27 (0.12)	
	FarmCPU-RAN	0.23 (0.11)	0.17 (0.11)	0.18 (0.11)	0.19 (0.11)	0.18 (0.11)	0.14 (0.12)	0.11 (0.12)	0.09 (0.12)
	FarmCPU-FIX	0.20 (0.12)	0.16 (0.11)	0.20 (0.11)	0.20 (0.11)	0.19 (0.11)	0.15 (0.12)	0.12 (0.12)	
GBLUP	CMLM-RAN	0.54 (0.09)	0.59 (0.08)	0.58 (0.08)	0.58 (0.09)	0.53 (0.09)	0.46 (0.10)	0.29 (0.12)	0.12 (0.12)
	FarmCPU-RAN	0.20 (0.12)	0.19 (0.11)	0.23 (0.11)	0.22 (0.11)	0.20 (0.11)	0.16 (0.12)	0.12 (0.12)	
	FarmCPU-FIX	0.18 (0.12)	0.17 (0.11)	0.20 (0.11)	0.20 (0.11)	0.19 (0.12)	0.16 (0.12)	0.12 (0.12)	
	FarmCPU-FIX-PS	0.17 (0.12)	0.16 (0.11)	0.19 (0.11)	0.20 (0.11)	0.19 (0.12)	0.15 (0.12)	0.12 (0.12)	
RKHS	CMLM-RAN	0.55 (0.09)	0.62 (0.08)	0.61 (0.08)	0.61 (0.09)	0.57 (0.09)	0.52 (0.11)	0.41 (0.13)	0.31 (0.14)
	FarmCPU-RAN	0.29 (0.13)	0.33 (0.13)	0.37 (0.12)	0.37 (0.13)	0.34 (0.13)	0.32 (0.14)	0.31 (0.14)	
	FarmCPU-FIX	0.28 (0.14)	0.28 (0.13)	0.31 (0.13)	0.37 (0.13)	0.36 (0.13)	0.33 (0.14)	0.31 (0.14)	
	FarmCPU-FIX-PS	0.27 (0.13)	0.27 (0.13)	0.31 (0.13)	0.36 (0.13)	0.36 (0.13)	0.33 (0.14)	0.31 (0.14)	

[*] *GBLUP, genomic best linear unbiased prediction; RKHS, reproducing kernel Hilbert space.*

[§] *CMLM-RAN and FarmCPU-RAN, traits-associated markers from compressed mixed linear model (CMLM) and fixed and random model Circulating Probability Unification (FarmCPU) are treated as random effects; FarmCPU-FIX, significant SNPs ($p < 1.72E-05$) are treated as the fixed effects and other remaining markers are treated as the random effects (fixed model); FarmCPU-FIX-PS, the Q matrix is treated as the fixed effect in the fixed model.*

[†] *100–40,000, the number of trait-associated markers.*

[#] *Prediction accuracy is represented by mean and standard deviation in brackets.*

in prediction accuracies for CMLM-TAMs over FarmCPU-TAMs was large in EW, with the percentage increase ranging from 32.26 to 383.33% across all scenarios (**Table 3**). With respect to TAMs, moderate and high prediction accuracies were achieved in five prediction models for the eight traits. The optimum number of TAMs for prediction differed greatly among the eight traits, two GWAS methods, and five GS models. These results indicate that it is necessary to determine the optimum SNP information that can represent sufficient variations to achieve high prediction accuracies for each trait before their application in GS breeding. Compared to all SNPs, higher prediction accuracies were achieved using TAMs in most scenarios. This indicates that TAMs could effectively improve the prediction accuracies of GYP and yield-related traits.

DISCUSSION

Genomic selection is a promising breeding method with the aim of accelerating the speed and efficiency of breeding processes. In contrast, GWAS is used to identify QTLs or genes that underlie important traits for breeding. They seek to model the different aspects of the genetic architecture of traits and have complementary advantages (Bian and Holland, 2017). Previous studies have shown the effectiveness of the GS method using important loci for target traits identified by

TABLE 4 | Prediction accuracy of random model, fixed model, and population structure model based on trait-associated markers in five prediction models for thousand kernel weight.

Model[*]	**Scenario**[§]	**Prediction accuracy**[#]							
		100[†]	**500**	**1,000**	**5,000**	**10,000**	**20,000**	**40,000**	**58,129**
Bayes A	CMLM-RAN	0.65 (0.06)	0.72 (0.06)	0.72 (0.05)	0.70 (0.05)	0.67 (0.06)	0.60 (0.07)	0.41 (0.08)	
	FarmCPU-RAN	0.55 (0.08)	0.58 (0.07)	0.59 (0.07)	0.58 (0.07)	0.54 (0.07)	0.48 (0.08)	0.35 (0.08)	0.20 (0.09)
	CMLM-FIX	0.66 (0.06)	0.72 (0.05)	0.73 (0.05)	0.71 (0.05)	0.67 (0.06)	0.60 (0.07)	0.42 (0.09)	
Bayes B	CMLM-RAN	0.63 (0.06)	0.70 (0.06)	0.71 (0.05)	0.68 (0.06)	0.64 (0.06)	0.57 (0.07)	0.39 (0.08)	0.21 (0.09)
	FarmCPU-RAN	0.53 (0.08)	0.56 (0.07)	0.58 (0.08)	0.56 (0.07)	0.52 (0.07)	0.46 (0.08)	0.34 (0.09)	
	CMLM-FIX	0.64 (0.06)	0.71 (0.05)	0.71 (0.05)	0.69 (0.06)	0.66 (0.06)	0.58 (0.07)	0.41 (0.09)	
Bayes C	CMLM-RAN	0.65 (0.06)	0.72 (0.06)	0.72 (0.05)	0.70 (0.05)	0.67 (0.06)	0.60 (0.07)	0.41 (0.08)	0.20 (0.09)
	FarmCPU-RAN	0.55 (0.08)	0.57 (0.07)	0.59 (0.07)	0.58 (0.07)	0.54 (0.07)	0.48 (0.08)	0.35 (0.08)	
	CMLM-FIX	0.66 (0.06)	0.72 (0.05)	0.73 (0.05)	0.71 (0.05)	0.67 (0.06)	0.60 (0.07)	0.42 (0.09)	
GBLUP	CMLM-RAN	0.67 (0.06)	0.73 (0.05)	0.73 (0.05)	0.71 (0.05)	0.68 (0.06)	0.60 (0.07)	0.40 (0.08)	0.20 (0.09)
	FarmCPU-RAN	0.56 (0.08)	0.60 (0.07)	0.60 (0.07)	0.58 (0.07)	0.54 (0.07)	0.48 (0.08)	0.34 (0.08)	
	CMLM-FIX	0.67 (0.06)	0.73 (0.05)	0.73 (0.05)	0.71 (0.05)	0.67 (0.06)	0.60 (0.07)	0.42 (0.09)	
	CMLM-FIX-PS	0.66 (0.06)	0.72 (0.05)	0.72 (0.05)	0.69 (0.05)	0.64 (0.06)	0.55 (0.07)	0.39 (0.09)	
RKHS	CMLM-RAN	0.66 (0.06)	0.72 (0.05)	0.72 (0.05)	0.69 (0.06)	0.65 (0.06)	0.56 (0.07)	0.37 (0.08)	0.24 (0.08)
	FarmCPU-RAN	0.54 (0.08)	0.58 (0.07)	0.58 (0.07)	0.55 (0.07)	0.51 (0.07)	0.45 (0.08)	0.33 (0.08)	
	CMLM-FIX	0.66 (0.06)	0.72 (0.05)	0.72 (0.05)	0.69 (0.06)	0.64 (0.06)	0.55 (0.07)	0.39 (0.09)	
	CMLM-FIX-PS	0.66 (0.06)	0.72 (0.05)	0.72 (0.05)	0.69 (0.05)	0.64 (0.06)	0.55 (0.07)	0.39 (0.09)	

[*] GBLUP, genomic best linear unbiased prediction; RKHS, reproducing kernel Hilbert space.

[§] CMLM-RAN and FarmCPU-RAN, traits-associated markers from compressed mixed linear model (CMLM) and fixed and random model Circulating Probability Unification (FarmCPU) are treated as the random effects; CMLM-FIX, significant SNPs ($p < 1.72E-05$) are treated as the fixed effects and other remaining markers are treated as the random effects (fixed model); CMLM-FIX-PS, the Q matrix is treated as the fixed effect in the fixed model.

[†] 100–40,000, the number of trait-associated markers.

[#] Prediction accuracy is represented by mean and standard deviation in brackets.

GWAS (Bian and Holland, 2017; Liu et al., 2019; Rice and Lipka, 2019). In this study, we demonstrated the potential of incorporating prior information for grain yield and seven yield-related traits explored by GWAS into GS in a maize association panel.

Prediction models are the major factors that affect the prediction accuracy of different traits. In this study, GBLUP, Bayes A, Bayes B, Bayes C, and RKHS were adopted to compare the prediction accuracies of eight traits based on GWAS-derived markers. The advantage of RKHS over the other four models was demonstrated using GYP, EW, KNE, KNR, EL, and ED in most TAM subsets, which was in line with many studies on maize, wheat, barley, and *Arabidopsis thaliana* (González-Camacho et al., 2012; Heslot et al., 2012; Pérez-Rodríguez et al., 2012; Liu et al., 2018; Li et al., 2020). RKHS, as one of the semiparametric methods, does not need to make most of the assumptions on the relationship between phenotype and genotype as do parametric models and was found to have the potential for capturing the total genetic effects from real data (Gianola et al., 2006; Gianola and van Kaam, 2008). The inferior performance of the RKHS over other models has also been reported in maize kernel oil traits

TABLE 5 | Prediction accuracy of random model, fixed model, and population structure model based on trait-associated markers in five prediction models for kernel row number.

Model*	Scenario§	Prediction accuracy#							
		100†	500	1,000	5,000	10,000	20,000	40,000	58,129
Bayes A	CMLM-RAN	0.71 (0.06)	0.77 (0.05)	0.79 (0.05)	0.78 (0.05)	0.76 (0.06)	0.69 (0.07)	0.52 (0.10)	0.34 (0.12)
	FarmCPU-RAN	0.70 (0.05)	0.68 (0.07)	0.64 (0.07)	0.57 (0.10)	0.55 (0.09)	0.49 (0.10)	0.41 (0.11)	
	FarmCPU-FIX	0.70 (0.05)	0.73 (0.06)	0.73 (0.08)	0.69 (0.07)	0.67 (0.07)	0.63 (0.08)	0.57 (0.09)	
Bayes B	CMLM-RAN	0.69 (0.07)	0.75 (0.05)	0.78 (0.05)	0.76 (0.06)	0.74 (0.06)	0.66 (0.08)	0.50 (0.10)	0.35 (0.12)
	FarmCPU-RAN	0.68 (0.05)	0.69 (0.06)	0.66 (0.07)	0.56 (0.09)	0.53 (0.10)	0.47 (0.11)	0.40 (0.11)	
	FarmCPU-FIX	0.69 (0.06)	0.73 (0.06)	0.73 (0.06)	0.68 (0.07)	0.66 (0.07)	0.62 (0.08)	0.56 (0.09)	
Bayes C	CMLM-RAN	0.70 (0.06)	0.77 (0.05)	0.79 (0.05)	0.78 (0.05)	0.76 (0.06)	0.69 (0.07)	0.52 (0.10)	0.35 (0.11)
	FarmCPU-RAN	0.69 (0.05)	0.69 (0.06)	0.64 (0.07)	0.56 (0.10)	0.55 (0.09)	0.48 (0.10)	0.41 (0.11)	
	FarmCPU-FIX	0.70 (0.05)	0.74 (0.06)	0.73 (0.06)	0.69 (0.07)	0.67 (0.07)	0.63 (0.08)	0.57 (0.09)	
GBLUP	CMLM-RAN	0.72 (0.06)	0.77 (0.05)	0.80 (0.05)	0.79 (0.05)	0.76 (0.06)	0.70 (0.07)	0.53 (0.10)	0.36 (0.12)
	FarmCPU-RAN	0.70 (0.05)	0.67 (0.07)	0.63 (0.08)	0.56 (0.10)	0.56 (0.10)	0.50 (0.11)	0.42 (0.11)	
	FarmCPU-FIX	0.70 (0.05)	0.73 (0.06)	0.73 (0.06)	0.69 (0.07)	0.67 (0.07)	0.63 (0.08)	0.57 (0.09)	
	FarmCPU-FIX-PS	0.71 (0.05)	0.73 (0.06)	0.73 (0.06)	0.69 (0.07)	0.67 (0.07)	0.63 (0.08)	0.57 (0.09)	
RKHS	CMLM-RAN	0.70 (0.06)	0.77 (0.05)	0.79 (0.05)	0.77 (0.06)	0.75 (0.06)	0.67 (0.07)	0.51 (0.09)	0.39 (0.10)
	FarmCPU-RAN	0.70 (0.05)	0.65 (0.07)	0.62 (0.08)	0.56 (0.09)	0.54 (0.09)	0.49 (0.10)	0.43 (0.10)	
	FarmCPU-FIX	0.71 (0.05)	0.72 (0.06)	0.72 (0.06)	0.67 (0.07)	0.65 (0.08)	0.61 (0.08)	0.56 (0.09)	
	FarmCPU-FIX-PS	0.71 (0.05)	0.72 (0.06)	0.72 (0.06)	0.67 (0.07)	0.65 (0.08)	0.61 (0.08)	0.56 (0.09)	

**GBLUP, genomic best linear unbiased prediction; RKHS, reproducing kernel Hilbert space.*

§CMLM-RAN and FarmCPU-RAN, traits-associated markers from compressed mixed linear model (CMLM) and fixed and random model Circulating Probability Unification (FarmCPU) are treated as the random effects; FarmCPU-FIX, significant SNPs (p < 1.72E−05) are treated as the fixed effects and other remaining markers are treated as the random effects (fixed model); FarmCPU-FIX-PS, the Q matrix is treated as the fixed effect in the fixed model.

†100–40,000, the number of trait-associated markers.

#Prediction accuracy is represented by mean and standard deviation in brackets.

(Hao et al., 2019) and cotton fiber quality traits (Islam et al., 2020). In this study, GBLUP showed a slight advantage over RKHS and the other models using TAMs for HKW and KRN. If additivity has a major effect, RKHS produces a similar performance as other methods, whereas if non-additive effects are present, it has a better prediction accuracy (Morota and Gianola, 2014). Although no single model was consistently performing better in all scenarios, RKHS could be the best choice when the computation time and prediction accuracy were comprehensively considered.

Except for GYP, the prediction accuracy of TAMs produced by CMLM was consistently higher than that by FarmCPU. In multiple species, FarmCPU outperformed CMLM and other methods by controlling the inflation of p values, identifying newly associated SNPs, and overlapping with the reported loci (Liu et al., 2016). CMLM and FarmCPU use different strategies to solve the confounding problem and improve statistical power for the mixed linear model methods (Zhang et al., 2010; Liu et al., 2016), which results in different marker information. Different markers, marker distributions, MAF, and multicollinearity might show the discrepancy in accuracies of the two GWAS methods. Except for EW, moderate and high accuracies were displayed in five models using FarmCPU-derived TAMs for GYP and other traits, which were high enough to make efficient predictions. GS can remarkably accelerate genetic gains by shortening the breeding cycle even if moderate accuracies are achieved (Heffner et al., 2010).

Genome-wide association study is a rapid and effective method for identifying genetic variations in important germplasms. Based on the prior knowledge of the underlying genetic architecture detected by GWAS, the advantage of integrating GWAS with GS was identified in our association panel. Our results showed that subsets of TAMs that treated significant SNP as the fixed effects or random effects could improve the prediction accuracies of GYP and yield-related traits compared with all markers. This was similar to the results of the studies by Liu et al. (2020) and Yuan et al. (2019), who reported that the prediction accuracy of marker trait-associated SNPs was higher than that of all markers or random genome-wide SNPs for maize grain yield, flowering time, and *Fusarium* ear rot resistance. The study by Lozada et al. (2019) proved that wheat yield achieved higher accuracies using three subsets of associated markers that were selected from GWAS in training populations compared with all markers. Compared with GS without marker selection by GWAS, TAMs as the random effects in GS increased the prediction accuracies, regardless of which TAMs were selected from in the full dataset or training set (Cericola et al., 2017; Liu et al., 2019; Ali et al., 2020). In most cases, the prediction accuracy was the highest at 100–5,000 TAMs and then decreased as the number of markers increased for the eight traits. A similar trend was observed in wheat grain yield based on GWAS-derived markers (Lozada et al., 2019). The decreased trend of the prediction accuracy was also found in many cases where evenly distributed SNPs were used and three examples where randomly selected markers were used in rice (Spindel et al., 2015). Higher marker density caused a lower prediction accuracy if significant SNPs were included, but resulted in a higher accuracy if significant SNPs were excluded for simple traits that were controlled by one or several genes with the large effects (Zhang et al., 2019a). The multicollinearity and complexity of GS models for the estimation of GEBVs became severe when an increasing number of markers were used (Ali et al., 2020), which might decrease the prediction accuracy. The smaller number of TAMs that benefited higher accuracies could be helpful to lower the costs of genotyping in GS-assisted breeding. In general, GS based on GWAS results from the full panel set could help to improve the prediction accuracies, although the "inside trading" effects lead to inflated values (Arruda et al., 2016).

In this study, treating one or several significant SNPs as the fixed effects in GS models resulted in higher accuracies in most cases, compared with those with only the random effects, which was in accordance with the trends in accuracy improvement shown in maize, wheat, and rice (Arruda et al., 2016; Spindel et al., 2016; Herter et al., 2019; Odilbekov et al., 2019). The incorporation of large-effect QTL or SNPs as the fixed effects was also a promising strategy to improve the prediction accuracy of GS (Bernardo, 2014; Herter et al., 2019). A slightly decreased accuracy was observed in the fixed model of GBLUP and RKHS for GYP, EW, and HKW. A similar result was also revealed in wheat yield stability using GBLUP (Sehgal et al., 2020). Except for HKW, the genetic architecture of GYP, EW, and yield stability was complex and hard to capture, which was supported by the fact that less robust SNPs with low phenotypic variation were identified. These could lead to the results obtained for these traits.

Integrating information on population structure into fixed models did not improve prediction performance and, in some cases, slightly decreased the accuracies. Similar results were found in the study by Rio et al. (2019); when taking genetic structure into account, the prediction accuracy of maize grain yield, grain moisture, yield index, and male flowering did not improve compared to standard GBLUP. However, Liu et al. (2019) showed that taking three principal components as the fixed effects in the random model could slightly improve the prediction accuracy. In fact, the impact of population structure on GS accuracy depends on many factors such as *a priori* indicators, prediction strategies, allele effects, allele frequencies between groups, the features of traits, and populations (Guo et al., 2014; Liu et al., 2019; Rio et al., 2019). Extended models that consider this information will guarantee high accuracies of GEBV.

The major limitation of incorporating TAMs into GS models depended on the accuracy of GWAS results. Marker selection strategies based on *p* values or marker effects might produce an improper marker set with low accuracies if the GWAS was incorrect (Jeong et al., 2020). GWAS results from the full data set that included the training set and testing sets might produce an overfitted markers set. In real GS-assisted breeding projects, the training set is used to conduct prediction models and predict other breeding populations that only have genotypes. Further investigation is needed in order to validate the application prospect of GS based on prior information from the GWAS results.

Despite these limitations, the combination of GWAS and GS offers an effective means for germplasm screening of traits with low heritability where, for instance, a 1% increase in prediction accuracy could improve genetic gains (Rice and Lipka, 2019). Furthermore, continued enlargement of the association panel by incorporating new fixed effects and high-quality phenotypic data from multi-environment trials is expected to improve the accuracy of GEBV. Besides, the marker information and training population will be used to obtain an optimum breeding design and improve genetic gains through reducing costs. Recently, GMStool is developed to present the best prediction model with the optimal marker set based on GWAS results (Jeong et al., 2020), which provides a useful tool for breeders. As GBS, SNP array technology, and other high-output genotyping strategies arise, the genotyping costs are likely to continue to decrease, whereas the phenotyping costs are usually steady or increasing (Spindel et al., 2015). Therefore, the combination of GWAS and GS will become a cost-effective method for selecting high-yield germplasms in maize and other species.

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.

AUTHOR CONTRIBUTIONS

JM collected phenotypic data, performed GWAS and GS analyses, and wrote the manuscript. YC provided help for the phenotypic measurement. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the Science and Technology Project of Henan Province (192102110008), the Science-Technology Foundation for Outstanding Young Scientists of Henan Academy of Agricultural Sciences (2020YQ04), and the Fund for Distinguished Young Scholars from Henan Academy of Agricultural Sciences (2020JQ01).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.690059/full#supplementary-material

REFERENCES

Ali, M., Zhang, Y., Rasheed, A., Wang, J., and Zhang, L. (2020). Genomic prediction for grain yield and yield-related traits in Chinese winter wheat. *Int. J. Mol. Sci.* 21:1342. doi: 10.3390/ijms21041342

Arruda, M. P., Lipka, A. E., Brown, P. J., Krill, A. M., Thurber, C., Brown-Guedira, G., et al. (2016). Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum* L.). *Mol Breed.* 36:84. doi: 10.1007/s11032-016-0508-5

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

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

Beyene, Y., Gowda, M., Olsen, M., Robbins, K. R., Pérez-Rodríguez, P., Alvarado, G., et al. (2019). Empirical comparison of tropical maize hybrids selected through genomic and phenotypic selections. *Front. Plant Sci.*10:1502. doi: 10.3389/fpls.2019.01502

Bian, Y., and Holland, J. B. (2017). Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. *Heredity* 118, 585–593. doi: 10.1038/hdy.2017.4

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, 2633–2635. doi: 10.1093/bioinformatics/btm308

Burgueño, J., Campos, G. D. L., 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

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

Cerrudo, D., Cao, S., Yuan, Y., Martinez, C., Suarez, E. A., Babu, R., et al. (2018). Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Front. Plant Sci.* 9:366. doi: 10.3389/fpls.2018.00366

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

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

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

Gianola, D., and van Kaam, J. B. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303. doi: 10.1534/genetics.107.084285

González-Camacho, J. M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125, 759–771. doi: 10.1007/s00122-012-1868-9

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

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

Guo, Z., Tucker, D. M., Lu, J., Kishore, V., and Gay, G. (2021). Evaluation of genome-wide selection efficiency in maize nested association mapping populations. *Theor. Appl. Genet.* 124, 261–275. doi: 10.1007/s00122-011-1702-9

Hao, Y., Wang, H., Yang, X., Zhang, H., He, C., Li, D., et al. (2019). Genomic prediction using existing historical data contributing to selection in biparental populations: a study of kernel oil in maize. *Plant Genome* 12:180025. doi: 10.3835/plantgenome2018.05.0025

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

Herter, C. P., Ebmeyer, E., Kollers, S., Korzun, V., Würschum, T., and Miedaner, T. (2019). Accuracy of within- and among-family genomic prediction for Fusarium head blight and *Septoria tritici* blotch in winter wheat. *Theor. Appl. Genet.* 132, 1121–1135. doi: 10.1007/s00122-018-3264-6

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52, 146–160. doi: 10.2135/cropsci2011.06.0297

Islam, M. S., Fang, D. D., Jenkins, J. N., Guo, J., McCarty, J. C., and Jones, D. C. (2020). Evaluation of genomic selection methods for predicting fiber quality traits in Upland cotton. *Mol. Genet. Genomics* 295, 67–79. doi: 10.1007/s00438-019-01599-z

Jeong, S., Kim, J. Y., and Kim, N. (2020). GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. *Sci. Rep.* 10, 19653–19665. doi: 10.1038/s41598-020-76759-y

Li, G., Dong, Y., Zhao, Y., Tian, X., and Liu, W. (2020). Genome-wide prediction in a hybrid maize population adapted to Northwest China. *Crop J.* 8, 830–842. doi: 10.1016/j.cj.2020.04.006

Lipka, A. E., Tian, F., Wang, Q., Peifer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12:e1005767. doi: 10.1371/journal.pgen.1005767

Liu, X., Wang, H., Hu, X., Li, K., Liu, Z., Wu, Y., et al. (2019). Improving genomic selection with quantitative trait loci and nonadditive effects revealed by empirical evidence in maize. *Front. Plant Sci.* 10:1129. doi: 10.3389/fpls.2019.01129

Liu, X., Wang, H., Wang, H., Guo, Z., Xu, X., Liu, J., et al. (2018). Factors affecting genomic selection revealed by empirical evidence in maize. *Crop J.* 6, 341–352. doi: 10.1016/j.cj.2018.03.005

Liu, Y., Hu,. G., Zhang, A., Loladze, A., Hu, Y., Wang, H., et al. (2020). Genome-wide association study and genomic prediction of *Fusarium* ear rot resistance in tropical maize germplasm. *Crop J.* 9, 325–341 doi: 10.1016/j.cj.2020.08.008

Lozada, D. N., Mason, R. E., Sarinelli, J. M., and Brown-Guedira, G. (2019). Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. *BMC Genetics* 20, 82–93. doi: 10.1186/s12863-019-0785-1

Ma, J., Wang, L., Cao, Y., Wang, H., and Li,. H. (2021). Association mapping and transcriptome analysis reveal the genetic architecture of maize kernel size. *Front. Plant Sci.* 12:632788. doi: 10.3389/fpls.2021.632788

Massman, J. M., Jung, H. J. G., and Bernardo, R. (2013). Genomewide selection verses marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Sci.* 53, 58–66. doi: 10.2135/cropsci2012.02.0112

Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

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

Millet, E. J., Welcker, C., Kruijer, W., Negro, S., Coupel-Ledru, A., Nicolas, S. D., et al. (2016). Genome-wide analysis of yield in Europe: Allelic effects vary with drought and heat scenarios. *Plant Physiol.* 172, 749–764. doi: 10.1104/pp.16.00621

Montesinos-López, A., Montesinos-López, O. A., Crossa, J., Burgueño, J., Eskridge, K. M., Falconi-Castillo, E., et al. (2016). Genomic Bayesian prediction model for count data with genotype × environment interaction. *G3-Genes Genom. Genet.* 6, 1165–1177. doi: 10.1534/g3.116.028118

Morota, G., and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Front. Genet.* 5, 363. doi: 10.3389/fgene.2014.00363

Odilbekov, F., Armonien,é, R., Koc, A., Svensson, J., and Chawade, A. (2019). GWAS-assisted genomic prediction to predict resistance to Septoria Tritici Blotch in Nordic winter wheat at seedling stage. *Front. Genet.* 10:1224. doi: 10.3389/fgene.2019.01224

Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., and Singh, A. K. (2019). Machine learning approach for prescriptive plant breeding. *Sci. Rep.* 9:17132. doi: 10.1038/s41598-019-53451-4

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-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., and Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3-Genes Genom. Genet.* 2, 1595–1605. doi: 10.1534/g3.112.003665

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Rice, B., and Lipka, A. E. (2019). Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. *Plant Genome*, 12:180052. doi: 10.3835/plantgenome2018.07.0052

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.* 4, 217–220. doi: 10.1038/ng.1033

Rio, S., Mary-Huard, T., Moreau, L., and Charcosset, A. (2019). Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. *Theor. Appl. Genet.* 132, 81–96. doi: 10.1007/s00122-018-3196-1

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

Schulthess, A. W., Zhao, Y., Longin, C. F. H., and Reif, J. C. (2018). Advantages and limitations of multiple-trait genomic prediction for *Fusarium* head blight severity in hybrid wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 131, 685–701. doi: 10.1007/s00122-017-3029-7

Sehgal, D., Rosyara, U., Mondal, S., Singh, R., Poland, J., and Dreisigacker, S. (2020). Incorporating genome-wide association mapping results into genomic prediction models for grain yield and yield stability in CIMMYT spring bread wheat. *Front. Plant Sci.* 11:197. doi: 10.3389/fpls.2020.00197

Shi, Z., Song, W., Xing, J., Duan, M., Wang, F., Tian, H., et al. (2017). Molecular mapping of quantitative trait loci for three kernel-related traits in maize using a double haploid population. *Mol. Breed.* 37:108. doi: 10.1007/s11032-017-0706-9

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:e1004982. doi: 10.1371/journal.pgen.1004982

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J. L., et al. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116, 395–408. doi: 10.1038/hdy.2015.113

Sun, S., Wang, C., Ding, H., and Zou, Q. (2020). Machine learning and its applications in plant molecular studies. *Brief Funct. Genomics* 19, 40–48. doi: 10.1093/bfgp/elz036

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Yuan, Y., Cairns, J. E., Babu, R., Gowda, M., Makumbi, D., Magorokosho, C., et al. (2019). Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in maize. *Front. Plant Sci.* 9:1919. doi: 10.3389/fpls.2018.01919

Zhang, C., Zhou, Z., Yong, H., Zhang, X., Hao, Z., Zhang, F., et al. (2017). Analysis of the genetic architecture of maize ear and grain morphological traits by combined linkage and association mapping. *Theor. Appl. Genet.* 130, 1011–1029. doi: 10.1007/s00122-017-2867-7

Zhang, H., Yin, L., Wang, M., Yuan, X., and Liu, X. (2019a). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Front. Genet.* 10:189. doi: 10.3389/fgene.2019.00189

Zhang, X., Guan, Z., Li, Z., Liu, P., Ma, L., Zhang, Y., et al. (2020). A combination of linkage mapping and GWAS brings new elements on the genetic basis of yield-related traits in maize across multiple environments. *Theor. Appl. Genet.* 133, 2881–2895. doi: 10.1007/s00122-020-03639-4

Zhang, Y. M., Jia, Z., and Dunwell, J. M. (2019b). The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* 10:100. doi: 10.3389/fpls.2019.00100

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H. P., Longin, F. H., et al. (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theor. Appl. Genet.* 124, 769–776. doi: 10.1007/s00122-011-1745-y

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 Plant Science

ORIGINAL RESEARCH
published: 16 July 2021
doi: 10.3389/fpls.2021.672525

Genomic Prediction of Resistance to Tar Spot Complex of Maize in Multiple Populations Using Genotyping-by-Sequencing SNPs

Shiliang Cao[1,2†], *Junqiao Song*[2,3,4†], *Yibing Yuan*[2,5], *Ao Zhang*[2,6], *Jiaojiao Ren*[2,7], *Yubo Liu*[2,6], *Jingtao Qu*[2,5], *Guanghui Hu*[1,2], *Jianguo Zhang*[1], *Chunping Wang*[4], *Jingsheng Cao*[1], *Michael Olsen*[8], *Boddupalli M. Prasanna*[8], *Felix San Vicente*[2*] and *Xuecai Zhang*[2*]

[1] Maize Research Institute, Heilongjiang Academy of Agricultural Sciences, Harbin, China, [2] International Maize and Wheat Improvement Center (CIMMYT), El Batan, Mexico, [3] College of Agronomy, Henan University of Science and Technology, Luoyang, China, [4] Maize Research Institute, Anyang Academy of Agricultural Sciences, Anyang, China, [5] Maize Research Institute, Sichuan Agricultural University, Chengdu, China, [6] College of Biological Science and Technology, Shenyang Agricultural University, Shenyang, China, [7] College of Agronomy, Xinjiang Agricultural University, Urumqi, China, [8] International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya

Edited by:
Brian Gardunia,
Bayer Crop Science, United States

Reviewed by:
Francis Chuks Ogbonnaya,
Grains Research and Development Corporation, Australia
Claudia Teixeira Guimaraes,
Brazilian Agricultural Research Corporation (EMBRAPA), Brazil
Marcio Resende,
University of Florida, United States

***Correspondence:**
Xuecai Zhang
XC.Zhang@cgiar.org
Felix San Vicente
F.SanVicente@cgiar.org

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *26 February 2021*
Accepted: *21 June 2021*
Published: *16 July 2021*

Citation:
Cao S, Song J, Yuan Y, Zhang A, Ren J, Liu Y, Qu J, Hu G, Zhang J, Wang C, Cao J, Olsen M, Prasanna BM, San Vicente F and Zhang X (2021) Genomic Prediction of Resistance to Tar Spot Complex of Maize in Multiple Populations Using Genotyping-by-Sequencing SNPs. Front. Plant Sci. 12:672525. doi: 10.3389/fpls.2021.672525

Tar spot complex (TSC) is one of the most important foliar diseases in tropical maize. TSC resistance could be furtherly improved by implementing marker-assisted selection (MAS) and genomic selection (GS) individually, or by implementing them stepwise. Implementation of GS requires a profound understanding of factors affecting genomic prediction accuracy. In the present study, an association-mapping panel and three doubled haploid populations, genotyped with genotyping-by-sequencing, were used to estimate the effectiveness of GS for improving TSC resistance. When the training and prediction sets were independent, moderate-to-high prediction accuracies were achieved across populations by using the training sets with broader genetic diversity, or in pairwise populations having closer genetic relationships. A collection of inbred lines with broader genetic diversity could be used as a permanent training set for TSC improvement, which can be updated by adding more phenotyped lines having closer genetic relationships with the prediction set. The prediction accuracies estimated with a few significantly associated SNPs were moderate-to-high, and continuously increased as more significantly associated SNPs were included. It confirmed that TSC resistance could be furtherly improved by implementing GS for selecting multiple stable genomic regions simultaneously, or by implementing MAS and GS stepwise. The factors of marker density, marker quality, and heterozygosity rate of samples had minor effects on the estimation of the genomic prediction accuracy. The training set size, the genetic relationship between training and prediction sets, phenotypic and genotypic diversity of the training sets, and incorporating known trait-marker associations played more important roles in improving prediction accuracy. The result of the present study provides insight into less complex trait improvement via GS in maize.

Keywords: maize, tar spot complex, genomic prediction, genomic selection, prediction accuracy, genotyping-by sequencing

INTRODUCTION

Tar spot complex (TSC), caused by an interaction of at least three fungal species: *Phyllachora maydis*; *Monographella maydis*; and *Coniothyrium phyllachorae*, is one of the most important foliar diseases of maize (*Zea mays* L. subsp. *mays*) in many Central and South American tropical and subtropical areas (Hock et al., 1992; Pereyda-Hernández et al., 2009). TSC can result in up to 75% grain yield loss, due to reduced ear weight, low kernel filling, and loose kernels. Development and deployment of maize varieties with genetic resistance is the most economical and effective strategy for controlling TSC (Ceballos and Deutsch, 1992).

Understanding the genetic architecture of TSC resistance will allow breeders to improve their breeding efficiency by the implementation of marker-assisted selection (MAS) or genomic selection (GS) to introgress the resistance genes into susceptible germplasm. A few studies have been conducted to dissect the genetic architecture of TSC resistance in maize (Mahuku et al., 2016; Cao et al., 2017). In a collection of 890 inbred lines genotyped with 56 K SNPs, three TSC resistance loci on chromosomes 2, 7, and 8 were identified through association mapping (AM) analysis. The major quantitative resistance locus (QTL) detected on maize chromosome bin 8.03, was furtherly validated in three bi-parental populations through linkage mapping analysis. Identification of the major QTL on bin 8.03 provides the foundation for fine mapping this major QTL and developing functional markers for implementing MAS (Mahuku et al., 2016). The genetic architecture of TSC resistance in maize was confirmed by combined AM and linkage mapping using higher marker density, the major QTL on bin 8.03 was narrowed down to a 33.6 million base pair region, and the results showed that TSC resistance in maize is controlled by a major QTL on bin 8.03, coupled with several minor QTL with smaller effects on other chromosomes (Cao et al., 2017).

Genomic selection is an extension of MAS that uses genome-wide markers to predict the genomic estimated breeding values (GEBVs) of the un-tested lines for selection, where the genome-wide markers are used for selection without detection QTL (Meuwissen et al., 2001; Edriss et al., 2017). In maize, GS has been investigated to improve several major diseases, e.g., maize lethal necrosis resistance (Gowda et al., 2015; Sitonik et al., 2019), northern corn leaf blight resistance (Technow et al., 2013), ear rot resistance (Han et al., 2018; Liu et al., 2020). These studies showed that GS is a promising approach to improve the major diseases, which are under polygenic control. Medium-to-high prediction accuracies were achieved in these studies, and the factors affecting prediction accuracy were assessed over a wide range of target traits. Key factors affecting prediction accuracy include the heritability of the predicted trait (Combs and Bernardo, 2013; Zhang et al., 2015), size of the training set (Zhang et al., 2017), marker density (Spindel et al., 2015), marker quality (Guo et al., 2020), phenotypic, and genotypic variations of the target trait (Gowda et al., 2015), the genetic relationship between training and prediction sets (Isidro et al., 2015; Santantonio et al., 2020; Atanda et al., 2021), and incorporating known trait-marker associations (Bernardo, 2014; Wang et al., 2019), etc. A preliminary genomic prediction analysis has been conducted to investigate the effectiveness of implementing GS for improving TSC resistance in maize, results showed that moderate-to-high prediction accuracies were achieved within different populations using various population sizes and marker densities (Cao et al., 2017). The accuracy of predicting TSC resistance across populations is still unknown under the different factors affecting prediction accuracy.

In the present study, an association-mapping panel and three doubled haploid (DH) populations, genotyped with genotyping-by-sequencing (GBS), were used to estimate the genomic prediction accuracy of TSC resistance in maize. The main objectives of the present study are to: (1) estimate the genomic prediction accuracy of TSC resistance across populations, where the training and prediction sets are different; (2) assess the effect of marker density, marker quality, heterozygosity rate (HT) of samples, the genetic relationship between training and prediction sets, incorporating known trait-marker associations on estimation the genomic prediction accuracy of TSC resistance; (3) explore training population development base on the phenotypic variation of TSC resistance.

MATERIALS AND METHODS

Plant Materials, Phenotyping, and Phenotypic Data Analysis

In the present study, an AM panel and three bi-parental DH populations were used. The AM panel, designated Drought Tolerant Maize for Africa (DTMA) AM panel, consists of 282 tropical and subtropical inbred lines developed by the Global Maize Program of International Maize and Wheat Improvement Center (CIMMYT).

The three DH populations, namely Pop1, Pop2, and Pop3, consists of 174, 100, and 111 lines, respectively. Each of the DH populations was derived from an F_1 cross formed between a TSC resistant line and a TSC susceptible line, the protocol of generating DH lines was described by Prasanna et al. (2012). The resistant parental lines are widely used CIMMYT maize lines showing good resistance to TSC, and the susceptible parental lines are drought or drought and heat stress-tolerant lines (Yuan et al., 2019) showing severe susceptibility to TSC. The Pop2 and Pop3 shared a common donor line, and the susceptible parental lines of these two populations were derived from the same genetic pool through population improvement. The detailed information of the parental lines was described by Cao et al. (2017).

The DTMA AM panel was evaluated for TSC response in Mexico at five environments, i.e., in Puebla (Latitude: 20°28′; Longitude: −97°38′; Mega environment: lowland tropical) in 2009, 2011 and 2012; in Guerrero (Latitude: 17°02′; Longitude: −99°38′; Mega environment: lowland tropical) in 2012; and in Veracruz (Latitude: 19°15′; Longitude: −96°12′; Mega environment: lowland tropical) in 2012. Pop1 was evaluated for TSC response at three environments, i.e., in Puebla in 2011 and 2014; and in Guerrero in 2013. Pop2 was evaluated for TSC response at four environments, i.e., in Puebla in 2012 and 2014, each year had two planting dates. Pop3 was evaluated for TSC response at three environments, i.e., in Puebla in 2012 with two planting dates; and in Puebla in 2014 (Cao et al., 2017). All the locations used for disease screening had high and consistent natural pressure of TSC. A randomised complete block design was used for all experiments with three replications per location. Each plot consisted of a single 2-m row with 10 plants per row. The TSC score evaluation was performed according to the methods described by Mahuku et al. (2016). The disease severity was recorded using a scale of 1–5 with a 0.5 increment, where 1 = highly resistant (HR), no visible disease symptoms or lesions identifiable on any of the leaves; 5 = highly susceptible (HS), all leaves are dead, no green leaf tissue remaining or disease symptoms on more than 80% of the leaf surface.

MEATA-R software[1] (Alvarado et al., 2020) was used to analyze multi-location trials using a mixed linear model to estimate the best linear unbiased prediction (BLUP) value of genotypes and the broad-send heritability of the target trait in each population based on the entry mean within trials. The mixed linear model was applied as follows:

$$Y_{ijk} = \mu + g_i + e_j + ge_{ij} + r_k e_j + \varepsilon_{ijk}$$

where Y_{ijk} is the target trait, μ is the overall mean, g_i, e_j, and ge_{ij} are the effects of the i-th genotype, j-th environment, and i-th genotype by j-th environment interaction, respectively. $r_k e_j$ is the effect of the k-th replication within the j-th environment. ε_{ijk} is the residual effect of the i-th genotype, j-th environment, and k-th replication. Genotype is considered as the fixed effect, whereas all other terms are declared as the random effects.

Broad-sense heritability (H^2) based on the entry means within trials was estimated as follows:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{ge}^2}{ne} + \frac{\sigma_e^2}{ne\ \ nr}}$$

where σ_g^2, σ_e^2, and σ_{ge}^2 are the genotypic variance, error variance, and genotype-by-environment interaction variance, respectively, and nr and ne are the numbers of replications and environments, respectively.

Genotyping and SNP Calling

A commonly used GBS protocol was applied in the present study, which was described in the previous studies (Elshire et al., 2011; Wu et al., 2016; Wang et al., 2020). The SNP calling and imputation was performed according to the methods previously described (Glaubitz et al., 2014; Swarts et al., 2014). Both the un-imputed and the imputed datasets were generated for all four populations of the present study. The un-imputed datasets were only used in the three DH populations to build the block maps to perform linkage mapping analyses. The rest of the analyses were performed with the imputed datasets. Initially, 955,690 SNPs, evenly distributed on the 10 maize chromosomes, were called for each of the genotyped samples.

Population Structure Analysis

The population structure analysis was performed with the principal components analysis (PCA) in all the four populations, where 232,538 SNPs, filtered with minor allele frequency (MAF) greater than 0.05 and missing rate (MR) less than 20%, were utilised. In the DTMA AM panel, the population structure analysis was applied in software Structure V2.3.3 using an admixture model-based clustering method (Hubisz et al., 2009), where a sub-set of 10,000 SNPs with no missing values were randomly selected to perform this analysis. The heat map of the number of SNPs within 1 Mb physical position was shown in **Supplementary Figure 1**, which indicates that the 10,000 SNPs almost evenly distribute in 10 maize chromosomes. The average linkage disequilibrium decay distance reported in the previous study was 3.5 Kb at $r^2 = 0.1$ (Cao et al., 2017). In the DTMA AM panel, evenly distributed SNPs and rapid linkage disequilibrium decay are able to avoid the introduction of bias of oversampling SNPs in the linkage disequilibrium regions in the population structure analysis. Hypotheses were tested for sub-population number K ranging from 1 to 10, and each K was run seven times with burn-in time and replications both to 100,000.

Genomic Prediction Analysis

The genomic prediction was implemented in the *rrBLUP* package (Endelman, 2011). In each population, a five-fold cross-validation scheme with 100 replications was used to estimate the prediction accuracy of r_{MG}. The 80% of the lines in each population were randomly assigned as a training set to estimate the effect of the molecular markers and train the prediction model, the rest of the 20% lines were assigned as a validation set in each replication to get the GEBV of each line in the validation set. The average correlation coefficient between the GEBVs and the observed breeding values of the lines in the validation set was defined as the prediction accuracy r_{MG}. Within each of the four populations, SNPs filtered with MAF greater than 0.05 and MR less than 20%, were used for the genomic prediction analyses with a five-fold cross-validation scheme.

Effect of the Genetic Relationship on the Estimation of the Prediction Accuracy

According to the changes of *ad hoc* statistic delta K (ΔK) value, the DTMA AM panel was divided into several subgroups. Within each subgroup, a five-fold cross-validation scheme was used to estimate the prediction accuracy of r_{MG}. Besides, the predictions were also conducted between pairwise subgroups, when one subgroup was used as a training set to predict the other subgroup.

Across all the four populations, the predictions were also conducted between pairwise populations, where SNPs filtered with MAF greater than 0.05 and MR less than 20% across all the four populations, were used for the genomic prediction analyses. When the predictions were made across subgroups or populations, the training and validation sets were independent,

[1] http://hdl.handle.net/11529/10201

and the prediction accuracy of r_{MG} estimated in the validation set was calculated from only a one-time analysis.

Effect of Marker Quantity and Quality on the Estimation of the Prediction Accuracy

To assess the effect of marker quantity and quality on the estimation of the prediction accuracy, different parameters were applied to filter the SNP dataset within each population to perform the genomic prediction analyses. In each population, a five-fold cross-validation scheme with 100 replications was used to estimate the prediction accuracies. The prediction accuracies were compared, when the SNP datasets filtered with different parameters, were used for genomic prediction analyses.

Different levels of MAF, MR, and HT of the SNPs were used to control the marker quantity and quality. Nine combinations between three MAF levels and three MR levels were used to filter the SNP dataset within each population, MAF setting at 0.05, 0.20, and 0.40; MR setting at 0.00, 10, and 20%. The HT of the SNPs in the DTMA AM panel was set at 1, 3, 5, and 10% after the SNPs were filtered with MAF of 0.05 and MR of 0%, and the prediction accuracies were estimated with a five-fold cross-validation scheme. The effect of the HT of the samples on the estimation of the prediction accuracy was evaluated by setting the HT of the samples at 1, 3, 5, and 10% in the populations of DTMA and Pop1, where the SNPs filtered with MAF greater than 0.05 and MR of 0% were used for prediction analyses. The software of TASSEL V5.0 (Bradbury et al., 2007) was used to filter the imputed dataset with MAF and MR. The customised R scripts were used to filter the HT of the SNPs and samples.

Genomic Prediction Analyses With the Significantly Associated Markers Detected From the Genetic Mapping

Genomic prediction analyses with significantly associated markers were performed to simulate MAS. In the previous study, 261,948 filtered SNPs were used to perform AM analysis in the DTMA AM panel. In total, 155 SNPs were identified that were significantly associated with TSC resistance in maize at the threshold of $-\log10$ $(P) > 4.53$ (Cao et al., 2017). A five-fold cross-validation scheme was used to assess the accuracies of genomic predictions conducted with the significantly associated markers and the same number of random-selected markers, the number of markers was set as 1, 2, 3, 4, 5, 10, 20, 155, 500, 1000, 3000, 5000, 10,000, 30,000, 50,000, 100,000, and 200,000. The significant markers were selected based on their $-\log10$ (P) value, and their chromosome positions. The most significantly associated SNPs were selected on all chromosomes firstly, and then the second significant-associated SNPs were selected.

A block map was constructed in each of the three DH populations to perform linkage mapping in a previous study (Cao et al., 2017), where the blocks were treated as genetic markers to construct the genetic map. In total, 437 blocks in Pop1, 494 blocks in Pop2, and 493 blocks in Pop3 were built with 20,473, 27,818 and 326,07 SNPs, respectively. In the software of QTL IciMapping Version 4.1 (Meng et al., 2015), the single-marker analysis method was used to perform the linkage mapping analyses and rank the scores of the log of the odds of all the blocks, the scores of the log of the odds representing the significant levels of the association between the block and the TSC resistance. A five-fold cross-validation scheme was used to assess the accuracies of genomic predictions conducted with the significantly associated markers and the same number of random-selected markers, the number of markers was set as 5, 10, 15, 20, 30, 50, 100, 200, 300, 400, and all the blocks in each population.

In the above analyses, the prediction accuracy could be overestimated, because the same population was used to identify the significantly associated markers firstly, and then it was used to calculate the prediction accuracy estimated with the significantly associated markers. To avoid the overestimated prediction accuracy, the 150 significantly associated markers detected from the DTMA AM panel were used for estimating the prediction accuracy in each of the three DH populations, when the DTMA AM panel was used as the training set, and the DH population was used as the validation set. For comparison, 150 randomly selected markers were also used to estimate the prediction accuracy in each of the three DH populations.

Training Set Development Based on the Phenotypic Variation of TSC Resistance

According to the phenotypic variation information of the TSC resistance in each population, training sets were formed. Four scenarios were simulated and compared within each of the four populations, where the training set was formed by sampling the same percentage of materials with a selection from both resistant and susceptible tails (R + S), with random selection (RD), with a selection from the resistant tail (R), with a selection from the susceptible tails (S), respectively. In each scenario, the validation set was the whole population, and the training set ranged from 20 to 60%, with an interval of 20%. In each of the four populations, a total of 12 combinations and comparisons were conducted between the four scenarios and the three percentage levels of the training set.

RESULTS

Phenotypic Variation, Heritability, and Phenotypic Correlation Between Locations

The BLUP value of TSC resistance of all the genotypes across the four populations ranged from 1.31 to 4.39. The Pop3 had the widest range of variation among the four populations. The minimum BLUP value was 1.31,1.81, 1.18, and 1.37 in the DTMA AM panel, Pop1, Pop2, and Pop3, respectively. The maximum BLUP value was 3.23, 3.00, 3.95, and 4.39 in the DTMA AM panel, Pop1, Pop2, and Pop3, respectively. The heritability of TSC resistance across locations was 0.80, 0.54, 0.88, and 0.93 in the DTMA AM panel, Pop1, Pop2, and Pop3, respectively. The average phenotypic correlation coefficient of TSC resistance between locations was 0.47, 0.37, 0.68, and 0.84 in the DTMA AM panel, Pop1, Pop2, and Pop3, respectively.

Population Structure Analysis Within and Among Populations

According to the *ad hoc* statistic ΔK value changes, the DTMA AM panel was divided into three subgroups, the number of lines was 40, 111, and 131 in Subgroups 1, 2, and 3, respectively (Cao et al., 2017). Most of the lines in Subgroup 1 were from the Mexico physiology research group, lines in Subgroup 2 were mainly from the subtropical breeding program, and lines in Subgroup 3 were mainly from the lowland tropical breeding program. The result of the structure split for all the Ks (1–10) was provided in **Supplementary File 1**. The population structure within the DTMA AM panel was illustrated with the first two principal components in **Figure 1A**, where the results showed the first and second principal components explained 4.48 and 3.66% of the total SNP variation, respectively. Some lines from each subgroup centrally clustered with each other, indicating the moderate level of genetic relatedness among the subgroups. The inbred lines in the Subgroup 3 were most widely scattered, implying the broadest genetic diversity presented in Subgroup 3 among all the three subgroups. These observations are consistent with the current germplasm exchange patterns where there is a constant flow of germplasm among the subgroups.

The genetic relationship among the four populations was illustrated with the first two principal components in **Figure 1B**, where the results showed the first and second principal components explained 14.11 and 7.57% of the total SNP variation, respectively. The inbred lines in the DTMA AM panel were most widely scattered, implying the broadest genetic diversity presented in the DTMA AM panel among all four populations. The DTMA AM panel was not overlapped with any of the three bi-parental populations, the Pop1 was not overlapped with either the Pop2 or the Pop3. The Pop2 was overlapped with the Pop3, due to the common parent shared by these two populations, it indicated the closest relationships between these two populations.

Genomic Prediction Accuracies Obtained Within and Across Populations and Subgroups

Genomic prediction accuracies obtained from five-fold cross-validations and 100 replications were high in all four populations, when the SNP datasets, filtered with MAF greater than 0.05 and MR less than 20%, were used to perform prediction within each population. The number of SNPs after filter in the DTMA AM panel, Pop1, Pop2, and Pop3 were 261,948, 98,018, 102,204, and 104,046, respectively. The r_{MG} values observed in the DTMA AM panel, Pop1, Pop2, and Pop3 were 0.56, 0.60, 0.75, and 0.69. The r_{MG} value observed in the DTMA AM panel was lower than those observed in the DH populations.

Genomic prediction accuracies obtained from five-fold cross-validations and 100 replications were low to moderate within the three subgroups of the DTMA AM panel (**Table 1**). The r_{MG} values observed in the Subgroup 1, Subgroup 2, and Subgroup 3 were 0.27, 0.55, 0.35, respectively. The r_{MG} values observed in the subgroups of the DTMA AM panel were lower than those observed in the DTMA AM panel.

Genomic prediction accuracies obtained across subgroups were relatively low when the predictions were performed between pairwise subgroups (**Table 1**). The r_{MG} values observed between pairwise subgroups ranged from −0.30 to 0.33, the relative high prediction accuracies were observed, when Subgroup 3 was used as a training set to predict the other two subgroups, because of the bigger population size and broadest genetic diversity presented in Subgroup 3 contributing to the improvement of prediction accuracy. The r_{MG} values observed between pairwise subgroups were lower than those observed within the subgroups.

Genomic prediction accuracies obtained across populations varied in different scenarios and ranged from 0.20 to 0.64 (**Table 2**). The plots of the correlation between the predicted and the observed BLUP values for these predictions were shown in **Supplementary Figure 2**. When the DTMA AM panel was used as the training set, the r_{MG} values observed in the Pop1, Pop2, and Pop3 were 0.45, 0.61, and 0.55, respectively. When the DH populations were used as the training set to predict the DTMA AM panel, the r_{MG} values observed in the DTMA AM panel were relatively low and ranged from 0.20 to 026. The r_{MG} values observed between the pairwise DH populations were moderate to high and ranged from 0.36 to 0.64. The highest r_{MG} values were observed in pairwise populations of Pop2 and Pop3, i.e., 0.64 and 0.60. The lowest r_{MG} values were observed in pairwise populations of Pop1 and Pop3, i.e., 0.36 and 0.40.

Genomic Prediction Accuracies Obtained From Different Levels of Marker Density, Marker Quality, and Heterozygosity Rate of Samples

Across all the populations, the number of markers decreased as the MAF increased and the MR decreased, the marker quality improved as the number of markers decreased. The maximum number of markers and the highest MD were observed by filtered the SNPs with the combination of MAF of 0.05 and MR of 20%, the minimum number of markers and the lowest MD were observed by filtered the SNPs with the combination of MAF of 0.40 and MR of 0%. The number of SNPs filtered with the combination of MAF of 0.05 and MR of 20% in the DTMA AM panel, Pop1, Pop2, and Pop3 was 261,948, 98,018, 102,204, and 104,046, respectively. The number of SNPs filtered with the combination of MAF of 0.40 and MR of 0% in the DTMA AM panel, Pop1, Pop2, and Pop3 was 1144, 61,471, 65,923, and 61,525, respectively.

The prediction accuracy results estimated in all the four populations under the nine marker datasets filtered with the combinations of MAF and MR were shown in **Figure 2**. Within each population, the r_{MG} values estimated with the different marker datasets were slightly different. The r_{MG} values ranged from 0.54 to 0.58 in the DTMA AM panel, from 0.59 to 0.61 in the Pop1 population, from 0.75 to 0.78 in the Pop2 population, and from 0.65 to 0.71 in the Pop3 population. Across all the populations, relatively high and similar prediction accuracies were obtained across all levels of MAF and MR, indicating that

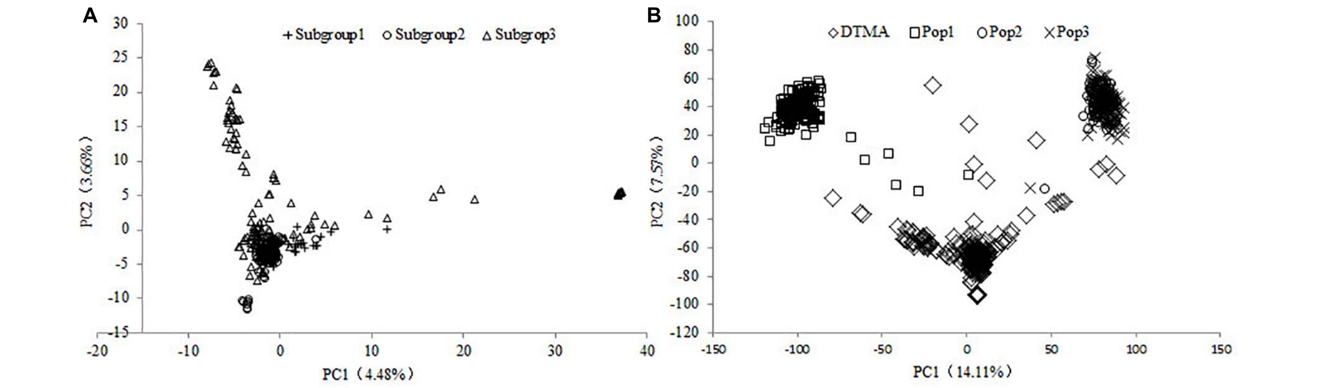


FIGURE 1 | Results of the principal components (PC) analysis in the **(A)** DTMA association mapping panel, and in **(B)** all the four populations of DTMA association mapping panel, Pop1, Pop2, and Pop3.

the levels of MAF and MR had minor effects on the estimation of the prediction accuracy.

The prediction accuracy results of all the four populations estimated at the four levels of HT of SNPs at 1, 3, 5, and 10% were shown in **Figure 3**. Under the combination of MAF of 0.05 and MR of 0%, the number of markers in the DTMA AM panel filtered with the HT of SNPs at 1, 3, 5, and 10% were 582, 4274, 7503, and 10,065, respectively. The r_{MG} values estimated from the number of SNPs of 582, 4274, 7503, and 10,065 were 0.45, 0.53, 0.53, and 0.54, respectively (**Figure 3A**). A significant increase of the r_{MG} value was observed in the DTMA AM panel, when the HT of SNPs changed from 1 to 3% and the number of SNPs increased from 582 to 4274. Under the combination of MAF of 0.05 and MR of 20%, the number of markers in all the DH populations were filtered with the HT of SNPs at 1, 3, 5, and 10%. The slight differences were observed on the r_{MG} values, as the HT of SNPs increased in all the DH populations (**Figures 3B–D**). These results indicated that the effect of HT of SNPs on the estimation of the prediction accuracy is mainly caused by the changes in the number of SNPs.

TABLE 1 | Genomic prediction accuracies for TSC resistance obtained between the three subgroups of the DTMA association mapping panel.

Training set (number of lines)	Validation set	Prediction accuracy
Subgroup 1 (40)	Subgroup 1	0.27
	Subgroup 2	−0.08
	Subgroup 3	−0.03
Subgroup 2 (111)	Subgroup 2	0.55
	Subgroup 1	−0.3
	Subgroup 3	0.07
Subgroup 3 (131)	Subgroup 3	0.35
	Subgroup 1	0.16
	Subgroup 2	0.33

TABLE 2 | Genomic prediction accuracies for TSC resistance obtained between all the four populations of DTMA association mapping panel, Pop1, Pop2, and Pop3.

Training set (number of lines)	Validation set	Prediction accuracy
DTMA (282)	Pop1	0.45
	Pop2	0.61
	Pop3	0.55
Pop1 (174)	DTMA	0.26
	Pop2	0.61
	Pop3	0.40
Pop2 (100)	DTMA	0.20
	Pop1	0.52
	Pop3	0.60
Pop3 (111)	DTMA	0.23
	Pop1	0.36
	Pop2	0.64

The prediction accuracy results of all the four populations estimated at the four levels of HT of samples at 1, 3, 5, and 10% were shown in **Figure 4**. Under the combination of MAF of 0.05 and MR of 0%, the number of samples in the DTMA AM panel filtered with the HT of the sample at 1, 3, 5, and 10% were 120, 184, 219, 250, respectively. The r_{MG} values estimated in the DTMA AM panel at the HT of samples of 1, 3, 5, and 10% were 0.53, 0.57, 0.56, and 0.56, respectively. In Pop1, the number of samples filtered with the HT of samples at 1, 3, 5, and 10% was 92, 165, 171, and 174, respectively. The r_{MG} values estimated in Pop1 at the HT of samples of 1, 3, 5, and 10% were 0.59, 0.59, 0.59, and 0.61, respectively. In Pop2, the number of samples filtered with the HT of samples at 1, 3, 5, and 10% was 46, 95, 100, and 100, respectively. The r_{MG} values estimated in the Pop2 at the HT of samples of 1, 3, 5, and 10% were 0.65, 0.76, 0.77, and 0.77, respectively. In the Pop3, the number of samples filtered with the HT of samples at 1, 3, 5, and 10% was 77, 111, 111, and 111, respectively. The r_{MG} values estimated in Pop3 at the HT of samples of 1, 3, 5, and 10% were 0.68, 0.69, 0.69, and 0.69, respectively. Similar trends were observed across all four populations, the slight increases were observed on the r_{MG} values, as the HT of samples increased. These results showed that the effect of HT of samples on the estimation of the prediction accuracy is mainly caused by the changes in the number of samples.

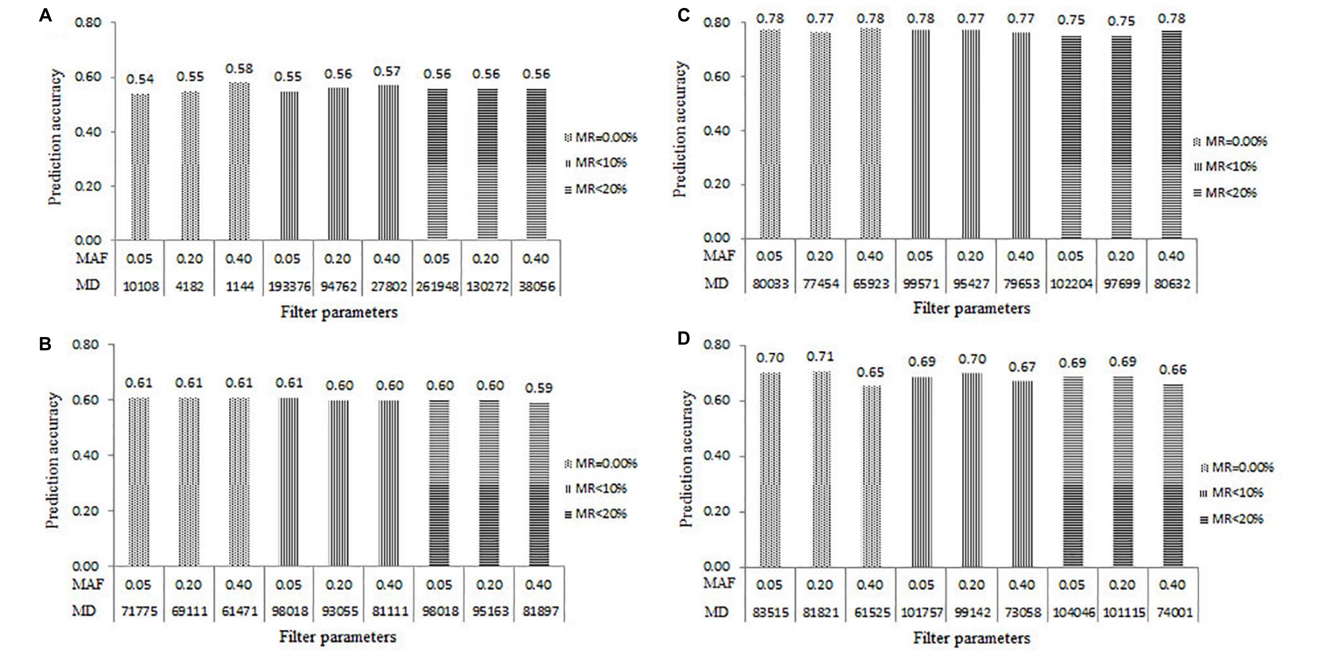


FIGURE 2 | Genomic prediction accuracies for TSC resistance estimated from the five-fold cross-validation scheme in all the four populations of **(A)** DTMA association mapping panel, **(B)** Pop1, **(C)** Pop2, and **(D)** Pop3, under the nine levels of marker density (MD) filtered with the combinations of three levels of minor allele frequency (MAF) and three levels of missing rate (MR).

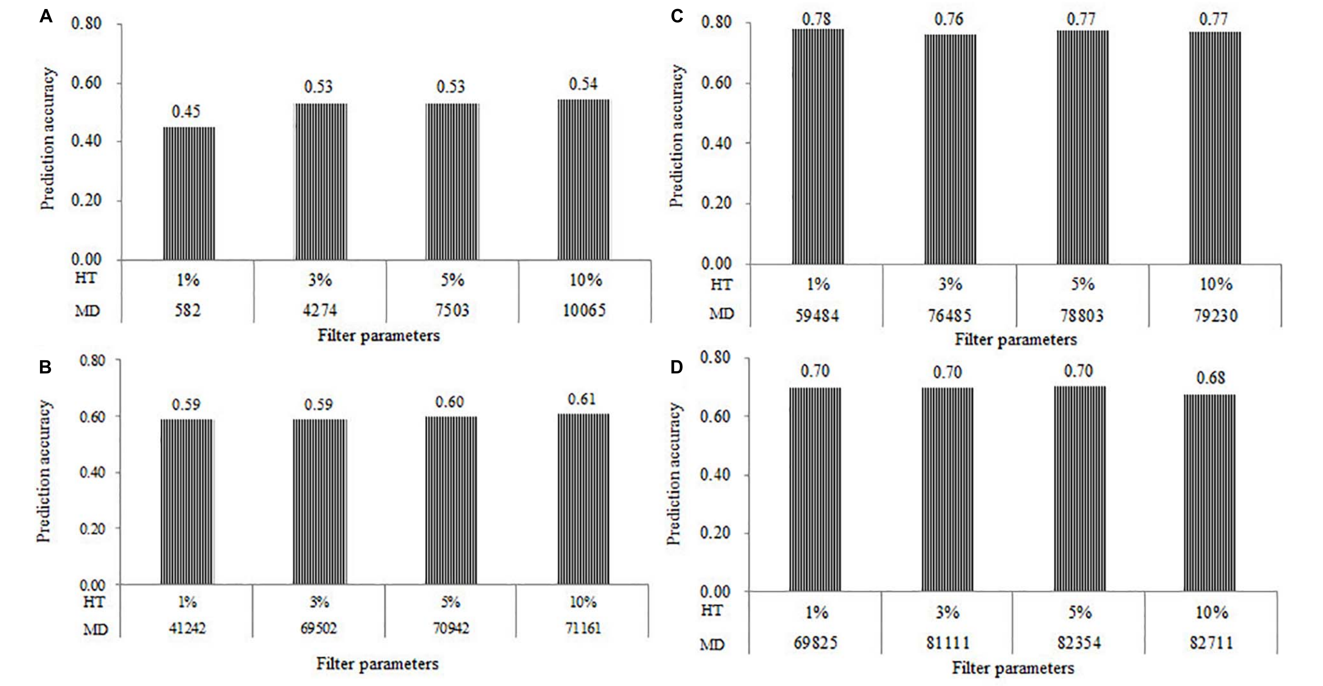


FIGURE 3 | Genomic prediction accuracies for TSC resistance obtained in the **(A)** DTMA association mapping panel, **(B)** Pop1; **(C)** Pop2; **(D)** Pop3, under the different levels of marker density (MD) at the four levels of heterozygosity rate (HT) of SNPs at 1, 3, 5, and 10%, and filtered with the combination of minor allele frequency (MAF) of 0.05 and missing rate (MR) of 0%.

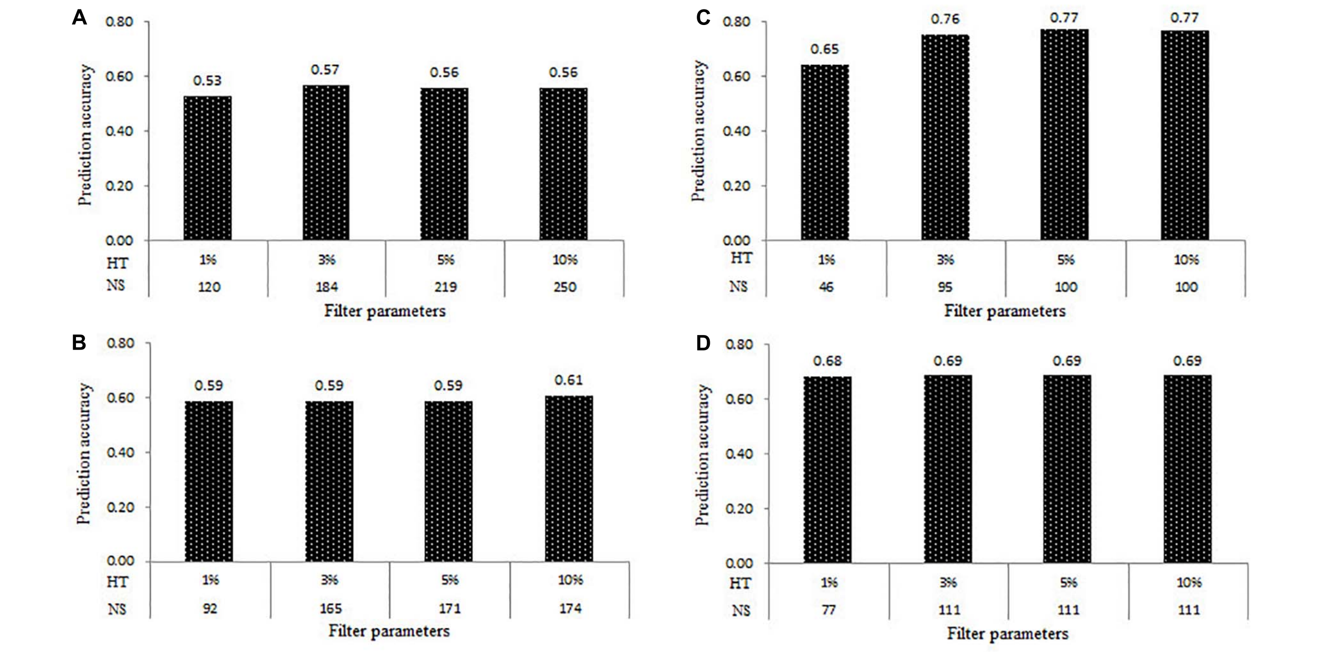


FIGURE 4 | Genomic prediction accuracies for TSC resistance obtained in the four populations of the **(A)** DTMA association mapping panel, **(B)** Pop1, **(C)** Pop2, and **(D)** Pop3, at the four levels of heterozygosity rate (HT) of samples of 1, 3, 5, and 10%, and the different number of samples (NS).

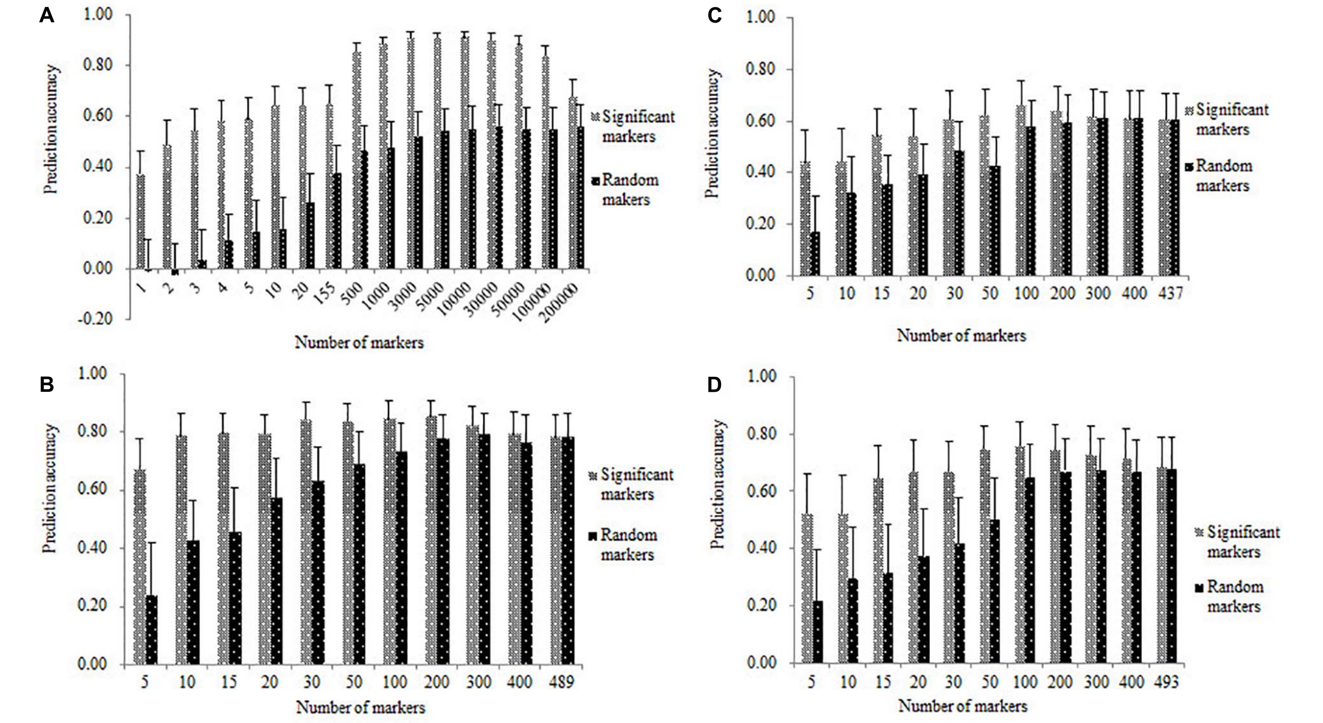


FIGURE 5 | Genomic prediction accuracies for TSC resistance estimated with the same number of significant and random markers in all the four populations of **(A)** DTMA association mapping panel, **(B)** Pop1, **(C)** Pop2, and **(D)** Pop3.

Genomic Prediction Accuracies Obtained From the Significantly Associated Marker

Genomic prediction accuracies in all the four populations estimated with the significantly associated SNPs were shown in **Figure 5**, where relatively high r_{MG} values were obtained with a few significantly associated markers in each of the four populations. The r_{MG} values obtained from the significantly associated SNPs were consistently higher than those obtained from the same number of randomly selected markers. In the DTMA AM panel, the number of significant-associated markers detected on the chromosomes of 2, 3, 7, and 8, were 1, 3, 1, and 150, respectively. The significantly associated SNPs used for prediction were ranked based on the information of their significant *p*-values and physical positions, and the top five significantly associated SNPs with the lowest *p*-values used for prediction were selected from the chromosomes of 8, 3, 2, 7, and 3, respectively. In the DTMA panel, the r_{MG} value obtained from the most significantly associated SNP on chromosome 8 was 0.37. The r_{MG} values obtained from the top two, top three, top four, and top five significantly associated SNPs were 0.49, 0.54, 0.58, and 0.59, respectively. The r_{MG} values consistently increased, as more significantly associated SNPs were used for prediction. The r_{MG} values reached the plateau, once the number of significantly associated SNPs used for prediction increased to more than 500. Similar trends were observed in the three DH populations, the r_{MG} values obtained from the significantly associated markers were consistently higher than those obtained from the same number of randomly selected markers, the r_{MG} values reached the plateaus in the DH populations, once the number of significantly associated markers used for prediction increased to more than 50. These results indicated that incorporating the significantly associated SNPs into GS has the potential for improving the prediction accuracy.

Genomic prediction accuracies in the DH populations of Pop1, Pop2, and Pop3 estimated with the 150 significantly associated SNPs were higher than those estimated with the same number of randomly selected SNPs (**Table 3**), when the DTMA AM panel was used as the training set to predict the DH population as the validation set. The genomic prediction accuracies estimated with the 150 significantly associated SNPs were 0.39, 0.49, and 0.43 in the Pop1, Pop2, and Pop3, respectively. The genomic prediction accuracies estimated with the 150 randomly selected SNPs were 0.09, 0.15, and 0.11 in the Pop1, Pop2, and Pop3, respectively.

TABLE 3 | Genomic prediction accuracies in the DH populations of Pop1, Pop2, and Pop3 estimated with the 150 significantly associated SNPs and the same number of randomly selected SNPs.

Training set	Validation set	Prediction accuracy estimated with the 150 significantly associated SNPs	Prediction accuracy estimated with the 150 randomly selected SNPs
DTMA	Pop1	0.39	0.09
DTMA	Pop2	0.49	0.15
DTMA	Pop3	0.43	0.11

Training Set Development Based on the Phenotypic Variation of the Target Trait

For all the four populations, the results of the prediction accuracies estimated in the 12 combinations between the four scenarios and the three percentage levels of the training set were presented in **Figure 6**. Across all four scenarios, the prediction accuracy increased in all the populations as the increase of percentage of the training set. For example, the prediction accuracies in the scenario of R+S were 0.72, 0.82, and 0.87, when the percentages of the training set in the DTMA panel were set as 20, 40, and 60%, respectively. Under the same percentage of the training set, the scenario of R+S outperformed the other three scenarios, and the scenario of RD outperformed the other two scenarios of R and S. For example, the prediction accuracy in the DTMA panel at the percentage of the training set at 60% were 0.87, 0.81, 0.59 and 0.71 for the scenario of R+S, RD, R, and S, respectively. Similar trends were also observed in the three DH populations. These results indicated that the training set development with broad phenotypic variation has the potential improving prediction accuracy.

DISCUSSION

In tropical and subtropical areas of Central and South America, TSC is one of the most destructive foliar diseases of maize, it may cause up to 75% grain yield loss. A few genetic studies have been conducted to dissect the genetic architecture of resistance to TSC of maize (Mahuku et al., 2016; Cao et al., 2017), where the heritabilities of TSC in different populations were medium-to-high, revealing that the phenotypic selection is effective for improving TSC resistance. However, improving TSC resistance through phenotypic selection is cost-intensive and time-consuming, because multiple location trials are required to improve TSC resistance through phenotypic selection.

Previously published studies revealed that TSC resistance in maize is controlled by a major QTL on bin 8.03, coupled with several minor QTL with smaller effects on other chromosomes. Fine mapping the major QTL on bin 8.03 and developing function markers associated with this major QTL will facilitate the implementation of MAS for improving breeding efficiency, and saving cost. In the present study, the effectiveness of MAS was simulated, when a few significantly associated SNPs were used for GS. In the DTMA panel, the prediction accuracy estimated with the most significantly associated SNPs on bin 8.03 was 0.37, and the prediction accuracy continuously increased as more significantly associated SNPs were used for GS. A similar trend was also observed in the three DH populations. These results implied that it is effective to improve the TSC resistance in maize by implementing MAS for introgression of the major QTL on bin 8.03 into susceptible germplasm. Moreover, TSC resistance in tropical maize could be furtherly improved by implementing GS for selecting multiple stable genomic regions simultaneously, or by implementing MAS and GS stepwise.

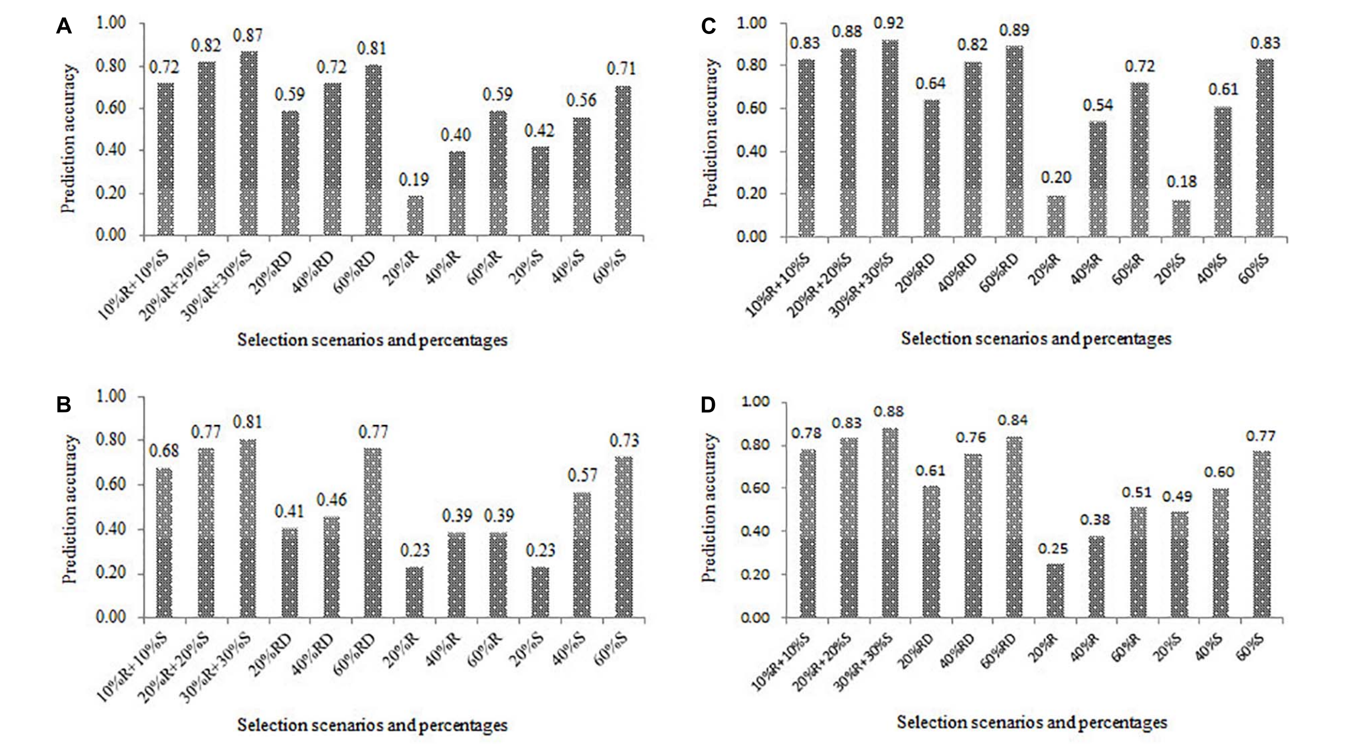


FIGURE 6 | Genomic prediction accuracies estimated in the 12 combinations between the four scenarios and the three percentage levels of the training set (20, 40, and 60%) in the four populations of **(A)** DTMA association mapping panel, **(B)** Pop1, **(C)** Pop2, and **(D)** Pop3. The scenario of R + S represents the selection from both resistant and susceptible tails, RD represents the random selection, R represents the selection from the resistant tail, S represents the selection from the susceptible tail.

In maize, GS has been shown as an effective genomic tool to improve breeding efficiency and accelerate genetic gain over a wide range of target traits with different levels of genetic complexity (Crossa et al., 2017). GS was implemented in various kinds of the population to estimate the genomic prediction accuracy of different target traits in several previous studies (Zhao et al., 2012; Vélez-Torres et al., 2018). In the previous study, moderate-to-high prediction accuracies of TSC resistance were achieved within each of the four populations (Cao et al., 2017). In the present study, moderate-to-high prediction accuracies were achieved across populations by using the training sets with broader genetic diversity, and in pairwise populations having closer genetic relationships. These results implied that a collection of inbred lines with broader genetic diversity could be phenotyped in multiple locations and used as a permanent training set, which will be employed to implement GP on the untested new populations. The training set could be updated by incorporating more new phenotyped lines, which have closer genetic relationships with the prediction set. Therefore, higher prediction accuracies can be achieved by strengthening the genetic relationship between the training and prediction sets and increasing the size of the training set (Riedelsheimer et al., 2013). This strategy will enhance breeding efficiency and save costs dramatically for improving TSC resistance in a breeding program. Moreover, a common training set also could be built for the implementation of GS on multiple traits improvement, especially for the less complex traits of foliar diseases or nutritional quality traits in maize, which can be predicted very well by using a collection of inbred lines with broad genetic diversity as the training set.

Implementation of GS requires a profound understanding of factors affecting genomic prediction accuracy (Zhang et al., 2017). In the previous study, the effects of training set size and marker density on the estimation of the genomic prediction accuracy of TSC resistance were investigated (Cao et al., 2017). In the present study, the effects of factors of marker density, marker quality, HT of samples, phenotypic diversity of the training set, incorporating known trait-marker associations on the estimation of the genomic prediction accuracy of TSC resistance were furtherly assessed. Results showed that the levels of MAF, MR, and HT of SNPs had minor effects on the estimation of the prediction accuracy. The effects of MAF, MR, and HT of SNPs on the estimation of the prediction accuracy are mainly caused by the changes in the number of SNPs. Once the number of SNPs is saturated on each chromosome, and at least one SNP per linkage disequilibrium block is selected for prediction, the prediction accuracy reaches a plateau (Lorenzana and Bernardo, 2009). There is a tradeoff between the number of markers and marker quality, marker quality becomes lower as the number of markers increases in a specific marker dataset. Appropriate levels of MAF, MR, and HT of SNPs should be considered and selected to improve the prediction accuracy and reduce the computational burden by balancing the number of markers and marker quality, this result is consistent with several previous studies (Guo et al., 2020). Within each of the four populations, slight increases in prediction accuracy were also observed, as the HT of samples increased and the training set size enlarged, indicating that training set size is an important factor improving prediction accuracy (Combs and Bernardo, 2013).

Selective genotyping is proposed to improve QTL mapping and save cost in bi-parental populations, where only the individuals from one or two tails with extreme phenotypic values are genotyped (Sun et al., 2010). In the present study, the R + S scenario built the training set by selecting the individuals from two tails with extreme phenotypic values, the R + S scenario had higher prediction accuracies than those in other scenarios. Taking the advantages of more accurate phenotyping and abundant phenotypic variation, the R + S scenario outperformed other scenarios. It implies that prediction accuracy can be improved by developing a training set with broad phenotypic variation, as well as broad genotypic diversity indicated in several previous studies (Gowda et al., 2015; Guo et al., 2020).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. These data can be found at the CIMMYT Research Data & Software Repository Network: https://hdl.handle.net/11529/10548579.

AUTHOR CONTRIBUTIONS

BP, MO, FS, and XZ conceived and designed the overall study. FS and XZ coordinated the phenotyping. BP, MO, and XZ coordinated the genotyping. SC, JS, YY, AZ, JR, YL, JQ, GH, JZ, CW, and JC analysed the data. SC, JS, FS, and XZ drafted the manuscript. SC, JS, MO, BP, FS, and XZ interpreted the results. All authors contributed to the manuscript editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the Bill & Melinda Gates Foundation (OPP1093167). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the author accepted manuscript version that might arise from this submission. The authors gratefully acknowledge the financial support from Mexico's Secretary of Agriculture and Rural Development (SADER), the Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods (AGG) project, and the CGIAR Research Program (CRP) on MAIZE. The CGIAR Research Program MAIZE receives W1&W2 support from the Governments of Australia, Belgium, Canada, China, France, India, Japan, Korea, Mexico, Netherlands, New Zealand, Norway, Sweden, Switzerland, United Kindom, United States, and the World Bank. The authors also thank the National Natural Foundation of China (Grant Numbers: 31771814 and 31801442), the China Scholarship Council, the China Postdoctoral Science Foundation (Grant Numbers: 2017M621318 and 2018M643774), the Natural Foundation of Heilongjiang Province (Grant Number: C2017063), the Postdoctoral Science Foundation of Heilongjiang Province (Grant Number: LBH-Z17205), and the Distinguished Youth Fund of Heilongjiang Academy of Agricultural Sciences (Grant Number: 2019JCQN004).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.672525/full#supplementary-material

REFERENCES

Alvarado, G., Rodríguez, F. M., Pacheco, A., Burgueño, J., Crossa, J., Vargas, M., et al. (2020). META-R: a software to analyze data from multi-environment plant breeding trials. *Crop J.* 8, 745–756. doi: 10.1016/j.cj.2020.03.010

Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., et al. (2021). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theor. Appl. Genet.* 134, 279–294. doi: 10.1007/s00122-020-03696-9

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

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, 2633–2635. doi: 10.1093/bioinformatics/btm308

Cao, S., Loladze, A., Yuan, Y., Wu, Y., Zhang, A., Chen, J., et al. (2017). Genome-wide analysis of tar spot complex resistance in maize using genotyping-by-sequencing SNPs and whole-genome prediction. *Plant Genome* 10, 1–15. doi: 10.3835/plantgenome2016.10.0099

Ceballos, H., and Deutsch, J. A. (1992). Inheritance of resistance to tar spot complex in maize. *Phytopathology* 82, 505–512. doi: 10.1094/phyto-82-505

Combs, E., and Bernardo, 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

Crossa, J., Perez-Rodriguez, P., Cuevas, J., Montesinos-Lopez, O., Jarquin, D., de Los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 11, 961–975. doi: 10.1016/j.tplants.2017.08.011

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

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:e19379. doi: 10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Gen.* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

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

Gowda, M., Das, B., Makumbi, D., Babu, R., Semagn, K., Mahuku, G., et al. (2015). Genome-wide association and genomic prediction of resistance to maize lethal necrosis disease in tropical maize germplasm. *Theor. Appl. Genet.* 128, 1957–1968. doi: 10.1007/s00122-015-2559-0

Guo, R., Dhliwayo, T., Mageto, E. K., Palacios-Rojas, N., Lee, M., Yu, D., et al. (2020). Genomic prediction of kernel zinc concentration in multiple maize populations using genotyping-by-sequencing and repeat amplification sequencing markers. *Front. Plant Sci.* 11:534. doi: 10.3389/fpls.2020.00534

Han, S., Miedaner, T., Utz, H. F., Schipprack, W., Schrag, T., and Melchinger, A. (2018). Genomic prediction and GWAS of Gibberella ear rot resistance traits in dent and flint lines of a public maize breeding program. *Euphytica* 214, 1–20. doi: 10.1007/s10681-005-6149-0

Hock, J., Dittrich, U., Renfro, B. L., and Kranz, J. (1992). Sequential development of pathogens in the maize tar spot disease complex. *Mycopathologia* 117, 157. doi: 10.1007/BF00442777

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9, 1322–1332. doi: 10.1111/j.1755-0998.2009.02591.x

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

Liu, Y., Hu, G., Zhang, A., Loladze, A., Hu, Y., Wang, H., et al. (2020). Genome-wide association study and genomic prediction of Fusarium ear rot resistance in tropical maize germplasm. *Crop J.* 9, 325–341. doi: 10.1016/j.cj.2020.08.008

Lorenzana, R. E., and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor. Appl. Genet.* 120, 151–161. doi: 10.1007/s00122-009-1166-3

Mahuku, G., Chen, J., Shrestha, R., Narro, L. A., Guerrero, K. V. O., Arcos, A. L., et al. (2016). Combined linkage and association mapping identifies a major QTL (qRtsc8-1), conferring tar spot complex resistance in maize. *Theor. Appl. Genet.* 129, 1217–1229. doi: 10.1007/s00122-016-2698-y

Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 265–279. doi: 10.1016/j.cj.2015.01.001

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

Pereyda-Hernández, J., Hernández-Morales, J., Sandoval-Islas, J. S., Aranda-Ocampo, S., de León, C., and ómez-Montiel, N. G. (2009). Etiología y manejo de la mancha de asfalto (Phyllachora maydis Maubl.) del maíz en Guerrero, México. *Agrociencia* 43, 511–519.

Prasanna, B. M., Chaikam, V., and Mahuku, G. (2012). *Doubled Haploid Technology in Maize Breeding: Theory and Practice*. Mexico: CIMMYT.

Riedelsheimer, C., Endelman, J. B., Stange, M., Sorrells, M. E., Jannink, J. L., and Melchinger, A. E. (2013). Genomic predictability of interconnected biparental maize populations. *Genetics* 194, 493–450. doi: 10.1534/genetics.113.150227

Santantonio, N., Atanda, S. A., Beyene, Y., Varshney, R. K., Olsen, M., Jones, E., et al. (2020). Strategies for effective use of genomic information in crop breeding programs serving Africa and south Asia. *Front. Plant Sci.* 11:353. doi: 10.3389/fpls.2020.00353

Sitonik, C., Suresh, L. M., Beyene, Y., Olsen, M., Makumbi, D., Oliver, K., et al. (2019). Genetic architecture of maize chlorotic mottle virus and maize lethal necrosis through GWAS, linkage analysis and genomic prediction in tropical maize germplasm. *Theor. Appl. Genet.* 132, 2381–2399. doi: 10.1007/s00122-019-03360-x

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, 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:e1004982. doi: 10.1371/journal.pgen.1004982

Sun, Y., Wang, J., Crouch, J. H., and Xu, Y. (2010). Efficiency of selective genotyping for genetic analysis and crop improvement of complex traits. *Mol. Breed.* 26, 493–511. doi: 10.1007/s11032-010-9390-8

Swarts, K., Li, H., Navarro, J. A. R., An, D., Romay, M. C., Hearne, S., et al. (2014). Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *Plant Genome* 7, 1–13. doi: 10.3835/plantgenome2014.05.0023

Technow, F., Bürger, A., and Melchinger, A. E. (2013). Genomic prediction of Northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3 Genes Genom. Genet.* 3, 197–203.

Vélez-Torres, M., García-Zavala, J. J., Hernández-Rodríguez, M., Lobato-Ortiz, R., López-Reynoso, J. J., Benítez-Riquelme, I., et al. (2018). Genomic prediction of the general combining ability of maize lines (*Zea mays* L.) and the performance of their single crosses. *Plant Breed.* 137, 379–387. doi: 10.1111/pbr.12597

Wang, N., Liu, B., Liang, X., Zhou, Y., Song, J., Yang, J., et al. (2019). Genome-wide association study and genomic prediction analyses of drought stress tolerance in China in a collection of off-PVP maize inbred lines. *Mol. Breed.* 39, 113.

Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., et al. (2020). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci. Rep.* 10:16308. doi: 10.1038/s41598-020-73321-8

Wu, Y., San Vicente, F., Huang, K., Dhliwayo, T., Costich, D. E., Semagn, K., et al. (2016). Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. *Theor. Appl. Genet.* 129, 753–765. doi: 10.1007/s00122-016-2664-8

Yuan, Y., Cairns, J. E., Babu, R., Gowda, M., Makumbi, D., Magorokosho, C., et al. (2019). Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in maize. *Front. Plant Sci.* 9:1919. doi: 10.3389/fpls.2018.01919

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

Zhang, X., Pérez-Rodríguez, P., Semagn, K., Beyene, Y., Babu, R., López-Cruz, M. A., et al. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low density and GBS SNPs. *Heredity* 114, 291–299. doi: 10.1038/hdy.2014.99

Zhao, Y. S., Gowda, M., Liu, W. X., Wurschum, T., Maurer, H. P., Longin, F. H., et al. (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theor. Appl. Genet.* 124, 769–776. doi: 10.1007/s00122-011-1745-y

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 Plant Science

OPINION
published: 06 August 2021
doi: 10.3389/fpls.2021.717284

Opinionated Views on Genome-Assisted Inference and Prediction During a Pandemic

*Daniel Gianola**

Department of Animal and Dairy Sciences, University of Wisconsin-Madison, Madison, WI, United States

Keywords: genomic selection, prediction, quantitative genomics, animal breeding, plant breeding, machine learning, bayesian methods, best linear unbiased prediction

OPEN ACCESS

Edited by:
Johannes W. R. Martini,
International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Rodolfo Juan Carlos Cantet,
University of Buenos Aires, Argentina
Giovanny Eduardo Covarrubias Pazaran,
Bayer CropScience, Belgium
Valentin Wimmer,
KWS Saat, Germany

****Correspondence:***
Daniel Gianola
gianola@ansci.wisc.edu

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *30 May 2021*
Accepted: *30 June 2021*
Published: *06 August 2021*

Citation:
Gianola D (2021) Opinionated Views on Genome-Assisted Inference and Prediction During a Pandemic. Front. Plant Sci. 12:717284. doi: 10.3389/fpls.2021.717284

Genome-assisted prediction of complex (e.g., quantitative) traits is an ingredient of "Genomic Selection," a paradigm adopted successfully in animals and plants of agricultural importance. The approach has impacted the timing of selection decisions, and it has delivered improvements in the quality of predictions ("accuracy") relative to what can be attained by the use of pedigrees and phenotypes. It has enhanced the rate of response to genetic selection and spectacularly so in dairy cattle, at least as suggested by genome-based estimates of genetic change. Researchers have spent much effort in developing and adapting prediction machines, and the author will focus on this matter, with mild excursions into tangential issues. The material is organized into nine sections and, since the author was to opine, it represents a set of personal views, rather than a review of literature, made retrospectively.

1. Deconstruction of "genetic architecture": Molecular genetics and biochemistry confirm that the theory of quantitative genetics provides just a linear (local) approximation to complexity with little (if any) mechanistic value. The intricate interactions and feedbacks inherent in biological systems cannot be captured by simple linear regressions, even if highly-dimensional regression models are fitted to the data. The effective dimension of a model cannot exceed the sample size. For example a model with 5 million parameters run with a sample size of 500 does not provide meaningful estimates of more than 500 distinct estimable functions of parameters: individual site effects are not likelihood-identified. The view that quantitative genomics can unravel the "genetic architecture" of complex traits by providing an inventory of allelic frequencies and allelic substitution effects, or by a decomposition of variance (typically complicated by strong linkage disequilibrium) is equivalent to stating that tons of bricks, steel, and glass can represent Zaha Hadid's new Beijing airport or Frank's Gehry's Guggenheim Museum at Bilbao. The author often refraines from using the buzz term "*genetic architecture*" and favores "*statistical architecture*" instead.

2. Crumbs are not bread: The QTL paradigm [superseded by zillions of genome-wide association study (GWAS) in human genetics] has had a minor impact on agricultural practices (fertilization, management, etc.), with few exceptions. GWAS with single-marker regression is also insufficient because it accounts for little genetic variation (except for major effect variants at intermediate frequencies, which are "caught" by observation anyhow), apart from ignoring interactions as stated above. Although a more complex model may improve learning, the author has not seen reports where variable selection methods and members of the Bayesian alphabet capture signals much more effectively than a simple GWAS run with large samples, as in human genetics consortia. Shrinkage methods are typically "vector optimized" (with ridge regression notoriously so), and the borrowing of information facilitated by proper priors tends to make signals similar to each other. Bayesian variable selection (BVS) with spike-slab distributions may be more powerful, but signals from large-effect variants are strengthened at the expense of mitigating small effects. In BVS or LASSO, the "richer get richer and the poorer get poorer" whereas ridge regression is more "social democrat," making effect-size estimates similar to each other. If a pizza for 500 persons is divided into 5 million unexpected guests, each will end up getting a crumb. The perception of the author is that advances in the resolution and causality of small-effect variants *via* GWAS and genome-enabled prediction have been marginal, at least in agriculture.

3. Corroboration vs. induction: The main contributions of quantitative genetics (genomics) have been in description, prediction, and decision, e.g., selection choices, inbreeding management, and optimum contribution theory, as opposed to inference. In predictive approaches, genomic heritability or correlations take the role of "regularization knobs" (i.e., not viewed as parameters with existential meaning) for constructing prediction machines. The objective is to make statements about yet-to-be-observed phenotypes based on some training data. Predictions can be calibrated empirically, but inferences cannot. How can one say that an estimate of an entelechy, such as heritability is bad or good? Following Descartes: "*I cannot be observed, therefore, I do not exist*." According to Encyclopedia Britannica, for Descartes to prove that heritability exists, one must assume it does. For prediction, "there is no need for that hypothesis" as often attributed to Laplace.

4. Occam's razor resurrected: A less recognized but important ingredient of the study on genomic selection of Meuwissen et al. (2001), was the use of predictive cross-validation employed earlier in plant breeding but almost completely ignored in animal breeding. In the latter field, the ideas of Henderson (1963, 1973, 1984) encouraged work in developing more complex and bigger models, based on the (incorrect) perception that bigger was better. An example is multiple-trait longitudinal models for dairy cattle, producing cow-specific curves at the genetic and environmental levels for several lactations in hundreds of thousands of genetically related cows! Little attention had been devoted to evaluating whether or not a simple model would predict better than a bigger one. The use of cross-validation in genome-enabled prediction debunked the widespread perception. Big complex models make more assumptions and, with finite sample sizes, it is not uncommon that such models lack robustness, thus failing to deliver better predictions. During the 20th century, model choice received scant formal consideration in animal breeding, a notable exception being a study by Sorensen and Waagepetersen (2003). Genomic selection with cross-validation helped to refute older views. Simplicity can be effective and is often elegant.

5. Prediction is inclusive: There is no universally best genome-based prediction machine for animal or plant breeding. The relative performance of the various methods depends mainly on the information content and structure of the training set, and on the extent to which a configuration of genotypes spanned in the training process will also appear in the testing set. These two aspects are difficult to evaluate *ex ante*. Often, the size of the training sample or functions thereof, e.g., Daetwyler et al. (2008), are used as a proxy for the "expected quality" of predictions. However, a sample may be huge and yet convey little information. The plant breeding group in Munich has worked (e.g., Auinger et al., 2021) in assessing genomic measures of information content, such as molecular diversity present in a training sample, and attempting to connect these metrics to predictive outcomes. For instance, a strong underlying structure may affect prediction adversely, even in large samples, so conceivably it could be modified to enhance the quality of outcomes. The larger the overlap between training and testing samples, the more relevant to a target population the statements made from training data will be. George Box and Norman Draper (my teacher in a regression course I took in 1972) taught: "*Never extrapolate beyond the experimental region*". Suppose a prediction machine "sees" 50% *AABB* and 50% *aabb* individuals in the training process. However, the testing set has the configuration $\frac{1}{3}AABB + \frac{1}{3}AaBb + \frac{1}{3}aabb$. Both sets have the same allelic frequencies, but the testing set contains a "novelty," *AaBb*, so the prediction machine would be extrapolating. Genetic relatedness is a measure of such overlap, but the driving force is the degree of molecular similarity between individuals in the corresponding data partitions. Random replication of cross-validation may produce an estimate of an upper bound for predictive ability. Even when both training and testing sets are representative of a target population, the performance of prediction methods often depends on cryptic interplays between environment, trait and model complexity (effective number of parameters fitted vis-a-vis effective training sample size).

It is futile to have information-rich training samples but unrepresentative and ridiculously small testing sets, as large variation among outcomes of similar prediction exercises is to be expected. Small testing sets and failure to replicate cross-validation in some studies have produced results where models accommodating dominance and epistasis appear as delivering a somewhat better performance than additive prediction models. Such results may be "false positives" reflecting chance, rather than signal.

6. And the Oscar goes to. . . : A simple method such as genomic best linear unbiased predictor (GBLUP) may tell something about the state of nature and perform adequately. An involved procedure, such as a deep neural network (DNN), may tell nothing and yet produce spectacular results, although it has failed miserably in some studies. Like all neural networks, a DNN is regarded as "universal approximator." An ongoing meta-analysis of hundreds of studies made in INIA, Spain (disclaimer: I will be a coauthor) places reproducing kernel Hilbert spaces regression (RKHS); e.g., Wahba (2007) methods ahead of others, but only slightly. Work with animals and plants and with various field crops in CIMMYT (e.g., Costa-Neto et al., 2021) has shown the flexibility of kernel methods for capturing genome-environment interactions and environmental similarities. In Wisconsin, RKHS has been extended to the single-step BLUP setting, and the CIMMYT group is developing a multiple-trait Bayesian RKHS. Last, but not the least, RKHS is the mother of GBLUP, along the lines that Gibbs sampling is a child of the Metropolis-Hastings algorithm for Markov chain Monte Carlo sampling.

Animal breeding industries have embraced GBLUP, and there seems to be little scope for adoption of the Bayesian alphabet models (the membership of this club is converging to infinity) for routine use, but there can be exceptions. GBLUP is a "good thing" as pointed out in the early '90s, and we have known for a while that it is not only a special case of RKHS, but also a maximum penalized likelihood estimator, a linear neural network, and that it has a Bayesian interpretation. It has been extended to cross-sectional, multi-trait, longitudinal situations and has been "robustified." Importantly, software developed mainly at the University of Georgia by Misztal allows crunching millions of predicted genomic breeding values. GBLUP with mild tweeks will probably remain the technology (term used deliberately) of choice for genetic evaluation of selection candidates. The science of genome-enabled prediction has arrived at a reasonable destination, but the voyage will continue, and new data will bring challenges.

7. Help needed: Despite an abundance of chips enabling large-scale genotyping, training samples are seldom drawn at random, thus unrepresentative. This situation constitutes a selection process that is often not considered in predictive models. Animal breeders have been widely influenced by the "selection bias" study of Henderson (1975), based on questionable assumptions, as pointed out first by Robin Thompson (1979). *Ad-hoc* approaches and arguments have been used for justifying some forms of analysis or modeling, such as the notion of treating a large number of contemporary groups as fixed, leading to inefficient estimation (James-Stein "inadmissibility" argument; Judge et al., 1985). The arguments were based on an obscure notion of bias removal advocated by Henderson. Such views have carried into genome-enabled prediction in animal breeding. The problem of employing selected samples for inference and prediction stands and should be studied with more rigor, e.g., *via* missing data theory.

8. Bias against bias: The notion of statistical "bias" continues to be misunderstood. GBLUP is believed (just consider its name) to be an unbiased predictor, but it is identical to ridge regression in some settings. However, ridge regression is a biased estimator. Is this a Dr. Jekyll Mr. Hide issue or some statistical bipolarity? The answer is that prediction and estimation unbiasedness have different definitions! Say you own a plant or a bull called "Charlie," a fixed entity with identity (e.g., if Charlie is *AA*, it has a specific breeding value that possibly differs from that of *Aa* or *aa*). You are not interested in learning the average of a (very) large sample of potential Charlies; rather, you seek the breeding value of the Charlie you have. If ridge regression is used to estimate the breeding value of Charlie, Dr. Jekyll says there is estimation bias, but Mr. Hide states that there would be none. The latter is wrong (in the bias sense) with respect to Charlie, but not with respect to an average of potential Charlies, some of which will be *AA*, some *Aa*, and some *aa*. All good members of the Bayesian alphabet including GBLUP, with its appealing Bayesian interpretation (Gianola and Fernando, 1986), and practically all machine learning methods (e.g., random forests) provide biased predictions that, on average, will be better than unbiased machines. A potential therapy for unbiasedness-obsession is "debiasing" (Breiman, 2001). However, predictions would be probably worse because, in addition to the extant uncertainty of prediction sets, there would be an extra error resulting from a deteriorated bias-variance trade-off. In the end, the debiased genome-enabled predictions may be much worse than prior to bias removal.

9. Use a GPS to map the road ahead: Defining pertinent breeding objectives (the classical Smith-Hazel problem) continues being crucial in practice, but it has become academically non glamorous at these times of massive genotyping, epigenotyping, proteomics, metabolomics, and (fine) phenotyping. It is important not to lose perspective as, otherwise, breeders will get inebriated with a cocktail "on apps." Another issue of (some) concern is that the current emphasis on "big data," "massive computing," and "visualization" may diminish basic science education, as it appears that current thesis students start crunching numbers before they know genomics, or the meaning of a probability distribution, or attain an elementary knowledge of randomization or causality. Foundational theory and concepts should continue being taught. Otherwise, the field may drown in a technology-induced maelstrom, and critical or even visionary perspectives may end up playing a role that becomes secondary to that of a beautiful visualization or, even worse, to a machine.

W. G. Hill ("Bill") noted in a 2010 study discussing from Lush to Genomics: "Opinions we can debate." I look forward to that conversation (Hill, 2010).

AUTHOR CONTRIBUTIONS

DG prepared the first draft of the opinion, provided critical comments, typed the definitive version, and checked the spelling. The author contributed to the article and approved the submitted version.

FUNDING

Financial support was provided by the MTA account, Department of Animal and Dairy Sciences, University of Wisconsin-Madison. Salary support was provided by the Social Security Administration, the Wisconsin Retirement System, and the SURS Illinois Retirement Program.

ACKNOWLEDGMENTS

Dr. Johannes Martini (Aardevo, the Netherlands), Professors Chris-Carolin Schön (Technische Universität München), Miguel Pérez-Enciso (Universitat Autònoma de Barcelona), Bruce Walsh (University of Arizona) and three reviewers are thanked for useful comments.

REFERENCES

Auinger, H.-J., Lehermeier, C., Gianola, D., Mayer, M., Melchinger, A. E., da Silva, S., et al. (2021). Calibration and validation of predicted genomic breeding values in an advanced cycle maize population. *Theor. Appl. Genet.* doi: 10.1007/s00122-021-03880-5. [Epub ahead of print].

Breiman, L. (2001). Using iterated bagging to debias regessions. *Mach. Learn.* 45, 261–277.

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2021). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trial. *Heredity* 126, 92–106. doi: 10.1038/s41437-020-00353-1

Daetwyler, H. D., 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

Gianola, D., and Fernando, R. L. (1986). Bayesian methods in animal breeding theory. *J. Anim. Sci.* 63, 217–244.

Henderson, C. R. (1963). "Selection index and expected genetic advance," in *Statistical Genetics and Plant Breeding*, eds W. D. Hanson and H. F. Robinson (Washington, DC: The National Academy of Sciences; The National Research Council), 141–163.

Henderson, C. R. (1973). "Sire evaluation and genetic trends," in *Proceedings of the Animal Breeding and Genetics Symposium* in Honor of Dr. Jay L. Lush (Champaign: American Society of Animal Science; American Dairy Science Association), 10–41.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–449.

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph, ON: University of Guelph.

Hill, W. G. (2010). Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. *Genetics* 196, 1–6. doi: 10.1534/genetics.112.147850

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., and Lee, T.-C. (1985). *The Theory and Practice of Econometrics*, *2nd Edn*. New York, NY: Wiley.

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

Sorensen, D., and Waagepetersen, R. (2003). Normal linear models with genetically structured residual variance heterogeneity: a case study. *Genet. Res.* 82, 207–222. doi: 10.1017/S0016672303006426

Thompson, R. (1979). Sire evaluation. *Biometrics* 35, 339–353.

Wahba, G. (2007). *Statistical learning in medical data analysis*. Technical Report 1136. Department of Statistics, University of Wisconsin, Madison, WI.

Conflict of Interest: The author declares 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 collaboration with the author DG.

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 Plant Science

ORIGINAL RESEARCH
published: 09 August 2021
doi: 10.3389/fpls.2021.672417

Understanding the Effectiveness of Genomic Prediction in Tetraploid Potato

Stefan Wilson[1], *Chaozhi Zheng*[1], *Chris Maliepaard*[2], *Han A. Mulder*[3], *Richard G. F. Visser*[2], *Ate van der Burgt*[4] and *Fred van Eeuwijk*[1]*

[1] Biometris, Wageningen University & Research Centre, Wageningen, Netherlands, [2] Plant Breeding, Wageningen University and Research, Wageningen, Netherlands, [3] Wageningen University and Research Animal Breeding and Genomics Centre, Wageningen, Netherlands, [4] Solynta, Wageningen, Netherlands

Use of genomic prediction (GP) in tetraploid is becoming more common. Therefore, we think it is the right time for a comparison of GP models for tetraploid potato. GP models were compared that contrasted shrinkage with variable selection, parametric vs. non-parametric models and different ways of accounting for non-additive genetic effects. As a complement to GP, association studies were carried out in an attempt to understand the differences in prediction accuracy. We compared our GP models on a data set consisting of 147 cultivars, representing worldwide diversity, with over 39 k GBS markers and measurements on four tuber traits collected in six trials at three locations during 2 years. GP accuracies ranged from 0.32 for tuber count to 0.77 for dry matter content. For all traits, differences between GP models that utilised shrinkage penalties and those that performed variable selection were negligible. This was surprising for dry matter, as only a few additive markers explained over 50% of phenotypic variation. Accuracy for tuber count increased from 0.35 to 0.41, when dominance was included in the model. This result is supported by Genome Wide Association Study (GWAS) that found additive and dominance effects accounted for 37% of phenotypic variation, while significant additive effects alone accounted for 14%. For tuber weight, the Reproducing Kernel Hilbert Space (RKHS) model gave a larger improvement in prediction accuracy than explicitly modelling epistatic effects. This is an indication that capturing the between locus epistatic effects of tuber weight can be done more effectively using the semi-parametric RKHS model. Our results show good opportunities for GP in 4x potato.

Keywords: tetraploid potato, genotype by sequencing, genomic prediction, genome wide association study, non-additive effects

OPEN ACCESS

Edited by:
Fernando H. Toledo,
International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Guilherme da Silva Pereira,
Federal University of Viçosa, Brazil
Li Li,
University of New England, Australia

***Correspondence:**
Fred van Eeuwijk
fred.vaneeuwijk@wur.nl

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *25 February 2021*
Accepted: *13 July 2021*
Published: *09 August 2021*

Citation:
Wilson S, Zheng C, Maliepaard C, Mulder HA, Visser RGF, van der Burgt A and van Eeuwijk F (2021) Understanding the Effectiveness of Genomic Prediction in Tetraploid Potato. Front. Plant Sci. 12:672417. doi: 10.3389/fpls.2021.672417

INTRODUCTION

Cultivated potato (*Solanum tuberosum* L.) is one of the most consumed food crops in the world, behind only rice and wheat (Birch et al., 2012). Since its discovery over 500 years ago, breeders have selected and hybridised this crop to adapt to various environmental conditions and satisfy numerous market desires. With its large genetic diversity, this was easily achieved making potato one of the most versatile food crops. Most of the environmental and market class adaptations, as well as genetic gains for simple traits, have been attained via phenotypic selection, which may take 10–12 years until a new cultivar is introduced (Jansky, 2009; Endelman et al., 2018). However, there has been limited progress for more quantitative traits with lower heritabilities, for example yield Jansky (2009). Genomic prediction (GP), where phenotypes are regressed on marker profiles (Bernardo, 1996; Whittaker et al., 2000; Meuwissen et al., 2001), allows for the early selection or discarding of favourable or unfavourable hybrids, and therefore significantly speeds up the breeding cycle (Hickey et al., 2017).

Genomic prediction has seen more application in animal breeding in comparison to plant breeding and has rarely been applied to polyploid species until recently. Cultivated potato is an autotetraploid, and the patterns of inheritance in autotetraploids are more complicated than diploids and allotetraploids (Gallais, 2003; Garcia et al., 2013; Dufresne et al., 2014), hence the reason for the smaller number of GP studies among these species. Despite the obstacles, GP has recently been put to use in a number of autopolyploid crops including alfalfa (Annicchiarico et al., 2015), potato (Habyarimana et al., 2017; Sverrisdóttir et al., 2017; Enciso-Rodriguez et al., 2018; Endelman et al., 2018; Amadeu et al., 2020), blueberry de Bem Oliveira et al. (2019), Amadeu et al. (2020), and tetraploid ryegrass Guo et al. (2018).

Despite the common theme of past studies, in that they look at GP in autopolyploids, they differ in more ways than just the species they focus on. This study intends to merge some of the principles used in previous studies. Genotype by sequencing (GBS) has been utilised previously in the study of GP of autopolyploid crops (Annicchiarico et al., 2015; Sverrisdóttir et al., 2017; Guo et al., 2018), and will be implemented in this study as the method for investigating DNA variation. One difficulty encountered in quantitative genetics for polyploids is the determination of allele dosage. Recent studies have investigated methods to deal with this problem (Endelman et al., 2018; Guo et al., 2018; de Bem Oliveira et al., 2019) by looking directly at allele frequencies and refraining from performing discrete genotype calling. This study also directly examines allele frequencies, but uses a probabilistic approach for determining the most likely dosage based on allele frequency ratios.

Statistical models used for GP face the scenario where $n << p$, therefore penalties are introduced for reliable estimation of marker effects, which require assumptions on the parametric distribution of these marker effects (Piepho, 2009). The most common GP model is known as GBLUP (Genomic best linear unbiased predictor), a mixed model, where the relationship between cultivars is used as input, and is equivalent to using a ridge regression penalty with an assumed normal distribution for marker effects (Piepho, 2009). A relationship matrix can be derived assuming additive effects and non-additive effects (dominance and epistasis). We investigate the impact of explicitly accounting for non-additive effects (Enciso-Rodriguez et al., 2018; Endelman et al., 2018; Amadeu et al., 2020) vs. implicitly modelling these non-additive effects using the semi-parametric Reproducing Kernel-Hilbert Space (RKHS) model (Gianola and van Kaam, 2008; Habyarimana et al., 2017). Another relationship matrix has been proposed for autotetraploids, that assumes separate genotype effects for each marker (Slater et al., 2016) which also implicitly captures non-additive effects and is included in this study. Bayesian models are also included in this study, to compare the impact of different prior assumptions on the distribution of marker effects (Pérez and de los Campos, 2014).

For GP, there is no "one-size-fits-all" model that works best, and instead the performance of models depends primarily on trait architecture (de los Campos et al., 2013). Unlike many GP studies, we extend this study to include a Genome Wide Association Study (GWAS), to describe the architecture of each trait and explain the differences in the performance of the various GP models. Applying GWAS to markers coded for different types of dominance (Rosyara et al., 2016), we attempt to identify the source of dominance effects, for those traits that were more accurately predicted with GP models that included non-additive effects. GWAS will also reveal the level of association between our markers and a particular trait, to understand why a GP model that estimates marker effects performs better than a model that estimates genotype effects or vice versa.

We aim to demonstrate the feasibility of GP in autotetraploid potato in this proof-of-concept study. Using four traits and GBS marker data, various modelling strategies will be compared to uncover the model or models most suitable for a given trait. To comprehend the relationship between a trait and its most suitable model, a GWAS is used to describe the genetic architecture of the traits, providing some insight as to why some modelling strategies might work better for particular traits.

MATERIALS AND METHODS

Plant Materials

A diversity panel of 147 tetraploid potato cultivars, including recent Dutch breeding material were chosen for this study. This subset of cultivars are representative of the worldwide commercial potato germplasm and were selected based on criteria such as: phenotypic diversity of important traits, country of origin, market category (chip and French fry processing, cooking and starch varieties), year of commercial introduction, and availability of the cultivars. Some of these varieties were analysed in previous studies that used similar criteria for selection (D'hoop et al., 2008, 2014). Propagation was done by two Dutch breeding companies, one of which had also performed phenotyping and collecting the biological material needed for genotyping.

Genotypic Information

DNA material (100 ng) was digested with ten units of EcoT22 (Clontech) and incubated at 37°C for 2 h and then heat killed. Samples were then ligated with 640 units of T4 ligase (NEB) and phased adaptors with TGCA overhangs at 22°C for 1 h and heat killed. The ligated samples were diluted in the ratio 1:10 with water, and then amplified for 18 cycles to add barcodes. Barcoded libraries were SPRI purified, quantified, and pooled in groups of 48 samples. Pooled samples were SPRI purified, quantified, and diluted to 2 nM for sequencing on the Illumina HiSeq 2500 using single-end 1×100 reads. Sequence reads were mapped against the potato reference genome sequence of DM v4.04, including the chloroplast and mitochondrial sequences using Burrows-Wheeler Aligner 0.7.12. After the removal of monomorphic markers, those with more than two alleles and

TABLE 1 | Two examples of genotype probabilities based on allele counts.

Allele count		Genotype probabilities				
Reference	Alternative	AAAA	AAAB	AABB	ABBB	BBBB
15	13	0	0.05	0.94	0.01	0
15	0	0.99	0.01	0	0	0

markers from repetitive regions of the genome, 870 thousand bi-allelic markers were available for further filtering. Markers with minor allele frequency <0.01 and those with read depths <10 or >100 were removed. From the remaining markers, the posterior probability of allele dosage, conditional on both allele counts and sequencing error, was calculated (see **Supplementary Material** for more details). This will be referred to as the genotype assignment probability. Tetraploid genotypes can belong to either of the classes AAAA, AAAB, AABB, ABBB, BBBB, where "A" and "B" are the reference and alternative allele, respectively. If there is an equal amount of counts for both alleles we would infer the genotype to be AABB (see example in **Table 1**). Similar methodology is applied in the PolyOrigin software (Zheng et al., 2020).

Genotype assignment probabilities were used as a filter criterion. For each individual, markers were removed when the highest genotype assignment probability was below a threshold. Stricter thresholds created more missing information and decreased the number of markers, since markers without information for more than 25% of the individuals were removed. Allele dosage was then determined as the dosage with maximum genotype probability. Probability thresholds of 0.85, 0.75, and 0.5 resulted in marker matrices of 19, 26, and 39 thousand markers, respectively. Using an additive GBLUP model, a preliminary GP analysis was performed to decide which marker matrix should be used as there may be a trade-off between the quantity and quality of markers. In almost all cases, the 39 K marker matrix gave the most accurate predictions and will henceforth be used for all analyses (**Supplementary Figure 1**). The larger number of markers lends itself to a more complete coverage of the genome (**Figure 1**).

Although linkage disequilibrium (LD) was not calculated in this study, it was calculated for an overlapping panel of tetraploid potato (Vos et al., 2017). In that study, it was found that LD falls quickly and suggested 40 K markers were needed for good coverage of the tetraploid potato genome for GWAS, which is comparable to the number of markers used here.

For all analyses performed in this study, we begin with genotype information contained in the marker matrix (X), with 147 rows and 39,000 columns. Each element of X gives the discrete count of alternative alleles (0, 1, 2, 3, 4) assigned by genotype probabilities, at a given marker position for a given cultivar. When these counts are entered in a design or relationship matrix and a single parameter is estimated to quantify the dependence of the phenotype on the allele count, then this implies that marker effects are additive.

For imputing the missing marker information, the mode was used. This was compared to mean imputation using the 39 K marker set mentioned above and a GBLUP model. The GP accuracies resulting from marker matrices imputed with the mode were slightly higher than those imputed from the mean.

Phenotypic Information

Field trials were performed in 2017 and 2018 at three locations: Spain, Poland, and the Netherlands. Seed tubers were planted in plots consisting of eight plants. A row-column resolvable design was implemented with two complete blocks, and varieties dispersed across the field using latinisation over rows and columns (Piepho et al., 2015). Checks of one particular variety were uniformly distributed throughout the trial in order to detect and correct for spatial trends. Randomisation was performed using the package DiGGer (Coombes, 2009) executed with the software R (R Core Team, 2019), where all analyses were conducted with this study.

Four traits will be discussed in this study: plot tuber weight (kilograms), plot tuber count (number of tubers), mean tuber length (millimetres), and dry matter content (percentage). Adjusted means were calculated by correcting for row and column trends, as well as block effects using the model:

$$y = block + rowinblock + colinblock + G + \epsilon, \tag{1}$$

where y is a vector of phenotypic observations. Equation (1) allows us to adjust for field trends (from blocks, rows within blocks, and columns within blocks) and extract the best linear unbiased estimate (BLUE) of each genotype (G). Complete blocks were used in each trial therefore a fixed term for the block effect is suitable in the statistical model. Rows and columns within blocks were incomplete and therefore treated as random effects having normal distributions as follows: $row \sim N(0, \sigma^2_{row})$ and $col \sim N(0, \sigma^2_{col})$ where σ^2_{row} and σ^2_{col}, are row and column variances, respectively. Other non-genetic factors are captured in the random term ϵ that is assumed to be normally distributed as $\epsilon \sim N(0, \sigma^2_{\epsilon})$ where the residual variance is represented by σ^2_{ϵ}.

For investigating genotype by environment interaction (GxE), the BLUEs from the six trials (three locations, 2 years) will be useful, however for this application of GP, we require one vector of observations for a given trait, as if they came from one single environment. To consolidate our six phenotypic values, we again calculated the adjusted means of each genotype, after correcting for the effect of different trials using Equation (2).

$$y = trial + G + \epsilon, \tag{2}$$

where y are the BLUEs calculated from Equation (1), and ϵ captures the variation from the interaction between genotype and trial as well as within trial error variation. Equation (2) is an across trial model, while Equation (1) was used for within trial analyses. This could have been combined in one statistical model, but for future GxE applications, and the ability to carefully assess each trial for outliers, it was conducted in two steps. A comparison of BLUEs calculated from the method described here vs. one single model was done and the results were the same. The BLUEs from Equation (2) will be used as the response variable for GP analyses going forward. This study is therefore a two-step

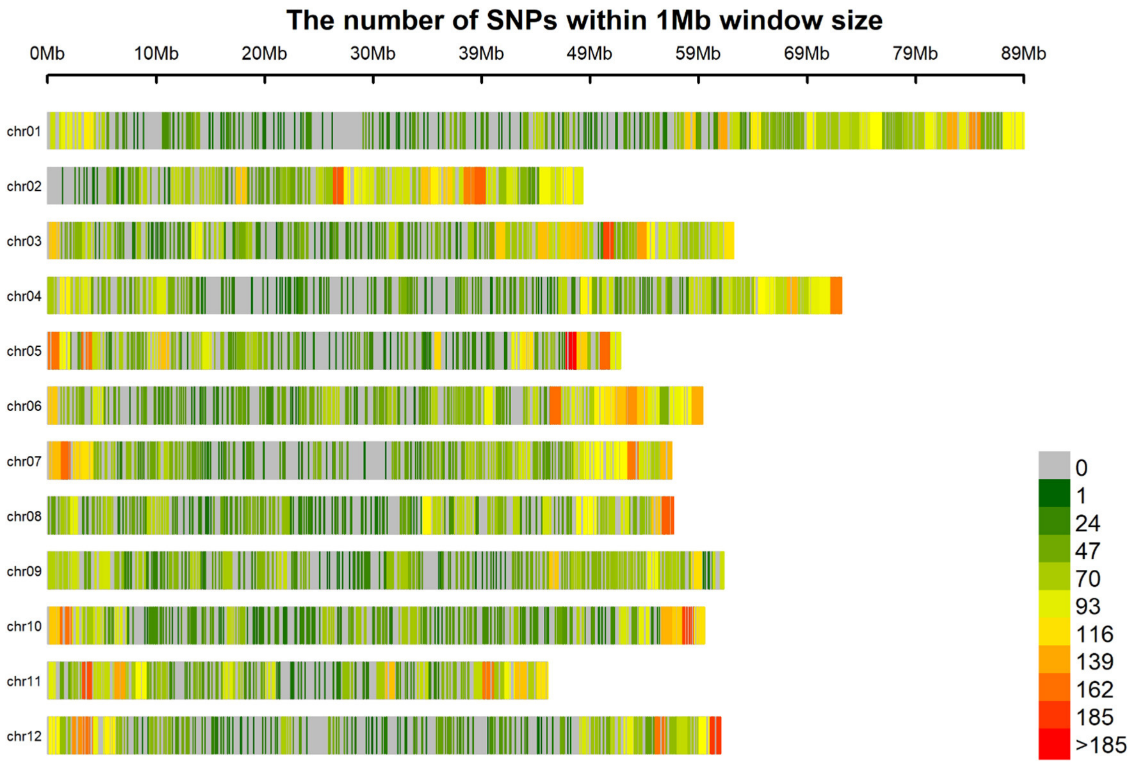


FIGURE 1 | Marker density of 39 K markers.

analysis since phenotypic adjusted means and GP are done with separate models. ASREML (Butler, 2009) was used to conduct all phenotypic analyses.

Heritability

To have an understanding of how much phenotypic variation can be attributed to between-genotype variation, broad-sense heritability was calculated. Using the BLUEs from Equation (1) as the response variable, we apply the following model across our three locations (L) and 2 years (T):

$$y = L + T + LT + \mathbf{G} + \mathbf{GL} + \mathbf{GT} + \boldsymbol{\epsilon}$$

Using the random terms of this model, highlighted in bold font, we can isolate the variability that is caused genetically from the variability that is caused from genotype by location and genotype by year interactions (GL, GT). The BLUEs from Equation (1) give only one value for each genotype per trial (year and location combined), for this reason our error term (ϵ) in the heritability equation captures the variation from the three way interaction of genotype, location, and year. Averaging a genotypic effect across multiple trials without including marker information, is closer to an estimation of repeatability than heritability (Falconer et al., 1996), but for now we shall use traditional nomenclature. Heritability was calculated from the variance components as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{gL}^2}{l} + \frac{\sigma_{gT}^2}{t} + \frac{\sigma_\epsilon^2}{l \times t}},$$

where l and t represent the total numbers of locations and years. Variation due to genotype by location and genotype by year interactions are represented by the terms $\sigma_{gL}^2, \sigma_{gT}^2$, respectively, while σ_g^2 represents genetic variance. The term σ_ϵ^2 is the variance of the three way interaction of genotype by year by location, and contains genetic signal alongside within trial variation.

Prediction Models

For GP, many types of statistical models are applicable; those that perform shrinkage vs. those that perform variable selection which is dependent on the assumed distribution of marker effects, and those models that account for non-additive effects in various ways.

- **Additive GBLUP:**

$$y = \mu + Z_a a + \epsilon, \tag{3}$$

where y is a vector of phenotypic values, μ is the overall mean, Z_a is a design matrix that relates the observations to genomic values and a is a vector of random additive genetic values with distribution $a \sim N(0, G_A\sigma_a^2)$. The additive genetic variance is given by σ_a^2, while ϵ is the vector of residual and non-modelled genetic effects, assumed to be normally distributed

TABLE 2 | Within marker locus coding for Full Tetraploid model (Slater).

Genotype	Locus coding				
AAAA	1	0	0	0	0
AAAB	0	1	0	0	0
AABB	0	0	1	0	0
ABBB	0	0	0	1	0
BBBB	0	0	0	0	1

For a genotype "AAAA," there are five columns with the first column assigned a 1 and the rest 0's.

$\epsilon \sim N(0, \sigma_\epsilon^2)$, with variance denoted by σ_ϵ^2. G_A is the additive genomic relationship matrix (from allele dosages) based on the work of VanRaden (2008) and extended by Ashraf et al. (2016). The calculation of this additive genomic relationship matrix is applicable to autotetraploids and was constructed with the R package AGHmatrix (Amadeu et al., 2016).

- **Additive + Dominance GBLUP:**

$$y = \mu + Z_a a + Z_d d + \epsilon, \tag{4}$$

where y, μ, a, and ϵ are the same as seen in Equation (3). Z_a and Z_d are design matrices to relate observations to additive genetic effects and dominance effects. The vector of dominance effects is indicated by d and follows a normal distribution: $d \sim N(0, G_D\sigma_d^2)$, where σ_d^2 is the dominant genetic variance. The digenic dominant relationship matrix G_D was built using the AGHmatrix R-package, as derived by Endelman et al. (2018).

- **Epistatic GBLUP:**

$$y = \mu + Z_a a + Z_d d + Z_e e + \epsilon \tag{5}$$

Equation (5) is an extension of Equation (4), with the inclusion of a term to capture epistatic effects. Z_e relates the observations to the epistatic effects e, which follow the normal distribution, $e \sim N(0, G_E\sigma_e^2)$ with epistatic genetic variance σ_e^2.

This paper considers first order epistasis (additive × additive), and to calculate G_E, the Hadamard product of G_A ($G_A \# G_A$) was used (Su et al., 2012; Endelman et al., 2018).

- **Full Auto-tetraploid GBLUP:**

$$y = \mu + Zf + \epsilon, \tag{6}$$

Proposed in the paper by Slater et al. (2016) is the full auto-tetraploid model which accounts for additive and non-additive effects by assuming each genotype has its own effect. Tetraploids have five possible genotypes (AAAA, AAAB, AABB, ABBB, BBBB), therefore f, the vector of effects in Equation (6), has length $5R$ where R is the number of markers (see **Table 2**). These effects f, follow the normal distribution, $f \sim N(0, G_F\sigma_f^2)$ with genetic variance σ_f^2. The details for calculation of the relationship matrix G_F can be found in the associated literature (Slater et al., 2016), and was constructed using the AGHmatrix R-package.

- **RKHS:** The model for Reproducing Kernel-Hilbert Space (RKHS) is the same as described in Equation (3), but the random genetic values have a different distribution: $a \sim N(0, K\sigma_g^2)$. The genomic relationship matrix G_A, is replaced by the kernel matrix, $K = exp^{-\frac{D}{\theta}}$, where D is a Euclidean distance matrix between genotypes, and θ a tuning parameter. The tuning parameter controls how fast the relationship between two genotypes decays as the distance between the corresponding pairs of marker vectors increases (Jiang and Reif, 2015) and is estimated from the data by maximizing the log-likelihood (Endelman, 2011). The genetic variance is no longer the result of allele substitution, as seen in the additive model (Equation 3) with additive genetic variance, σ_a^2. The genetic variance captured by RKHS (σ_g^2), includes additive and first order epistatic (additive × additive) effects (Gianola and van Kaam, 2008).
- **Bayesian LASSO:**

$$y = \mu + Xb + \epsilon \tag{7}$$

The first five genomic predictions described above estimate genotypic effects, the Bayesian models however estimate marker effects. Equation (7) includes terms for phenotype (y), overall mean (μ), and non genetic (or unmodelled) influences plus error (ϵ). Where it differs from our previous GP models is in the term Xb, which directly links the marker design matrix X, to the marker effects b. The marker effects are assumed to come from a distribution and in the case of Bayesian LASSO, a double exponential (Laplace) distribution, $b \sim Lap(0, \lambda)$, or alternatively:

$$b \sim \Pi_{j=1}^{R} \frac{\lambda}{2} e^{-\lambda|b_j|}$$

The λ parameter is inversely proportional to the variance of the distribution and is estimated from the data. The probability density function is multiplied across all markers (each indicated by subscript j), up to a total of R markers.

- **Bayes A:** The statistical model is similar to that seen in Equation (7), however Bayes A assumes that marker effects come from a scaled-t distribution with ν degrees of freedom, $b_j \sim t_\nu(0, \sigma_b^2)$, where σ_b^2 is the variance of marker effects.
- **BAYES C$\boldsymbol{\pi}$:** The Bayes Cπ model assumes that marker effects come from a mixture distribution where a proportion of markers (π) have zero effect and the remainder ($1 - \pi$) have non-zero effects from a normal distribution, such that:

$$b_j = \begin{cases} 0 & : \text{with probability } \pi \\ \sim N(0, \sigma_b^2) & : \text{with probability } 1 - \pi \end{cases}$$

Because markers are separated into either having an effect or having no effect, this model is performing marker selection. The proportion of zero effect markers π, is estimated from the data.

The three Bayesian models are better suited for traits controlled by few large effect loci, whereas the models mentioned before are better for predicting traits with many small effect loci

TABLE 3 | Marker configurations under different effect assumptions.

	AAAA	AAAB	AABB	ABBB	BBBB
Additive	0	1	2	3	4
Simplex dominant (B>A)	0	4	4	4	4
Duplex dominant (B>A)	0	0	4	4	4

(de los Campos et al., 2013). For all GP models, except RKHS, parameters were estimated using Bayesian statistics (Gibbs Sampler) with the package BGLR (Pérez and de los Campos, 2014), with 10,000 iterations and 2,500 iterations used as burn-in. Maximum likelihood was used to implement the RKHS model, and choose the most likely value for the tuning parameter.

Assessing Prediction Accuracy

With 147 varieties containing both phenotypic and genotypic information, cross-validation was performed by sampling a training set of 105 individuals to train the model, and using the trained model to predict the remainder of individuals (validation set) (Wilson et al., n.d.). These 105 individuals were sampled in order to minimise the genetic distance between the training and validation sets, using a sampling method based on the coefficient of determination (Rincent et al., 2012). This training set construction procedure uses marker information in the form of a genomic relationship matrix, as well as phenotypic information to construct the training set. Prediction accuracy is defined as the Pearson correlation between the BLUEs and the predicted genotypic values, and was averaged over the 100 repetitions.

Genome Wide Association Study (GWAS)

To suggest an explanation for the differences between GP models, a GWAS was performed to investigate the genetic architecture of the traits analysed. The proposed GP models assume different biological processes for controlling trait expression: many small effect loci vs. a few large QTLs as well as additive vs. dominant effects. For a given trait, the genetic architecture uncovered by GWAS will help explain why a particular GP model has higher prediction accuracy than another.

$$y = \mu + X\beta + \boldsymbol{g} + \epsilon \tag{8}$$

In Equation (8), y is the vector of BLUEs, μ is the overall mean. The polygenic effect is captured by the term g, and is distributed $g \sim N(0, G_A\sigma_g^2)$, where G_A is the same genomic relationship matrix across all chromosomes, used for the GBLUP prediction in Equation (3). The error term ϵ captures non-genetic residuals plus error, and is assumed to follow a normal distribution as seen in prior models. The term β represents the marker effect and X is the marker matrix containing genetic information that may be coded differently depending on the assumed type of effect (see **Table 3**).

From **Table 3**, we see the coding of the design matrix, where the additive effect assumes the size of the effect is proportional to the number of alternative alleles present. Simplex dominant (for the alternative allele) indicates that there are two levels for effects: when there is no alternative allele present and another for when there is at least one alternative allele. This simplex dominant configuration of allele effects corresponds with our GBLUP dominance prediction model (Equation 4). Duplex dominance means that the second level of effect occurs when at least two alternative alleles are present. Duplex dominance was not included in any GP models, however exploring the level of dominance can reveal genetic architecture information and therefore, help explain the differences between GP accuracies, allowing for expansion in future studies. For both simplex and duplex dominance, the reference allele was also regarded as the dominant allele, and therefore five different SNP design matrices (additive, simplex dominance for the reference allele, simplex dominance for the alternative allele, duplex dominance for the reference allele and duplex dominance for the alternative allele) were used in the GWAS of each trait. This analysis was done using the GWASpoly package (Rosyara et al., 2016).

The impact of population structure on the GWAS analysis was evaluated by looking at the quantile-quantile plots of the p-values for marker effects transformed to a log scale ($-log_{10}p$). Not correcting for population structure will result in spurious associations, and this was investigated by a visual assessment for inflation of p-values.

The threshold for identifying significantly associated markers was corrected for multiple testing using the method proposed by Li and Ji (2005). This is calculated as the significance level divided by the number of effective regions ($\frac{\alpha}{N_{eff}}$), where N_{eff} is estimated from the eigen values of the marker matrix. This resulted in 222 effective regions from the 39,000 markers. For each marker effect assumption (additive, simplex dominance etc.), significant markers were extracted and used as explanatory variables, along with the first three principal components (extracted from the relationship matrix constructed on allelic dosage), in a linear regression model. The R^2 statistic of this model is the fraction of the total sum of squares due to genotypic differences, that can be explained by markers (Wallace et al., 2016; Inostroza et al., 2018). For a given trait, we will be able to distinguish which effect, additive, dominance, or the effect of population structure, explains more of the phenotypic variance.

RESULTS

Population Structure

Using the marker matrix X (described previously in the Materials and Methods) an assessment of population structure was conducted via Principal Components analysis (**Figure 2**), analysis of molecular variance (AMOVA) and Wright's F_{ST} statistic (**Table 4**). A list of the seven distinct market classes are as follows, with the number of individuals belonging to each class given in parentheses: ancient (1), chip processing (39), French fry processing (42), fresh consumption (1), cooking (56), starch (7), and the rest (1).

Figure 2 illustrates that there is a lack of separation between market classes. For F_{ST} and AMOVA calculations, the three small market classes were not included as they did not

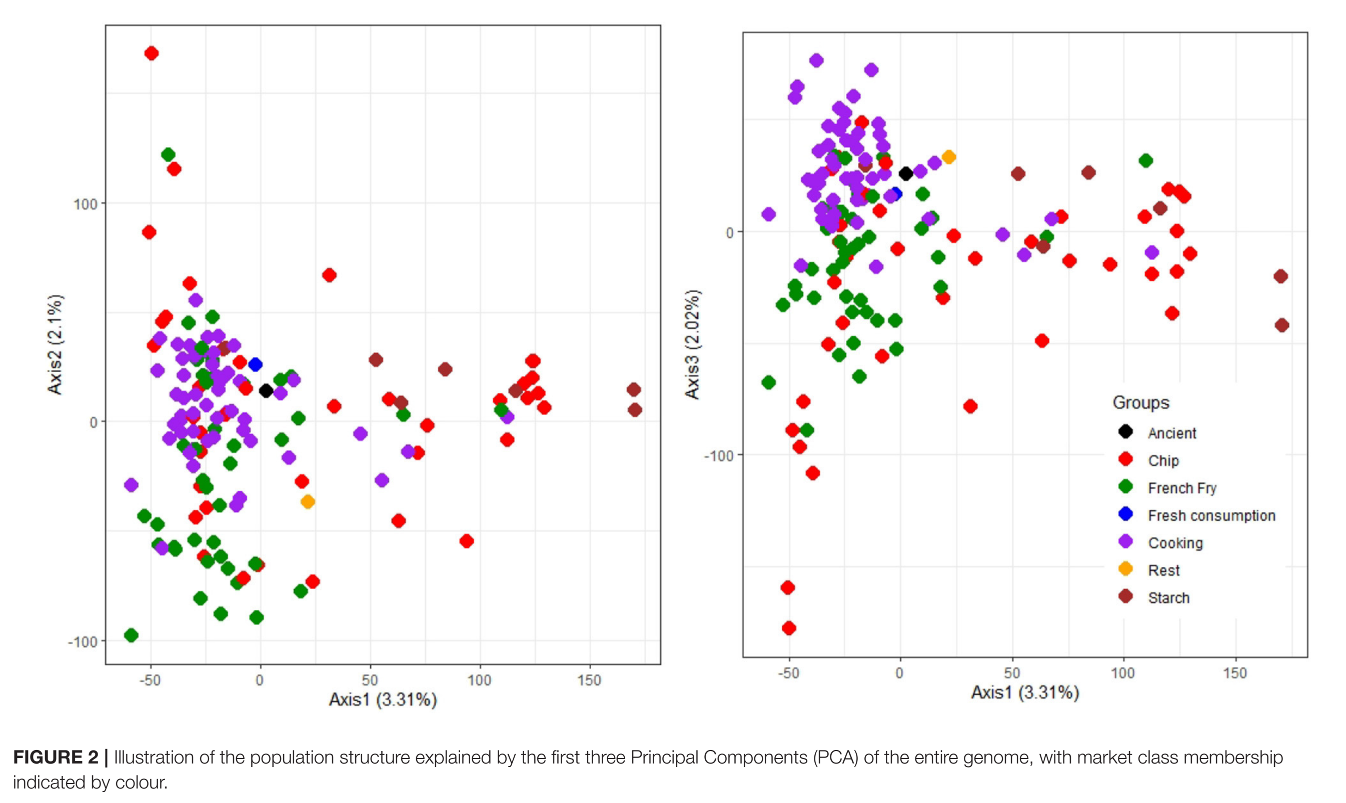


FIGURE 2 | Illustration of the population structure explained by the first three Principal Components (PCA) of the entire genome, with market class membership indicated by colour.

TABLE 4 | F_{ST} statistic between sub-populations.

F_{ST}	Cooking	French fry	Chip
French fry	0.0088		
Chip	0.0116	0.0098	
Starch	0.0323	0.0341	0.0130

Numbers close to zero indicate populations that are more genetically similar.

meet the requirement of minimum population size for these analyses (Willing et al., 2012; Nazareno et al., 2017). Population classifications contributed only 6.7% of the total molecular variation according to the results of AMOVA, further supporting what we see in **Figure 2**. The four major market classes showed very little separation with F_{ST} values close to zero (**Table 4**), indicating that these sub-populations are genetically similar. The starch market class is closer to the chip processing group than the cooking and French fry processing classes as shown in **Table 4**, and illustrated in **Figure 2**.

All population structure analyses were performed using the R packages StaMPP (Pembleton et al., 2013) and Adegenet (Jombart and Ahmed, 2011), because of their suitability for polyploid population genetics (Dufresne et al., 2014).

Phenotypic Analysis

Phenotypes were first adjusted for local trends within each trial as seen in Equation (1). At this level of analysis, outliers were detected and removed and the extracted BLUEs were then pooled across all trials as described in the Materials and Methods section. The resulting distributions and correlations between phenotypic values can be seen in **Figure 3**.

Broad-sense heritability was calculated for tuber weight, tuber count, tuber length, and dry matter resulting in H^2 values of 0.78, 0.79, 0.91, and 0.96, respectively. These heritability estimates are quite high and most likely because of the repeated trials at three locations and 2 years.

Genomic Prediction

The results of GP analyses on the four traits, compared across eight statistical models can be seen in **Figure 4**. Accuracies ranged from 0.32, when tuber count was predicted with a Bayesian LASSO model, to 0.77 when dry matter content was predicted with a Bayes-A model. With the highest heritability, it is not surprising that dry matter has the highest prediction accuracy. Tuber length was predicted more accurately than tuber count, and this corresponds with the ordering of their heritability estimates. The trait with the second highest prediction accuracy was tuber weight, which was unexpected as it had the lowest heritability, and was the only trait that did not agree with the order of heritability estimates.

There is no clear ranking of model performance across all traits, however **Figure 4** allows us to observe some trends. The three Bayesian models, that differ in their assumed distribution of marker effects, show little difference between them across all traits (differences of at least 0.03 will be considered relevant).

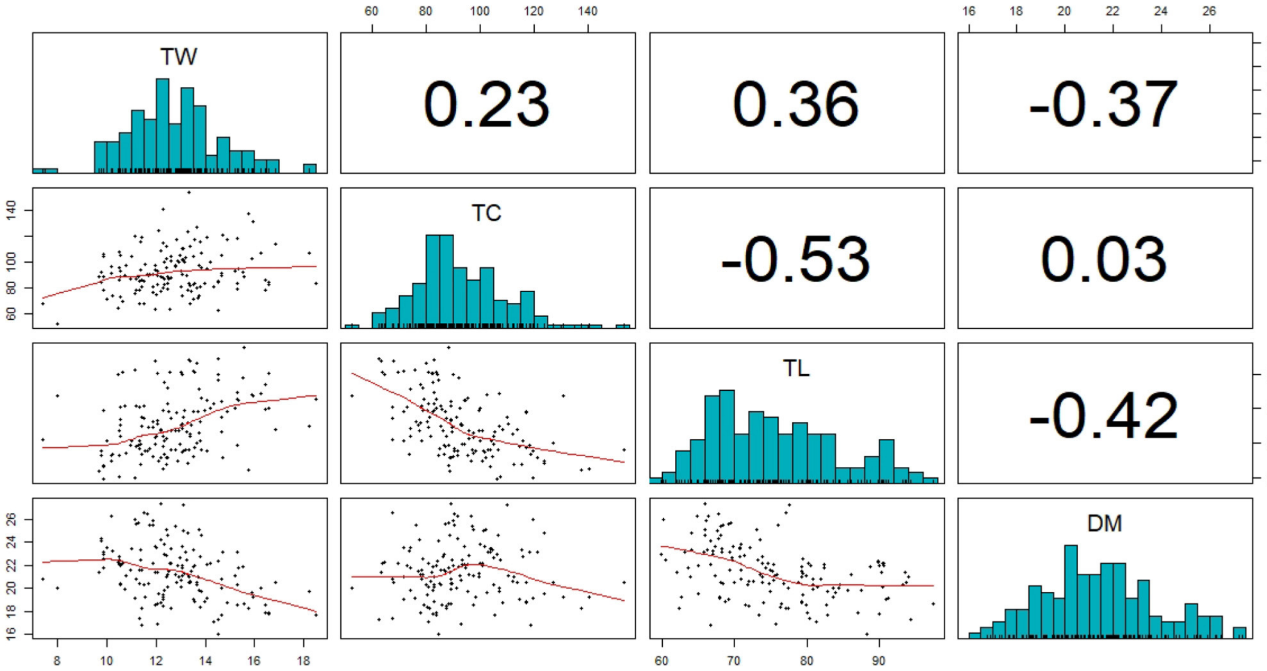


FIGURE 3 | Distribution and correlation between the four analysed traits: Tuber Weight (TW), Tuber Count (TC), Tuber Length (TL), Dry Matter (DM).

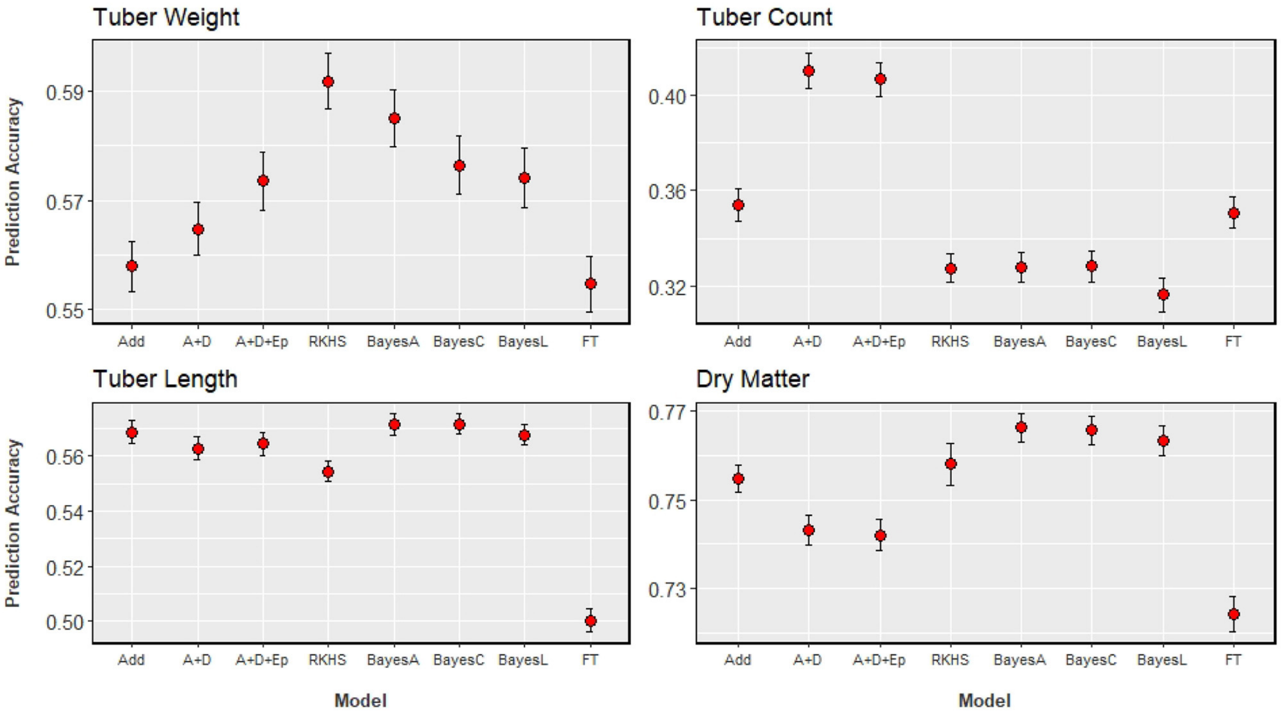


FIGURE 4 | GP results of the four analysed traits, with prediction accuracy on the y-axis, and the x-axis indicating the model used: Add (GBLUP with additive genomic relationship matrix), A+D (GBLUP with additive and dominance relationship matrices), A+D+Ep (GBLUP with additive, dominance, and epistatic relationship matrices), RKHS (Reproducing Kernel-Hilbert Space model), BayesC (Bayes Cπ model), BayesL (Bayesian LASSO model), FT (Full Tetraploid as proposed by Slater et al., 2016). Standard errors of estimates are illustrated with the bars around the points.

Bayesian models were among the better performing models, for those traits that were also predicted well by an additive GBLUP model. Extending an additive model to include dominance or dominance + epistasis did not significantly improve prediction accuracies for the traits analysed, except for tuber count. For dry matter content and tuber length, the addition of these non-additive effects decreased prediction accuracy, but these decreases were not relevant. With tuber count again being the exception, the performance of the RKHS model was comparable to the model that best predicted a given trait. The full tetraploid model (FT) was generally outperformed by all the other models, more so for dry matter content and tuber length.

Model ranking can better be assessed on a per trait basis, as the best performing model depends on the trait analysed. **Figure 4** shows that models that directly estimate marker effects are the most suitable for predicting dry matter. Tuber length can be predicted efficiently with either an additive GBLUP or one of the Bayesian models. Tuber weight prediction appears to benefit from modelling non-additive effects, but the source of that effect is not completely clear. There is a small increase in prediction accuracy as we move from additive, to an additive-dominant model, to a model that includes additive, dominance and additive × additive epistatic effects. This trend could suggest the presence of a non-additive effect not explicitly modelled by the GBLUP models, such as an effect that is of a higher order than the additive × additive epistatic interaction. The RKHS model produced the highest accuracies for this trait, and is unique among the models tested as all other models are parametric while the RKHS model is semi-parametric. The RKHS model captures the same first order epistasis as the parametric model, however it gives the most noticeable improvement in comparison to the standard additive GBLUP model (from 0.56 to 0.59). For predicting tuber count, extending the additive GBLUP model to include dominance effects improved the accuracy of prediction by 17%, from 0.35 to 0.41, which we consider as a relevant change. The explicit modelling of this particular non-additive effect is clearly beneficial for the prediction of this trait, more so than any other trait analysed.

An additional result from the Bayes Cπ model is the fraction of markers selected because of their potential QTL effects. For dry matter 0.27 markers were selected while the for tuber weight, half (0.5) of the markers were selected. The proportion of selected markers for tuber length and count was 0.35 and 0.34, respectively. This gives an idea of trait architecture which is investigated further in the next section.

GWAS

Trait architecture is responsible for the particularity between the accuracy of a GP statistical model and the trait analysed. To uncover some of the underlying genetic behaviour responsible for the expression of our four traits, Equation (8) was applied. Two further models were tested, one that included fixed effects for market class assignments and another that included fixed effects for the first three principal component axes. A look at the QQ-plot showed no significant inflation of *p*-values when these fixed effects were excluded (results not shown), and no difference when they were introduced to the GWAS model. This can be attributed to the lack of population structure as reported previously, thus a model simply with a genomic relationship matrix was enough to avoid spurious associations between markers and traits. The threshold for identifying significant markers was $-log_{10}p = 3.65$, after the 0.05 threshold was adapted for multiple testing ($\frac{0.05}{222}$). The signals detected when coding markers for additive or non-additive effects (in this case two levels of dominance) can be seen for dry matter and tuber count in **Figure 5**. Manhattan plots for tuber length and tuber weight are not shown as these plots were not very informative, however analyses on the significant markers for these traits still follow.

Across the five tested GWAS models for dry matter content, the most significant association with markers is observed when an additive effect is assumed (**Figure 5**). When compared to the plots assuming dominance we see that additivity gives both the highest scores $[-log_{10}(p)]$ and the most abundant markers appearing above the threshold. For tuber count we see significant markers in more abundance when a dominant coding of the marker matrix is considered. There are multiple flanking "hits" on chromosomes 1 and 3 when we assume simplex dominance for the reference allele. Chromosomes 4 (the two significant markers overlap on the plot) and 8 also show neighbouring markers with significant associations to tuber count when dominance is assumed to occur in the presence of a single alternative allele.

Manhattan plots of tuber weight did not show much evidence of significant QTLs in this analysis (**Supplementary Figure 2**). There are a few markers associated when dominant effects are modelled: these occur when the alternative allele is simplex or duplex dominant. Similarly, GWAS for tuber length analysis did not show any clear profile of associated markers (**Supplementary Figure 2**). Still we observed more significant markers when they were coded as duplex dominant for the alternative allele, however the highest scores were observed for additive effects and duplex dominance for the reference allele.

Manhattan plots can be ambiguous, therefore further analysis was done by performing a linear regression to uncover which marker effect type is more important for trait expression. Reported in **Table 5** is the fit statistic for each regression (R^2), which can be interpreted as the amount of phenotypic variance that can be explained by the significant markers and/or population structure.

Of the four traits, the variance of dry matter is best explained from marker information, and also has the biggest influence from principal components alone. This does not agree with previous population structure analysis (**Figure 2**), but those previous analyses were across the entire genome. For this trait, using only the significant additive markers found on chromosome 3, we can explain over 50% of phenotypic variation (not shown). For dry matter, the inclusion of dominance adds no information as seen in the GP results (**Figure 4** and **Table 5**). Tuber count, as seen in GP, is controlled by dominance effects. This dominance effect comes from the alternative allele as opposed to the reference allele which was not clear from **Figure 5**, and more than doubles the explained phenotypic variance (12.50–35.92%) in comparison to additive effects. Explained phenotypic variation of tuber weight remains unchanged under simplex dominance assumptions. We

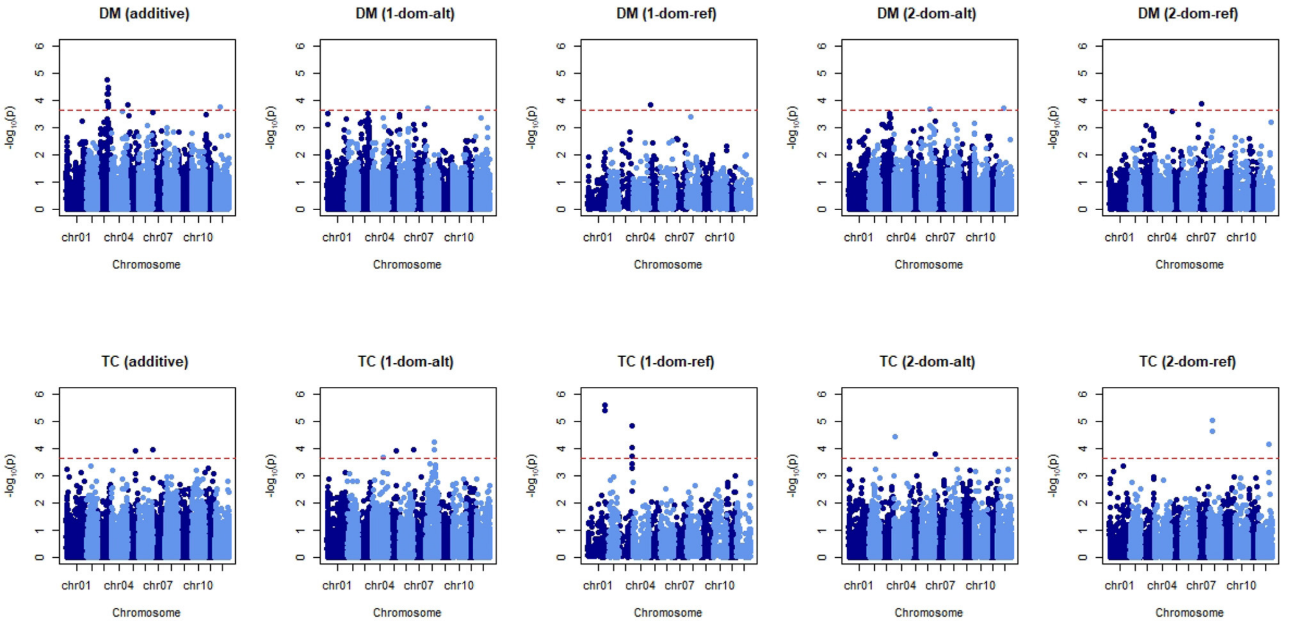


FIGURE 5 | Manhattan plots for dry matter (DM) and tuber count (TC). Five marker matrices were tested: additive, simplex dominant in favour of the alternative allele and reference allele (1-dom-alt and 1-dom-ref, respectively), duplex dominant in favour of the alternative allele and reference allele (2-dom-alt and 2-dom-ref, respectively). The red horizontal line indicates the threshold for significant markers.

TABLE 5 | Percentage of variance explained (R^2) from regression of each trait against: first three principal components only, significant additive markers after correcting for the first three principal components, significant dominance effect markers (under various configurations) after correcting for additive effects and the first three principal components.

Trait	3 PCs only	3 PCs + Add markers	3 PCs + Add + Dom markers			
			1-dom-alt	1-dom-ref	2-dom-alt	2-dom-ref
Tuber weight	5.55	12.92	12.26	12.26	30.40	16.88
Tuber count	0.07	12.50	35.92	14.99	23.43	15.78
Tuber length	6.13	42.43	53.24	44.68	57.75	47.73
Dry matter	41.77	67.01	67.96	66.64	68.49	66.24

see strong evidence that the level of dominance occurs at a duplex level, where the explained phenotypic variation increases from 12.92 to 30.40%, when compared to additive assumptions. A significant portion of the phenotypic variation of tuber length can be explained by additive effects (42%) and we do see an increase when dominance from the alternative allele is modelled (simplex or duplex), but this increase was not as noticeable as the increase shown when tuber count and weight are coded for dominant effects. In general the effect of population structure on explaining the variation of tuber traits (length, weight, and count) is small.

DISCUSSION

The primary focus of this study was to explore and compare statistical models for genomic prediction in tetraploid potato. As a secondary focus, a genome wide association analysis was conducted to identify trait architecture and thus explain the reasons for differences in prediction accuracies from trait to trait. The same marker profile was used for both GP and GWAS. It is worth noting that Bayes-R and Genome-wide Complex Trait Analysis (GCTA) can simultaneously perform GP and GWAS, therefore deserving of further study. However, for this study, we focused on a tetraploid species and therefore wanted to ensure that both the GP and GWAS analyses were tailored for tetraploids.

Translating these findings to a traditional breeding program expose the limitations of this study. Only 147 cultivars were analysed in this study, spread over several market classes. In a traditional breeding program thousands of new hybrids can be evaluated within one particular market class. Despite these limitations, the work done here still shows that there is merit using genomic selection, especially after the first round of phenotypic selection where a majority of the material has been discarded. After this stage genomic selection can then significantly speed up the breeding cycle.

Heritability

Broad-sense Heritability estimates were quite high. As mentioned previously, our heritability estimates can more accurately be defined as repeatability estimates (Falconer et al., 1996), because we are averaging across six trials. Our high estimates show that there is not too much genotype by environment interaction (GxE) and thus high repeatability. The order of heritability estimates was unexpectedly not in agreement with the order of GP accuracies, and this was also found in a similar study (Stich and Van Inghelandt, 2018). The order of our heritability estimates is not a ranking of which traits would best be explained by marker information, although it does give some indication. Instead it is a ranking of which traits show the least to most GxE, with dry matter having the least GxE and tuber weight having the most. Regardless, one would expect heritability to translate to marker effects and GP accuracies not be so low in relation to heritability estimates.

Genomic Prediction Models

For most traits, the differences between GP models were very small (Habyarimana et al., 2017; Sverrisdóttir et al., 2017; Amadeu et al., 2020). Amadeu et al. (2020) concluded that there is little difference in prediction accuracy between modelling strategies, and use of an additive GBLUP model would be sufficient for GP in auto-tetraploids. Only two traits in potato were analysed in that study: yield and specific gravity. Specific gravity is closely related to dry matter (Simmonds, 1977; Kumar et al., 2005), and in this study we also found that the additive GBLUP model is suitable. Tuber length also supports the conclusion by Amadeu et al. (2020), where a GBLUP additive model performs whole genome prediction as well as other models. For the other two traits analysed in this study, tuber count and yield, we have shown that other model considerations should be made to maximise prediction accuracy.

Capturing Dominance and Epistatic Effects

For tuber count, the modelling of dominance gave a 17% increase in prediction accuracy (from 0.35 to 0.41). The trait architecture revealed in the GWAS section, showed that significant dominant markers explained the most phenotypic variation, therefore targetting these non-additive effects resulted in the highest prediction accuracy. Trying to capture these non-additive effects with the full tetraploid model did not increase prediction accuracy.

Yield was one of the traits analysed by Amadeu et al. (2020), however, the RKHS model was not tested in that study. In this study, we show that the parametric models that included a term to capture epistasis did show evidence that there is an epistatic effect controlling tuber weight (which we consider as yield). The semi-parametric RKHS model produced the highest prediction accuracies for this trait. This is in agreement with findings from other studies that concluded epistasis is better captured by semi-parametric (and non-parametric) in comparison to parametric models (Howard et al., 2014; Jacquin et al., 2016; Momen et al., 2018). Based on GWAS results, there may be important dominance effects that were not explicitly modelled in our GP analyses of this trait. The GWAS results showed that, like tuber count, yield had the highest explained phenotypic variance when markers were coded as dominant. However, these markers that explained a significant portion of phenotypic variance for yield were coded as duplex dominant (instead of simplex). The dominance relation matrix used in our GP models assume simplex dominance, and there are no current adaptations to expand to higher levels of dominance (Amadeu et al., 2020).

The explicit modelling of epistasis for specific gravity in Endelman et al. (2018) gave a substantial increase in prediction accuracy, however that study also did not include the RKHS model. It would have been interesting to see if an RKHS model would have performed better, based on what we observed for epistasis in tuber weight and other previous studies as mentioned before. Interestingly, dry matter was not improved with the modelling of epistasis in this study, which contradicts the results of Endelman et al. (2018).

The full tetraploid model (Slater et al., 2016), developed to implicitly capture non-additive effects in auto-tetraploids, did not improve accuracies in this study and one other (Amadeu et al., 2020). In our analyses, this model performed least favourably for most traits. A possible reason for this is the use of genotype frequency instead of allele frequency. The marker data was dominated significantly with nulliplex and simplex dosages ($>$ 75% of information) and therefore genotype frequencies for other dosages may be severely under-represented. GBS data for tetraploid data has been said to bias against the alternate allele (Endelman, personal communication, November 07, 2019), which is most likely the cause of the imbalance of dosage classes for the data in this study. For this reason, it would be worthwhile to have another look at this model in a study with a more balanced marker profile.

Mixed Models (GBLUP) vs. Bayesian Models

Marker effect models are expected to perform better than mixed model GP models (GBLUP) when traits are controlled by a few high impact loci (de los Campos et al., 2013). Like the previous GP studies for potato (Habyarimana et al., 2017; Sverrisdóttir et al., 2017; Amadeu et al., 2020), this study also revealed very little difference between these two model classes. Despite the negligible differences, two traits did give some surprising results.

GWAS for dry matter found a few markers that were able to explain more than 50% of phenotypic variability. Therefore, for this trait, we would have expected a relevant increase going from a mixed model to marker effect model. Tuber yield showed a small but irrelevant increase ($<$ 0.03) when moving from mixed to marker effect modelling, however GWAS findings were unable to explain this result. The significant additive markers for yield explained very little phenotypic variation (12.92%), therefore it was unexpected that a marker effect model would have even slightly outperformed a GBLUP model. Habyarimana et al. (2017) also found that the Bayesian model gave more accurate predictions for yield than the traditional GBLUP model.

CONCLUSIONS

- For GP in auto-tetraploids, there are very little differences between different types of shrinkage methods, and models that do both shrinkage and variable selection.
- GWAS can assist in deciding what model strategies should be considered, especially when considering capturing non-additive effects. When GWAS reveals significant dominant effect markers (simplex), this should be modelled specifically in GP models.
- Tuber weight shows evidence of epistasis, therefore semi- and non-parametric models should be used to predict this trait. Further investigation can include extending the dominance relationship matrix to include duplex dominance and modelling higher levels of epistasis.
- There is no one-size-fits-all model, especially when capturing non-additive effects. Understanding the nature of these effects, example dominance in tuber count vs. epistasis in tuber weight, is important information when choosing the most suitable model.

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 performed the analyses and drafted the manuscript. CZ developed the scripts for calculating genotype probabilities. CM, HM, and RV contributed to the discussion on analytical models and data preparation. AB performed bioinformatic analyses on the sequence data. FE guided analyses and was the general overseer for the project. All authors significantly contributed to the present study and read and approved the final manuscript.

FUNDING

This study was funded by the following sources: Solynta, Meijer Potato, Pepsico, and the Dutch Research Council (NWO).

ACKNOWLEDGMENTS

This work was part of the research programme PredAPloid with project number 14520, which is financed by the NWO. We would like to thank Solynta and Meijer BV for making the data available to us.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.672417/full#supplementary-material

Supplementary Data Sheet 1 | Phenotypic data.

Supplementary Data Sheet 2 | Cross validation scheme.

Supplementary Data Sheet 3 | Genotypic data.

Supplementary Data Sheet 4 | Genotype probability details.

Supplementary Data Sheet 5 | Supplementary figures.

REFERENCES

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A. F., Resende, M. F. R. Jr, and Muñoz, P. R. (2016). AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example. *Plant Genome* 9:plantgenome2016.01.0009. doi: 10.3835/plantgenome2016.01.0009

Amadeu, R. R., Ferrão, L., F. V., Oliveira, I. B., Benevenuto, J., Endelman, J. B., and Munoz P. R. (2020). Impact of dominance effects on autotetraploid genomic prediction. *Crop Sci*, 60. doi: 10.2135/cropsci2019.02.0138

Annicchiarico, P., Nazzicari, N., Li, X., Wei, Y., Pecetti, L., and Brummer, E. C. (2015). Accuracy of genomic selection for alfalfa biomass yield in different reference populations. *BMC Genomics* 16, 1020–1020. doi: 10.1186/s12864-015-2212-y

Ashraf, B. H., Byrne, S., Fé, D., Czaban, A., Asp, T., Pedersen, M. G., et al. (2016). Estimating genomic heritabilities at the level of family-pool samples of perennial ryegrass using genotyping-by-sequencing. *Theor. Appl. Genet*. 129, 45–52. doi: 10.1007/s00122-015-2607-9

Bernardo, R. (1996). Best linear unbiased prediction of maize single-cross performance. *Crop Sci*. 36:cropsci1996.0011183X003600010009x. doi: 10.2135/cropsci1996.0011183X003600010009x

Birch, P. R. J., Bryan, G., Fenton, B., Gilroy, E. M., Hein, I., Jones, J. T., et al. (2012). Crops that feed the world 8: potato: are the trends of increased global production sustainable? *Food Secur*. 4, 477–508. doi: 10.1007/s12571-012-0220-1

Butler, D. (2009). *ASREML: ASREML() Fits the Linear Mixed Model*. Hemel Hampstead: R Package Version 3.0.

Coombes, N. E. (2009). *Digger Design Search Tool in R*. Available online at: http://nswdpibiom.org/austatgen/software/ (accessed July 10, 2020).

de Bem Oliveira, I., Resende Marcio, F. R. J., Ferro, L. F. V., Amadeu, R. R., Endelman, J. B., Kirst, M., et al. (2019). Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3* 9, 1189–1198. doi: 10.1534/g3.119.400059

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

D'hoop, B. B., Keizer, P. L. C., Paulo, M. J., Visser, R. G. F., van Eeuwijk, F. A., and van Eck, H. J. (2014). Identification of agronomically important qtl in tetraploid potato cultivars using a marker-trait association analysis. *Theor. Appl. Genet*. 127, 731–748. doi: 10.1007/s00122-013-2254-y

D'hoop, B. B., Paulo, M. J., Mank, R. A., van Eck, H. J., and van Eeuwijk, F. A. (2008). Association mapping of quality traits in potato (*Solanum tuberosum* l.). *Euphytica* 161, 47–60. doi: 10.1007/s10681-007-9565-5

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol*. 23, 40–69. doi: 10.1111/mec.12581

Enciso-Rodriguez, F., Douches, D., Lopez-Cruz, M., Coombs, J., and de los Campos, G. (2018). Genomic selection for late blight and common scab resistance in tetraploid potato (*Solanum tuberosum*). *G3* 8, 2471–2481. doi: 10.1534/g3.118.200273

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Endelman, J. B., Carley, C. A. S., Bethke, P. C., Coombs, J. J., Clough, M. E., da Silva, W. L., et al. (2018). Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. *Genetics* 209, 77–87. doi: 10.1534/genetics.118.300685

Falconer, D. S. D. S., Mackay, T. F. C., Falconer, D., and Mackay, T. F. (1996). *Introduction to Quantitative Genetics, 4th Edn*. Burnt Mill: Longman.

Gallais, A. (2003). *Quantitative Genetics and Breeding Methods in Autopolyploid Plants*. Paris: Quae.

Garcia, A. A. F., Mollinari, M., Marconi, T. G., Serang, O. R., Silva, R. R., Vieira, M. L. C., et al. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Sci. Rep*. 3:3399. doi: 10.1038/srep03399

Gianola, D., and van Kaam, J. B. C. H. M. (2008). Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303. doi: 10.1534/genetics.107.084285

Guo, X., Cericola, F., Fè, D., Pedersen, M. G., Lenk, I., Jensen, C. S., Jensen, J., and Janss, L. L. (2018). Genomic prediction in tetraploid ryegrass using allele frequencies based on genotyping by sequencing. *Front. Plant Sci*. 9:1165. doi: 10.3389/fpls.2018.01165

Habyarimana, E., Parisi, B., and Mandolino, G. (2017). Genomic prediction for yields, processing and nutritional quality traits in cultivated potato (*Solanum tuberosum* l.). *Plant Breed* 136, 245–252. doi: 10.1111/pbr.12461

Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., Hickey, J. M., Chiurugwi, T., 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

Howard, R., Carriquiry, A. L., and Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3* 4, 1027–1046. doi: 10.1534/g3.114.010298

Inostroza, L., Bhakta, M., Acuña, H., Vásquez, C., Ibáñez, J., Tapia, G., et al. (2018). Understanding the complexity of cold tolerance in white clover using temperature gradient locations and a GWAS approach. *Plant Genome* 11. doi: 10.3835/plantgenome2017.11.0096

Jacquin, L., Cao, T.-V., and Ahmadi, N. (2016). A unified and comprehensible view of parametric and kernel methods for genomic prediction with application to rice. *Front. Genet*. 7:145. doi: 10.3389/fgene.2016.00145

Jansky, S. (2009). "Chapter 2 - breeding, genetics, and cultivar development," in *Advances in Potato Chemistry and Technology*, eds J. Singh and L. Kaur (San Diego, CA: Academic Press), 27–62. doi: 10.1016/B978-0-12-374349-7.00002-7

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

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

Kumar, D., Ezekiel, R., Singh, B., and Ahmed, I. (2005). Conversion table for specific gravity, dry matter and starch content from under water weight of potatoes grown in North-Indian plains. *Potato J*. 32, 79–84.

Li, J., and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95, 221–227. doi: 10.1038/sj.hdy.6800717

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

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*. 8:12309. doi: 10.1038/s41598-018-30089-2

Nazareno, A. G., Bemmels, J. B., Dick, C. W., and Lohmann, L. G. (2017). Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Mol. Ecol. Resour*. 17, 1136–1147. doi: 10.1111/1755-0998.12654

Pembleton, L. W., Cogan, N. O. I., and Forster, J. W. (2013). Stampp: an r package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Mol. Ecol. Resour*. 13, 946–952. doi: 10.1111/1755-0998.12129

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

Piepho, H.-P., Williams, E., and Michel, V. (2015). Beyond latin squares: a brief tour of row-column designs. *Agron. J*. 107, 2263–2270. doi: 10.2134/agronj15.0144

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 (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statstical Computing.

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

Rosyara, U. R., De Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome* 9. doi: 10.3835/plantgenome2015.08.0073

Simmonds, N. W. (1977). Relations between specific gravity, dry matter content and starch content of potatoes. *Potato Res*. 20, 137–140. doi: 10.1007/BF02360274

Slater, A. T., Cogan, N. O. I., Forster, J. W., Hayes, B. J., and Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *Plant Genome* 9. doi: 10.3835/plantgenome2016.02.0021

Stich, B., and Van Inghelandt, D. (2018). Prospects and potential uses of genomic prediction of key performance traits in tetraploid potato. *Front. Plant Sci*. 9:159. doi: 10.3389/fpls.2018.00159

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., and Lund, M. S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS ONE* 7:e45293. doi: 10.1371/journal.pone.0045293

Sverrisdóttir, E., Byrne, S., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., et al. (2017). Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. *Theor. Appl. Genet*. 130, 2091–2108. doi: 10.1007/s00122-017-2944-y

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

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G. F., van Eck, H. J., and van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and snp-array data of tetraploid potato. *Theor. Appl. Genet*. 130, 123–135. doi: 10.1007/s00122-016-2798-8

Wallace, J. G., Zhang, X., Beyene, Y., Semagn, K., Olsen, M., Prasanna, B. M., et al. (2016). Genome-wide association for plant height and flowering time across 15 tropical maize populations under managed drought stress and well-watered conditions in Sub-Saharan Africa. *Crop Sci*. 56, 2365–2378. doi: 10.2135/cropsci2015.10.0632

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

Willing, E.-M., Dreyer, C., and van Oosterhout, C. (2012). Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PLoS ONE* 7:e42649. doi: 10.1371/journal.pone.0042649

Zheng, C., Amadeu, R. R., Munoz, P. R., and Endelman, J. B. (2020). Haplotype reconstruction in connected tetraploid f1 populations. *bioRxiv [Preprint]*. doi: 10.1101/2020.12.18.423519

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 Plant Science

ORIGINAL RESEARCH
published: 24 September 2021
doi: 10.3389/fpls.2021.703419

Reciprocal Recurrent Genomic Selection Is Impacted by Genotype-by-Environment Interactions

Maximilian Rembe[1], Jochen Christoph Reif[1]*, Erhard Ebmeyer[2], Patrick Thorwarth[3], Viktor Korzun[3,4], Johannes Schacht[5], Philipp H. G. Boeven[5], Pierrick Varenne[5], Ebrahim Kazman[6], Norman Philipp[6], Sonja Kollers[3], Nina Pfeiffer[2], C. Friedrich H. Longin[7], Niklas Hartwig[8], Mario Gils[8†] and Yusheng Zhao[1]

[1] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany, [2] KWS LOCHOW GmbH, Bergen, Germany, [3] KWS SAAT SE & Co. KGaA, Einbeck, Germany, [4] Federal State Budgetary Institution of Science Federal Research Center "Kazan Scientific Center of Russian Academy of Sciences", Kazan, Russia, [5] Limagrain Europe, Ferme de l'Etang – BP3–77390, Verneuil-l'Ètang, France, [6] Syngenta Seeds GmbH, Hadmersleben, Germany, [7] State Plant Breeding Institute, University of Hohenheim, Stuttgart, Germany, [8] Nordsaat Saatzucht GmbH, Langenstein, Germany

OPEN ACCESS

Edited by:
Sarah J. Hearne,
International Maize and Wheat
Improvement Center, Mexico

Reviewed by:
Mahmoud Said,
Academy of Sciences of the Czech
Republic (ASCR), Czechia
Andrea Ariani,
BASF, Belgium

***Correspondence:**
Jochen Christoph Reif
reif@ipk-gatersleben.de

†Present address:
Mario Gils,
Strube Research GmbH & Co. KG,
Söllingen, Germany

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: 30 April 2021
Accepted: 02 August 2021
Published: 24 September 2021

Citation:
Rembe M, Reif JC, Ebmeyer E,
Thorwarth P, Korzun V, Schacht J,
Boeven PHG, Varenne P, Kazman E,
Philipp N, Kollers S, Pfeiffer N,
Longin CFH, Hartwig N, Gils M and
Zhao Y (2021) Reciprocal Recurrent
Genomic Selection Is Impacted by
Genotype-by-Environment
Interactions.
Front. Plant Sci. 12:703419.
doi: 10.3389/fpls.2021.703419

Reciprocal recurrent genomic selection is a breeding strategy aimed at improving the hybrid performance of two base populations. It promises to significantly advance hybrid breeding in wheat. Against this backdrop, the main objective of this study was to empirically investigate the potential and limitations of reciprocal recurrent genomic selection. Genome-wide predictive equations were developed using genomic and phenotypic data from a comprehensive population of 1,604 single crosses between 120 female and 15 male wheat lines. Twenty superior female lines were selected for initiation of the reciprocal recurrent genomic selection program. Focusing on the female pool, one cycle was performed with genomic selection steps at the F_2 (60 out of 629 plants) and the F_5 stage (49 out of 382 plants). Selection gain for grain yield was evaluated at six locations. Analyses of the phenotypic data showed pronounced genotype-by-environment interactions with two environments that formed an outgroup compared to the environments used for the genome-wide prediction equations. Removing these two environments for further analysis resulted in a selection gain of 1.0 dt ha^{-1} compared to the hybrids of the original 20 parental lines. This underscores the potential of reciprocal recurrent genomic selection to promote hybrid wheat breeding, but also highlights the need to develop robust genome-wide predictive equations.

Keywords: grain yield, hybrid breeding, long-term selection gain, genotype-times-year interaction, abiotic stress

INTRODUCTION

Since the discovery of the advantages of hybrid breeding through increased performances due to the exploitation of heterosis (Shull, 1908), it has proven to be a successful strategy in allogamous species such as maize (Troyer, 1999), sunflower (Reif et al., 2013), sugar beet (Li et al., 2010), and rye (Geiger and Miedaner, 2015). Besides, hybrids display higher yield stabilities (Mühleisen et al., 2014), especially in marginal environments (Hallauer et al., 1988) and facilitate the stacking of major genes (Longin et al., 2012). These advantages stimulated investments in the implementation of hybrid breeding also in autogamous species, with the main challenge to develop economically competitive varieties that can compete against the line varieties on the market as the autogamous biology makes economic seed production challenging. Therefore, hybrid varieties must outperform significantly line varieties and the yield surplus must compensate for the higher costs in seed production. Recent advances enabled the introduction of hybrid breeding in autogamous species such as barley (Mühleisen et al., 2013), wheat (Melonek et al., 2021), and most successfully rice (Huang et al., 2017) but a major challenge is the selection gain per unit time: Classical hybrid breeding uses heterosis but exploits less additive variance and the breeding schemes are longer compared to line breeding (Longin et al., 2012).

A promising approach to breed high-yielding hybrids is to maximize the exploitation of beneficial heterosis. The concept of reciprocal recurrent selection (RRS) was originally proposed by Comstock et al. (1949) and optimizes the use of general and specific combining ability by selecting genotypes from one population based on the performance of their progeny resulting from crosses with another population. Ideally, this selection strategy results in a reciprocal shift in gene frequencies among the two populations from which female and male genotypes shall derive. Recurrent selection cycles are applied to further manifest this tendency. The success of RRS has been demonstrated in outcrossing species such as maize (Eyherabide and Hallauer, 1991; Tardin et al., 2007; Souza et al., 2010; Kolawole et al., 2018) and sugar beet (Doney and Theurer, 1978; Hecker, 1985). To the authors knowledge, no studies were published that investigate the potentials and limits of RRS in autogamous cereals such as wheat.

A disadvantage of RRS compared to recurrent selection is the elongation of breeding cycles due to the need to produce sufficient progeny based on which genotypes can be rated. In recurrent selection, the implementation of genomic selection has the potential to shorten the length of selection cycles and raise selection gain (Santantonio et al., 2020; Atanda et al., 2021), but empirical studies providing insights into the long-term effect in recurrent genomic selection are still missing. Research in animal breeding has suggested to complement RRS with genomic selection (Kinghorn et al., 2010). In oil palm, simulations have shown that genomic selection could potentially reduce the generation time of an RRS breeding cycle from 20 to 6 years (Cros et al., 2015). Integration of genomic selection into RRS would furthermore allow the combination of RRS and speed breeding approaches as proposed by Watson et al. (2018). Empirical evidence of the superiority of reciprocal recurrent genomic selection (RRGS) breeding programs, however, is still missing.

Many breeding programs are aimed at producing genotypes adapted to so-called mega-environments. Mega-environments are geographic regions that show similar growing conditions limiting the variance of the interaction effects between genotype and environments (Braun et al., 1996). In Germany, breeders generally aim for genotypes that are capable to meet the requirement criteria of the Federal Plant Variety Office (Bundessortenamt, Hannover), to release registered varieties. The Federal Plant Variety Office tests candidate genotypes in its official trials at up to 15 locations representing wheat growing regions in Germany. It is important to note here that Germany is not further subdivided in the Federal Plant Variety Office tests into target mega-environments for wheat breeding.

This study provides the first empirical results on the potential and limits of an RRGS breeding program in wheat targeted for Germany. The objectives were to (1) investigate the utility of genomic selection to identify superior females through genomic estimation of the general combining ability, (2) evaluate the selection gain for grain yield achieved by an RRGS breeding strategy, and (3) examine the impact of genotype-by-environment interaction on the effectiveness of a long-term breeding strategy.

MATERIALS AND METHODS

Design of the Reciprocal Recurrent Genomic Selection Program

We implemented an RRGS program based on genomic and phenotypic data of a large hybrid wheat population (further denoted as HYWHEAT population) presented in detail in previous studies (Longin et al., 2013; Zhao et al., 2013, 2015; Gowda et al., 2014; Liu et al., 2016, 2020a,b; Jiang et al., 2017; Schulthess et al., 2018; Thorwarth et al., 2018, 2019). Briefly, 120 female and 15 male winter wheat lines adapted to Central Europe were crossed using chemical hybridization agents (e.g., Croisor 100; Kempe et al., 2014) applying standard in house protocols. 1,604 single-cross hybrids were produced. The 1,604 hybrids, their 135 parents, and 10 commercial varieties (As de Coeur, Colonia, Genius, Hystar, JB Asano, Julius, Kredo, Tabasco, Tobak, Tuerkis) were evaluated for grain yield in 11 environments, i.e., 5 and 6 locations (Adenstedt, Boehnshausen, Hadmersleben, Harzhof, Hohenheim, and Seligenstadt), in the growing seasons 2011/2012 and 2012/2013, respectively, in Central Europe, resulting in high quality phenotypic data (Supplementary Table 2 in Zhao et al., 2015). The 135 parental lines were genotyped using a 90,000 SNP array based on an Illumina Infinium assay and after quality tests, 17,372 high-quality SNP markers were retained. The phenotypic and the genomic data were combined, and a ridge regression best linear unbiased prediction (RRBLUP) model was trained fitting additive and dominance effects using the package rrBLUP (Endelman, 2011) in the R software environment (R Core Team, 2020). The implementation of the RRBLUP model was described in detail elsewhere (Zhao et al., 2015). Briefly, the model was:

$$Y = 1_n\mu + Z_A a + Z_D d + e, \tag{1}$$

where Y refers to the grain yield data of the 135 parent lines and their 1,604 hybrids, μ was the overall mean, 1_n was an n-dimensional vector of ones, a and Z_A denoted the additive effects and the corresponding design matrix, and d and Z_D denoted the dominance effects and the corresponding design matrix. The estimated a and d effects were used to predict the genotypic values of the hybrid performances when crossed with the 15 male lines.

In the recurrent genomic selection program, we focused on the female pool and selected 20 out of the 120 female lines. The selection was based on the first-year estimates of general combining abilities and further criteria such as for example being carrier of the dwarfing gene *Rht2*. The 20 female lines formed the C_0 cycle and were crossed following a single round robin design (A x B, B x C, C x D, ..., T x A), i.e., every line was used in two crosses resulting in 20 F_1's. The 20 F_1's were grown in the following season and selfed to the F_2 generation in the green house. Seeds were harvested and around 30 F_2 plants were grown for each of the 20 biparental families amounting to a total of 629 F_2 plants. The 629 F_2 plants were genotyped before flowering using the above-mentioned SNP array. The general combining abilities of the 629 F_2 plants when crossed with the 15 original male lines were estimated using the SNP profiles and the above outlined RRBLUP model. The best 3 F_2 plants per family, i.e., 60 F_2 plants in total, were selected and selfed toward the F_5 generation resulting in 2,886 F_5 genotypes. Descendants from each of the 20 initial crosses were represented in this panel with a mean number of genotypes of 144, ranging from 76 to 277. Seeds of the 2,886 F_5 genotypes were grown in single row plots in the season 2016/2017 and a fraction of 382 $F_{5:6}$ families were visually selected based on overall agronomic performance (disease resistance) and considering plant height and flowering time to facilitate hybrid seed production when crossed with three out of the 15 above outlined male lines. The 382 $F_{5:6}$ families were genotyped using the above-mentioned SNP array. The general combining abilities of the 382 $F_{5:6}$ families when crossed with the 15 original male lines were estimated using the SNP profiles and the above outlined RRBLUP model. Based on the estimated general combining ability effects, 50 outstanding $F_{5:6}$ families were selected (denoted as C_1S). All of the 20 biparental F_2 families were represented in this set of families.

As further reference point besides C_0, 60 F_2 plants out of the above outlined 629 F_2 plants of the 20 biparental families were randomly selected. Here, a total of 3 F_2 plants were randomly drawn from each of the 20 biparental families and selfed toward the F_5 generation resulting in 714 F_5 genotypes. Seeds of the 714 F_5 genotypes were multiplied in single row plots in the season 2016/2017. A subfraction of 30 $F_{5:6}$ families were visually selected considering plant height and flowering time to facilitate hybrid seed production when crossed with three out of the above outlined 15 male lines. The subfraction of 30 $F_{5:6}$ families were denoted as C_1R. The 30 genotypes of the C_1R cycle were genotyped using the above-mentioned SNP array. The integrated data set was filtered by excluding markers with more than 5% missing values, resulting in 4,031 unique and polymorphic markers.

Evaluation of the Selection Gain in Field Trials and Phenotypic Data Analyses

The data set comprised 376 genotypes, including 3 male lines previously used to produce the 1,604 original F_1 hybrids, 20 female lines from C_0, 49 female lines (one out of the above mentioned 50 lines were discarded because hybrid seed production failed entirely) from C_1S, 30 female lines from C_1R, 267 F_1 hybrids, and 7 commercial varieties (Julius, Colonia, Tobak, Elixer, RGT Reform, Hystar, and Genius). The hybrids were derived by crossing the 99 female and the 3 male lines using a factorial mating design. For 267 of the potential 297 single-cross hybrids, enough seeds were harvested for intensive field trials.

All 376 genotypes were evaluated in yield plots for grain yield and plant height at 6 locations in the growing season 2018/2019. The locations were Hadmersleben (latitude 51.98 N, longitude 11.30 E), Mintraching (latitude 48.95 N, longitude 12.25 E), Adenstedt (latitude 52.20 N, longitude 10.18 E), Sossmar (latitude 52.2 N, longitude 10.08 E), Wohlde (latitude 52.8 N, longitude 9.98 E), and Boehnshausen (latitude 51.85 N, longitude 10.95) (**Supplementary Table 1**). The same seeding rate of 230 grains per m^2 was used for both parental lines and hybrids. The plot size ranged from 7.2 to 12 m². Harvesting was performed mechanically and adjusted to a moisture concentration of 140 g H_2O kg^{-1}. The field design was an alpha lattice with block size 11 where each environment corresponded to one replication. The yield trials were treated with fertilizers, fungicides, and herbicides according to farmers practice for intensive wheat production.

The quality of the outlier-controlled phenotypic data from each environment was assessed by estimating the genomic repeatability employing the package BGLR (Perez and de los Campos, 2014) in the software environment R (R Core Team, 2020). For this purpose, the following genomic prediction model was used for lines:

$$y = 1_n\mu + g + e, \tag{2}$$

where y was the n-dimensional vector of phenotypic records of each environment, 1_n was an n-dimensional vector of ones, u was a common intercept, g was an n-dimensional vector of additive genotypic values and e was the residual term. It was assumed that u was a fixed parameter, $g \sim N(0, G\sigma_g^2)$ and $e \sim N(0, I_n\sigma_e^2)$, where I_n denoted the $n \times n$ identity matrix and G denoted the $n \times n$ genomic relationship matrix among genotypes as proposed by VanRaden (2008). For each environment, a 5-fold cross-validation scheme was implemented. Therefore, the population of tested lines was randomly divided into five subsets of equal size. One subset was predicted after the model was trained based on the phenotypic and genotypic data from the remaining four subsets. The correlation between the observed and predicted values defined the prediction ability. After performing 100 5-fold cross-validations, genomic repeatability was obtained by the mean of the prediction abilities.

For assessing the quality of the outlier-controlled phenotypic data for the hybrids tested in each environment, genomic repeatability was estimated employing the following model using the package BGLR (Perez and de los Campos, 2014) in the software environment R (R Core Team, 2020):

$$y = 1_n\mu + Z_A a + Z_D d + e, \tag{3}$$

where y was the n-dimensional vector of phenotypic records of each environment, $1_n$was an n-dimensional vector of ones, μ was the common intercept, a and Z_A denoted the additive effects and the corresponding design matrix, and d and Z_D denoted the dominance effects and the corresponding design matrix. The cross validation of hybrids was executed in the same manner as described for lines.

After outlier tests, the following model was used to obtain best linear unbiased estimations (BLUEs) across environments:

$$y_{ijk} = \mu + g_i + r_j + b_k + e_{ijk}, \tag{4}$$

where y_{ijk} referred to the phenotypic performance of the *ith* genotype at the *jth* location in the *kth* block, μ referred to the intercept, g_i referred to the genetic effect of the *ith* genotype, r_j referred to the effect of the *jth* location, b_k referred to the *kth* block in the *jth* location and e_{ijk} denoted the residual. Genotype was treated as fixed and the remaining effects as random. Outlier detection test was performed following the method M4r as described by Bernal-Vasquez et al. (2016), where the standardized residuals were used in combination with the Bonferroni-Holm test to identify an outlier. The detected outliers (3 for grain yield) were removed for further analysis. Moreover, we estimated variance components with the following model:

$$y_{imfnk} = \mu + a + l_n + b_{nk} + p_i + g'_f + g''_m + g_{fm} + (g'l)_{fn} + (g''l)_{mn} + (pl)_{in} + e_{mfink}, \tag{5}$$

where y_{ifmnk} referred to the phenotypic performance of the *ith* genotype at the *nth* location in the *kth* block, l_n referred to the *nth* location, b_{nk} referred to the *kth* block at the *nth* location, p_i referred to the effect of the *ith* parental line, g'_f referred to the general combining ability (GCA) effect of the *fth* female line, g''_m referred of the GCA effect of the *mth* male line, g_{fm} referred to the specific combining ability (SCA) effect of the *fmth* genotype, $(g'l)_{fn}$ referred to the interaction effect between the GCA of the *fth* female and the *nth* environment, $(g''l)_{mn}$ referred to the interaction effect between the GCA of the *mth* male and the *nth* environment, $(pl)_{in}$ referred to the interaction effect of the *ith* parental line and the *nth* environment e_{mfink} referred to the residual. Dummy variables were used to distinguish between checks, lines, and hybrids. Based on the variance components, heritability (h^2) was estimated separately for lines and hybrids as $h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GxE}^2 + \sigma_e^2}{l}}$, where σ_G^2 refers to the genetic variance of lines or hybrids, σ_{GxE}^2 refers to the genotype-by-environment variance σ_e^2 refers to the residual variance, and l denotes the average number of environments in which the genotypes were tested. Linear mixed models have been executed using ASReml version 4.0 (Butler et al., 2017) in the software environment R (R Core Team, 2020).

GCA$_{\text{Female}}$-by-environment interaction effects were estimated by using the same model as in Equation (5) to further characterize the environments in which the genotypes were evaluated. The GCA$_{\text{Female}}$-by-environment interaction effects were estimated for the experiments of the growing season 2018/2019 only and furthermore in a combined data set consisting of the training environments of the growing seasons 2011/2012 and 2012/2013 and the test environments of the growing season 2018/2019. The GCA$_{\text{Female}}$-by-environment interaction effects were used to perform principal component analyses (PCA) and obtain Euclidean distances based on which the environments were clustered in a complete-linkage approach.

The observed response to selection was estimated as $R_{obs} = \hat{S}$, where $\hat{S} = \mu_{sel} - \mu_{pop}$ denoted the observed selection differential, with μ_{sel} being the phenotypic mean of the selected genotypes and μ_{pop} being the mean of the population from which the selected genotypes were drawn. The C_1 hybrids of the underlying RRGS breeding program have been produced using female lines deriving from a population of 629 genotypes. The capacity for all of the 629 genotypes to produce hybrids has not been estimated in field experiments but only through genomic prediction. For this reason, the mean performance of the C_0 hybrids evaluated in the growing season 2018/2019 has been considered as an approximation for μ_{pop}.

The expected response to selection was estimated as $R_{\exp} = i \bullet h \bullet \sigma_A$, where i denoted the intensity of selection, h refers to the square root of the heritability, and σ_A denoted the standard deviation of the breeding values. Selection intensity was calculated as $i\,(N, G) = i\,(\alpha) - \frac{G-N}{2N(G+1)i(\alpha)}$, where N was the number of selected genotypes, G was the size of the population from which the selected genotypes were drawn, and $i\,(\alpha) = i\left(\frac{N}{G}\right)$ referred to the standardized selection differential according to tabulated values (e.g., Becker, 1975).

Selection was performed in two steps. In the first step, 60 F_2 plants were selected out of a population of 629, resulting in a selection intensity of $i\,(N, G) = i\,(60,\ 629) = 1.78$. Since the selection was based on genomic predictions of the GCA effects of the female lines evaluated in the HYWHEAT experiments, the relevant variance of breeding values corresponds to σ_{GCA}^2, estimated in the experiments of the growing seasons 2011/2012 and 2012/2013 (Zhao et al., 2015). The selection was performed in a population of F_2 plants derived from crosses of genotypes from the aforementioned population. Specifically, three F_2 plants were selected from each family. From quantitative genetic theory, it can be inferred that half of the genetic variance can be exploited if a selection is performed within an F_2 family (Hallauer et al., 2010). It follows that for the first step of selection, $\sigma_{GCA\_F2} = \sqrt{\frac{1}{2}\sigma_{GCA}^2} = 1.2$. The square root of the heritability, h, was assessed using as a conservative estimate the prediction abilities obtained in a chessboard-like cross-validation considering two out of the three different test sets T_2, T_1, and T_0: T_2 test sets included hybrids sharing both parental lines, T_1 test sets comprised hybrids sharing one parental line, and T_0 test sets contained hybrids having no parental line in common with the hybrids in the related training sets. In the RRGS program, male testers were not changed and thus, the C_1 lines reflected a mix between the T_1 and T_2 scenario with a prediction ability of 0.55 and 0.76, respectively. For simplicity, the mean of the prediction abilities for scenarios T_1 and T_2 was considered, resulting in $h = 0.66$.

In the second step of selection, 50 plants were selected from a population of 382 $F_{5:6}$ plants. While h is considered equal to the first step, $i\,(N, G)$ and σ_{GCA} changed, with $i\,(50, 382) = 1.63$. According to quantitative genetic theory (Hallauer et al., 2010), the σ_{GCA} exploited in the second step amounted to $\sigma_{CGA_{F5:6}} = \sqrt{\frac{7}{8}\sigma^{2}_{GCA\_F2}} = 1.1$. The total response to selection was the sum of the responses of the first and second step.

Characterization of Field Locations

In the recent decades, Germany has become more prone to drought events with harmful effects to agro-ecosystems. Personal communication with responsible field technicians indicated adverse field conditions in some of the environments in which the genotypes of the RRGS program were tested. Therefore, GCA_{Female}-by-environment interaction effects were obtained from model (5) to estimate Euclidean distances between each pair of environments.

To further investigate the range in which the environments differed regarding physical stress, we used data from meteorological and satellite-based approaches estimating the plant available water and the condition of the regional vegetation, respectively. The German drought monitor provides data on plant available water beginning from 2015. Information for the plant available water at each location was extracted from the German drought monitor for the growing season 2018/2019 (Zink et al., 2016). In addition, the Vegetation Condition Index (VCI) was employed to quantify the severity of drought stress around the test locations. Geospatial data sets based on the MOD13Q1 images were accessed from the Application for Extracting and Exploring Analysis Ready Samples (https://lpdaacsvc.cr.usgs.gov/appeears/) by USGS. Data from MOD13Q1 images were available for the growing seasons 2011/2012, 2012/2013, and 2018/2019, qualifying them for the comparison of the HYWHEAT and RRGS environments. For each location, an area of 500 ha centered for the coordinates of the test site was selected. The VCI based on the Enhanced Vegetation Index (EVI) was obtained from the equation:

$$VCI_i = \frac{EVI_i - EVI_{\min}}{EVI_{\max} - EVI_{\min}}, \tag{6}$$

where VCI_i referred to the VCI on day i, EVI_i referred to the EVI on day i, $EVI_{\min}$ referred to the minimum EVI in the area observed in the period 2010–2019, and $EVI_{\max}$ referred to the maximum EVI in the area observed in the period 2010–2019. The recommended practice for drought monitoring using the VCI was applied as suggested by the United Nations Office for Outer Space Affairs (2021). The mean value of the selected area around the test site was applied in further considerations.

Based on the data for PAW and VCI, matrices with the individual weather profile of each environment were constructed. From these matrices, principal component analyses were performed, and complete-linkage clusters based on the Euclidean distances were obtained to identify environments with special conditions.

RESULTS

Analysis of Population Structure Revealed Genomic Traces of Selection

The population structure of the 3 male tester lines, the 20 founder female lines (C_0) of the RRGS program, their 30 resulting randomly drawn (C_1R) recombined, and 49 selected progenies (C_1S) was analyzed based on 4,031 polymorphic SNP markers. The principal component analysis derived from the eigenvectors of the parental lines revealed that male and female lines tended to be separated by the first principal component (**Figure 1**). With respect to the second principal component, C_1R was more widely spread than C_1S. Overall, C_1S appeared to be more separated from the male parents than C_1R.

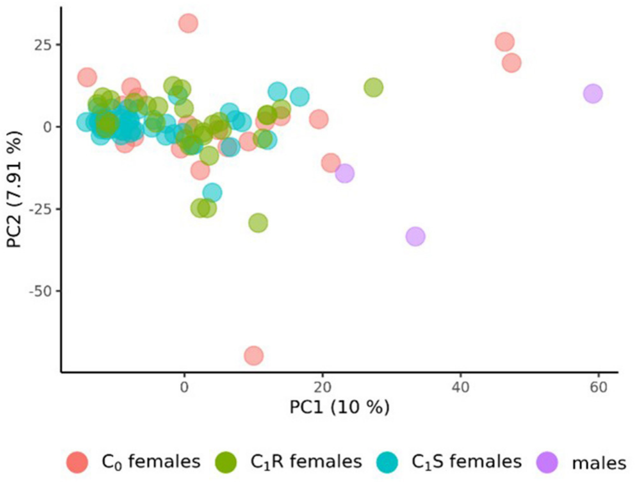


FIGURE 1 | Principal Component Analysis (PCA) of the 20 founder wheat lines (C_0 females), the 3 male lines, the 30 female lines drawn from random after recombining the 20 founder lines (C_1R), and the 49 female lines from the first selection cycles (C_1S). PCA were derived from the eigenvectors of the 3 male and 20 female founder lines. The proportion of variance displayed by the principal components (PC) were presented in brackets.

Phenotypic Data Indicated Pronounced Interactions Between Genotypes and Environments

Genomic repeatabilities were moderate to high, ranging from 0.13 in Wohlde to 0.51 in Hadmersleben with an average of 0.34 in lines and ranging from 0.17 in Mintraching to 0.58 in Adenstedt with an average of 0.34 in hybrids (**Supplementary Table 1**). This underlines the overall high quality of the yield trials. Interestingly, we observed that correlations between grain yields in each environment were low for some pairs (**Table 1**). For example, grain yields of lines and hybrids studied at Wohlde and Hadmersleben were not significantly correlated ($r = 0.09$; $P > 0.36$ for lines; and $r = -0.08$; $P > 0.20$ for hybrids). The grain yield trial conducted at Hadmersleben was not an outlier but correlated significantly with the grain yield trial conducted at Boehnshausen ($r = 0.51$; $P < 0.001$ for the lines; and $r = 0.23$; $P < 0.001$ for the hybrids), a second location in Saxony-Anhalt. These pronounced differences among locations were also visible in the contribution of genotype-by-environment interaction effects (G×E) to the phenotypic variance (**Table 2**). Genotypic variances σ_G^2 were significantly greater than zero ($P < 0.01$, **Table 2**) for lines as well as hybrids, with σ_G^2 being 5.85-times smaller in hybrids than in lines. The ratio of $\sigma_{GxE}^2/\sigma_G^2$ amounted to 0.81 in lines and the ratio of $\sigma_{GCA(Female)xE}^2/\sigma_{GCA\ (Female)}^2$ to 1.13 for general combining

TABLE 1 | Pearson moment correlations between grain yield of 109 wheat lines (below diagonal) and 264 hybrids (above diagonal) evaluated at six locations in the year 2019 to assess the selection gain of the reciprocal recurrent genomic selection program.

Inbred/hybrid	Adenstedt	Boehnshausen	Hadmersleben	Mintraching	Sossmar	Wohlde
Adenstedt	1.00	−0.01	0.05	0.10	−0.02	0.17**
Boehnshausen	0.42***	1.00	0.23***	0.15*	0.12*	−0.07
Hadmersleben	0.22*	0.51***	1.00	0.13*	0.14*	−0.08
Mintraching	0.29**	0.22*	0.26**	1.00	0.14*	−0.01
Sossmar	0.54***	0.55***	0.39***	0.17”	1.00	−0.01
Wohlde	0.44***	0.12	0.09	0.32***	0.24*	1.00

*”, *, **, and *** significantly different from zero at the 0.05, 0.01, 0.001, and 0.0001 level of probability.*

TABLE 2 | Estimates of variance components (residual variance indicated as σ_e) and heritability (*h*2) for winter wheat for grain yield (dt/ha).

Source	Grain yield (dt/ha) 6 locations	Grain yield (dt/ha) 4 locations
Lines		
σ^2_{LINES}	17.21***	17.91***
$\sigma^2_{LINESxE}$	14.01***	10.05***
h^2(Lines)	0.84	0.76
F_1 hybrids		
σ^2_{SCA}	1.07**	1.05
σ^2_{SCAxE}	6.86**	7.50
$\sigma^2_{GCA(Female)}$	1.73**	2.14*
$\sigma^2_{GCAxE(Female)}$	1.97***	2.20*
$\sigma^2_{GCA(Male)}$	0.00	0.00
$\sigma^2_{GCAxE(Male)}$	1.57^{NS}	1.95^{NS}
$\sigma^2_{HYBRIDS}$	2.94	3.20
$\sigma^2_{HYBRIDSxE}$	10.40	11.65
σ^2_e	5.73***	5.77***
h^2(hybrids)	0.54	0.44

Parents and checks were grouped together as lines. The panel was evaluated at 6 locations and comprised 109 lines (7 checks, 99 females and 3 males) and 264 hybrids. In a further analysis, only 4 locations with no stressful growing conditions were investigated. NS, Not significant.

, **, and * significantly different from zero at the 0.01, 0.001, and 0.0001 level of probability.*

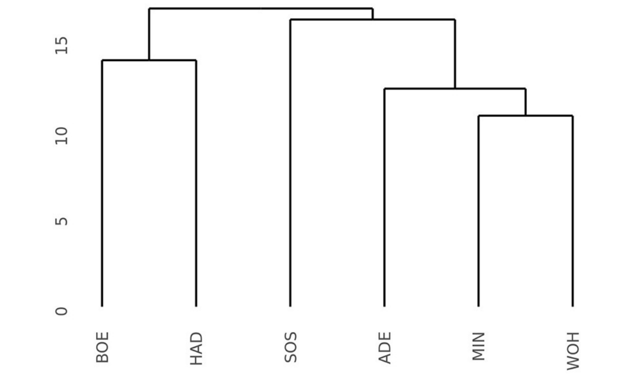


FIGURE 2 | Dendrogram based on the Euclidean distances among six locations estimated using the GCA_{Female}-by-environment interaction effects from the grain yield trials performed in the year 2019 to assess the selection gain of the reciprocal recurrent genomic selection program. The locations were ADE, Adenstedt; BOE, Boehnshausen; HAD, Hadmersleben; MIN, Mintraching; SOS, Sossmar; WOH, Wohlde.

ability effects of the females, which was of special interest during the selection. This underlines the substantial contribution of genotype-by-environment-interaction effects to the phenotypic variance. The estimated heritability (h^2) was high for lines (0.84) and moderate (0.54) for hybrids.

Drought Stress Was Associated With the Pattern of Genotype-by-Environment Interactions

The pronounced differences among locations encouraged us to investigate the pattern of interaction effects between genotypes and environments in more detail. Due to the exploitation of additive effects in the recurrent genomic selection program, we focused on the interaction effects between the GCA effects of females with environments and performed a cluster analysis. The analysis revealed that the Boehnshausen and Hadmersleben locations formed a distinct group, separate from the other locations of the RRGS experiment (**Figure 2**). We assessed the clustering of the locations in more detail by analyzing two published meteorological and satellite-based parameters: the plant available water in the soil (PAW) and vegetation condition index (VCI). Boehnshausen and Hadmersleben were the locations with the lowest PAW during the early growing season (**Figure 3A**) and both locations also clearly clustered separately from the remaining locations when applying a principal component analyses based on the PAW of the entire growing season (**Figure 3B**). A similar picture was observed for the VCI profiles. Boehnshausen and Hadmersleben showed low VCI values throughout the growing season and distinguished from the other locations in particular during the autumn and winter months of the growing season (**Figure 3C**). The principal component analyses based on the VCI profiles of the entire growing season separated the Boehnshausen and Hadmersleben locations from the remaining ones (**Figure 3D**). Thus, the pronounced genotype-by-environment interactions were most

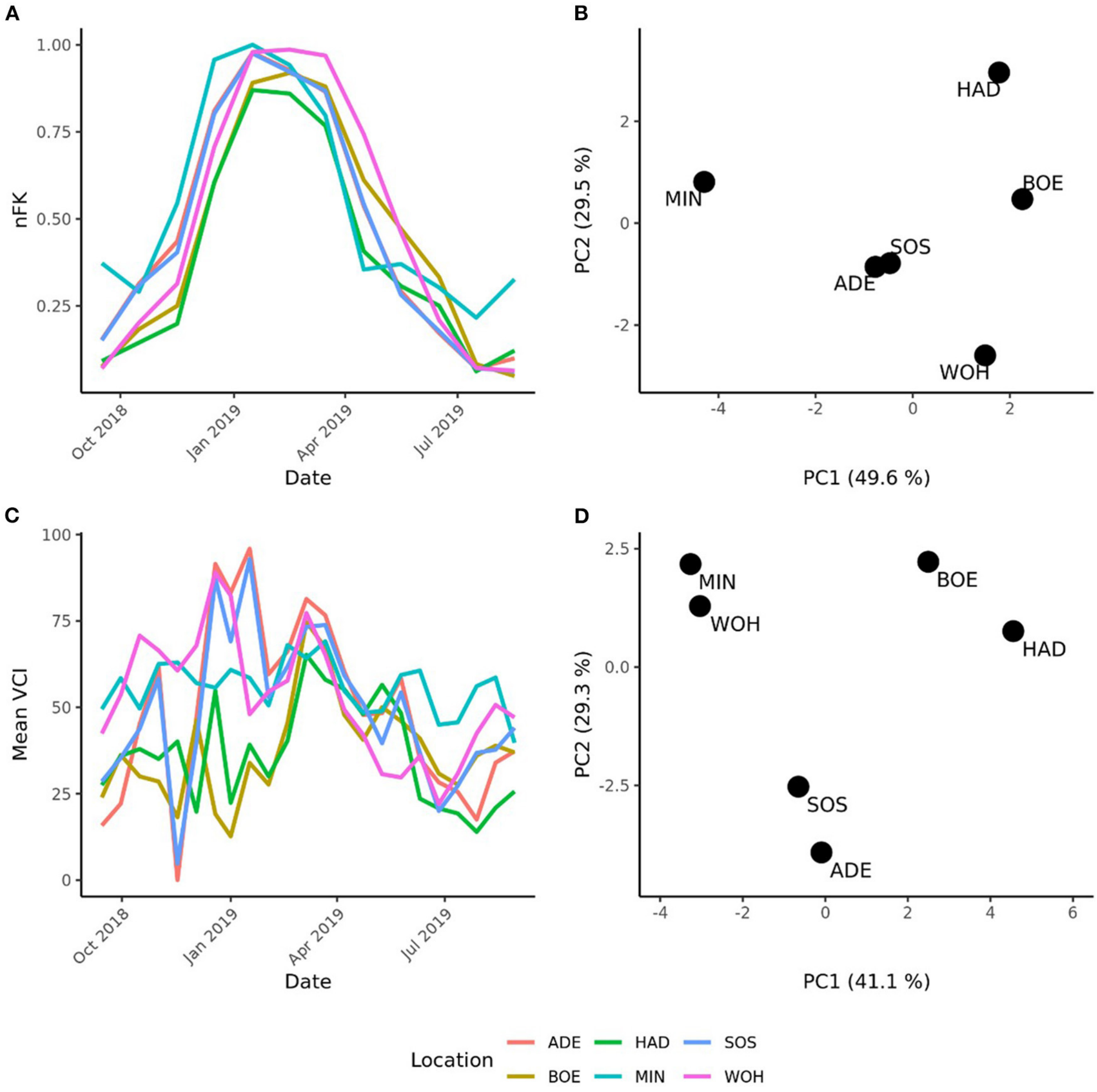


FIGURE 3 | Characterization of the locations used to assess the selection gain of the reciprocal recurrent genomic selection program. **(A)** Line plot of the plant available water (PAW) in the soil and **(B)** a principal component analyses (PCA) based on the PAW profiles of the locations recorded in the growing season [September 1st in the year of sowing (2018) to September 1st in the year of harvest (2019)]. **(C)** Line plot of the mean vegetation condition index (VCI), and PCA based on the mean VCI profiles of the locations recorded in the growing season **(D)**. The locations were indicated as ADE, Adenstedt; BOE, Boehnshausen; HAD, Hadmersleben; MIN, Mintraching; SOS, Sossmar; WOH, Wohlde.

likely caused by severe drought stress occurring in the region of Saxony-Anhalt in the growing season 2018/2019.

Pattern of Genotype-by-Environment Interactions for Integrated Phenotypic Data of the Training and the RRGS Populations

The HYWHEAT training population was phenotyped at five locations in the 2011/2012 season and at six locations in the season 2012/2013, and the RRGS program was evaluated at six locations in the 2018/2019 season. Three overlapping locations albeit in different years were used for both, the HYWHEAT and for the RRGS trials. Interestingly, for the overlapping genotypes (27 for lines and 48 for hybrids) between the HYWHEAT and the RRGS experiments, we observed a much higher correlation between grain yield estimated in the growing seasons 2011/2012 and 2012/2013 within the HYWHEAT experiment ($r = 0.49$; $P < 0.00$ for lines and $r = 0.43$; $P < 0.00$ for hybrids) than between the RRGS experiment and the HYWHEAT experiment in 2011/2012 ($r = -0.04$; $P < 0.80$, for lines and $r = 0.08$; $P < 0.80$, for hybrids) and in 2012/2013 ($r = 0.05$; $P < 0.40$ for lines and $r = -0.17$; $P < 0.80$, for hybrids). A closer look at the correlations between grain yield of the RRGS experiment in each environment

TABLE 3 | Correlations of phenotypic data from single environments of the RRGS experiments (2018–2019) with phenotypic data from HYWHEAT experiments and with single years of the HYWHEAT experiment.

	RRGS: 2018–2019	Hywheat: 2012	Hywheat: 2013	Hywheat: total
Lines	Adenstedt	0.24	0.30	0.40*
	Boehnshausen	0.04	−0.14	−0.11
	Hadmersleben	0.03	−0.30	−0.24
	Mintraching	0.38	−0.07	0.11
	Sossmar	0.11	−0.11	0.03
	Wohlde	0.43*	0.41*	0.54**
Hybrids	Adenstedt	0.37*	0.37**	0.47***
	Boehnshausen	-0.26”	-0.27”	−0.32*
	Hadmersleben	−0.20	−0.23	−0.32*
	Mintraching	−0.20	−0.04	−0.13
	Sossmar	−0.09	−0.07	−0.07
	Wohlde	0.13	0.27”	0.24

A number of 27 overlapping lines and 48 overlapping hybrids were included into the estimation.
*”, *, **, and *** significantly different from zero at the 0.05, 0.01, 0.001, and 0.0001 level of probability.*

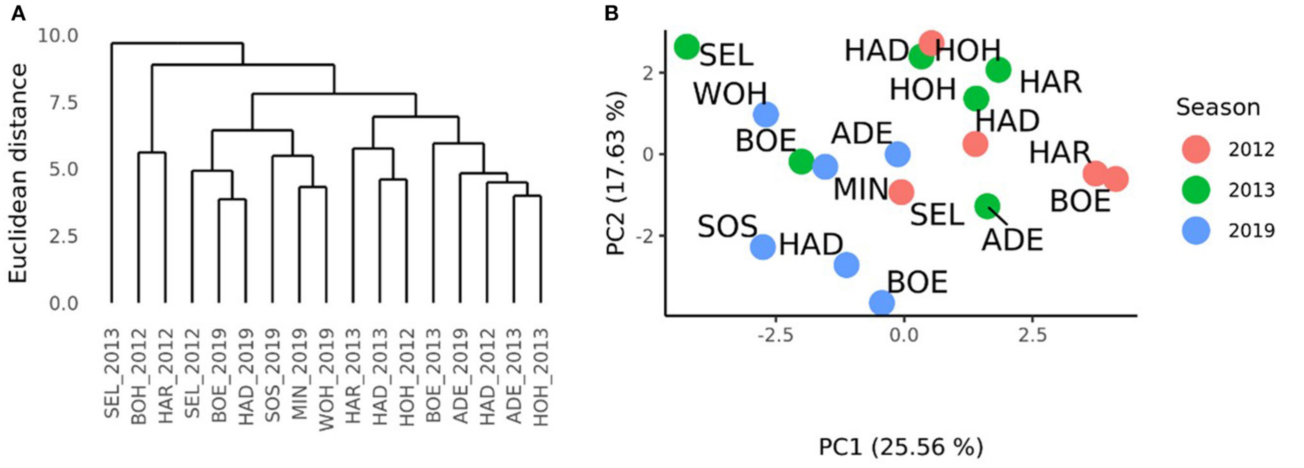


FIGURE 4 | Characterization of the environments of the HYWHEAT and RRGS experiments of the growing seasons 2011/2012, 2012/2013, and 2018/2019, based on the phenotypic performances of overlapping tested hybrids. **(A)** Dendrogram based on the Euclidean distances among 17 location times year combinations (location_year) estimated using the GCA$_{\text{Female}}$-by-environment interaction effects from the grain yield trials performed in the year 2012 and 2013 for the training population (HYWHEAT) and in the year 2019 to assess the selection gain of the reciprocal recurrent genomic selection program. **(B)** PCA based on the GCA$_{\text{Female}}$-by-environment interaction effects of 16 location times year combinations. The locations were indicated as ADE, Adenstedt; BOE, Boehnshausen; HAD, Hadmersleben; HA*R,* Harzhof; HOH, Hohenheim; MIN, Mintraching; SEL, Seligenstadt; SOS, Sossmar; WOH, Wohlde.

and the HYWHEAT experiments revealed strong interaction effects with years (**Table 3**). The RRGS experiment conducted in Wohlde and Adenstedt showed the highest correlations with the HYWHEAT experiments with a decreasing trend toward Mintraching, Sossmar, Boehnshausen, and Hadmersleben.

A complete-linkage clustering based on the Euclidean distances estimated using the GCA$_{\text{Female}}$-by-environment interaction effects was performed to further investigate the relationships among the environments of the HYWHEAT and the RRGS experiments (**Figure 4A**). The location Seligenstadt in 2013, and Boehnshausen in 2012 and Harzhof in 2012 formed outgroups. Apart from Seligenstadt in 2012, which grouped together with the environments Seligenstadt, Boehnshausen, Hadmersleben, Sossmar, Mintraching, and Wohlde from the RRGS experiment, the remaining HYWHEAT environments constituted a distinguished cluster including the environment of Adenstedt in 2019. A PCA based on the GCA$_{\text{Female}}$-by-environment interaction effects showed that apart from Seligenstadt in 2013, the environments of the HYWHEAT experiment grouped together with the RRGS environments Adenstedt, Mintraching and Wohlde in 2019 (**Figure 4B**). The RRGS environments Boehnshausen, Hadmersleben and Sossmar grouped separately from the remaining environments of the RRGS and the HYWHEAT experiments.

A distance matrix obtained from the VCI profiles of the 17 environments of the RRGS and the HYWHEAT experiments was calculated. The comparison to the distance matrix derived from the $\mathrm{GCA}_{\mathrm{Female}}$-by-environment interaction effects revealed a correlation of 0.17 which was significantly different from zero ($P < 0.01$) according to a Mantel test (Mantel, 1967). The cluster which was derived from the VCI profiles of the 17 environments indicated the presence of two subgroups among the HYWHEAT and RRGS experiments (**Figure 5A**). The environments of the RRGS experiment grouped apart from the HYWHEAT experiments, with the environment of Mintraching in 2019 behaving exceptionally as it was situated within the HYWHEAT experiments. Within the HYWHEAT experiments, the location Adenstedt of the growing season 2011/2012 appeared as outgroup. The remaining HYWHEAT environments formed two subgroups distinguished mostly by the year of the evaluation. A PCA was executed based on the VCI profiles of all environments in which the genotypes were tested during the HYWHEAT and RRGS experiments (**Figure 5B**). This analysis exposed shifts of the growing conditions across the growing seasons in which the genotypes were evaluated. Based on the 1st principal component, the environments in the RRGS experiment showed to be largely separated from all remaining environments from the HYWHEAT experiments. Only Mintraching situated closely to some of the HYWHEAT experiments. The 2nd principal component separated the RRGS experiments into three groups: Mintraching and Seligenstadt, Sossmar and Adenstedt, and Boehnshausen and Hadmersleben. The first principal component explained 32.71% of the variance, the second principal component explained 16.78% of the variance.

Selection of Test Locations Affected the Assessment of Breeding Success

Evaluation of effectiveness of RRGS was conducted at six locations during the 2018/2019 growing season, between which pronounced genotype-by-environment interaction effects were observed. Moreover, the 2018/2019 growing season locations showed high genotype-by-year interactions compared to the HYWHEAT experiments conducted in the 2011/2012 and 2012/2013 growing seasons, based on which the genomic selection model was trained. In particular, the Boehnshausen and Hadmersleben locations of the 2018/2019 growing season showed low correlations to the environments of the HYWHEAT experiment (**Table 3**). By comparing the BLUEs for the overlapping genotypes of the RRGS experiment with the BLUEs from the HYWHEAT experiment, correlations of 0.13 and −0.10 were observed for lines and hybrids, respectively. After excluding the locations Boehnshausen and Hadmersleben from the RRGS experiment, correlations between the RRGS experiment and the HYWHEAT experiment based on overlapping genotypes increased to 0.37 for lines and 0.21 for hybrids. Furthermore, exclusion of the Boehnshausen and Hadmersleben locations resulted in a drop of $\sigma^2_{GxE}/\sigma^2_G$ from 1.13 to 1.02 for the GCA of the female lines, indicating a lower proportion of genotype-by-environment interactions among the remaining locations of the RRGS experiment (**Table 2**). These findings encouraged us to investigate the influence of genotype-by-environment interactions on the selection gain of the RRGS breeding programs. To this end, we estimated the selection gain based on phenotypic data collected in all six environments of the RRGS experiment and alternatively we excluded two environments with negative average correlations to the single environments of the HYWHEAT data set and estimated the selection gain based on the remaining four locations.

Including all six environments from the growing season 2018/2019, the randomly drawn female lines of the C_1 cycle showed comparable ($P > 0.1$) average yields as the female parent lines of the C_0 cycle (**Figure 6A**). The genomically selected females showed no significant differences of 1.0 dt ha^{-1} ($P > 0.1$) average yields compared to the randomly selected female lines. Surprisingly, genomically selected female lines of the C_1 cycle showed lower ($P > 0.1$) average yields than the female lines of the C_0 cycle. Both differed by 1.15 dt ha^{-1}. The average yield of the C_0-hybrids, the genomic-selected fraction of the C_1-hybrids (C_1S) and the randomly drawn fraction of the C_1-hybrids (C_1R) did not show any significant ($P > 0.1$) difference. The midparent heterosis was not significantly ($P > 0.1$) larger for C_1S (10.3%) as compared to C_1R (9.7%) and C_0-hybrids (9.8%) (**Figure 7A**). The same was observed for better parent heterosis (**Figure 7C**).

Excluding the two outlier locations from the growing season 2018/2019, randomly drawn female lines of the C_1 cycle showed comparable ($P > 0.1$) average yields as the female parent lines of the C_0 cycle (**Figure 6B**). Genomically selected female lines of the C_1 cycle and randomly selected female lines of the C_1 cycle showed no significantly different ($P > 0.1$) grain yield performance. The female parent lines of the C_1 cycle performed comparable ($P > 0.1$) to the female parent lines of the C_0 cycle. While C_1R hybrids showed no significant difference ($P > 0.1$) in average yield performance compared to C_0 hybrids, C_1S hybrids outperformed ($P < 0.05$) C_0 hybrids by 1.0 dt ha^{-1}, achieving a selection gain of 1%. Moreover, C_1S hybrids outperformed ($P < 0.1$) C_1R hybrids by 0.7 dt ha^{-1}. Midparent heterosis was not significantly different ($P > 0.1$) in C_1R (11.5%) compared to C_0 (11.3%), while C_1S (12.8%) showed a clear advancement and performed significantly better than C_0 ($P < 0.05$) and C_1R ($P < 0.05$) (**Figure 7B**). A different pattern was observed for better parent heterosis. C_0 (11.3%) and C_1R performed comparable ($P > 0.1$). C_1S (10.0%) did not perform significantly different from C_0 ($P > 0.1$) and C_1R ($P > 0.1$) (**Figure 7D**).

The observed selection differential and hence the observed response to selection varied depending on which environments were considered for the evaluation. When all six environments were included, it amounted to $R_{obs\_6E} = -0.4\ dt\ ha^{-1}$. When environments with severe stress conditions were excluded and only four environments were considered, the observed selection differential and hence observed response to selection was $R_{obs\_4E} = 1.0\ dt\ ha^{-1}$.

DISCUSSION

We conducted one cycle of an RRGS program in wheat, including field evaluation of the resulting hybrids, which took a total of 6 years from the first crosses. It is important to note that each subsequent selection cycle lasts only one additional year at most,

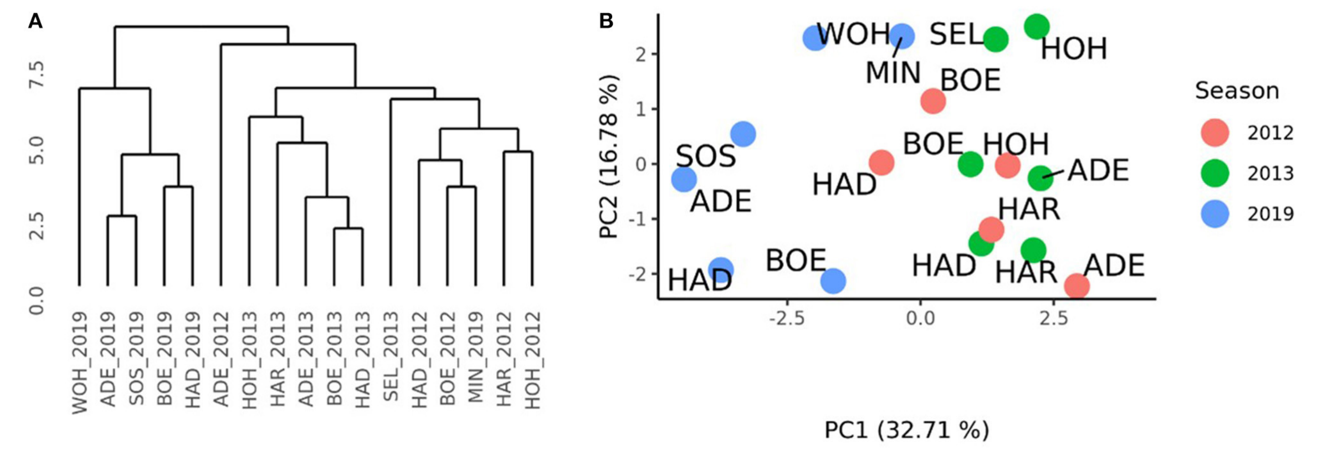


FIGURE 5 | Characterization of the environments of the HYWHEAT and RRGS experiments of the growing seasons 2011/2012, 2012/2013, and 2018/2019, based on satellite-based images. **(A)** Dendrogram based on the mean vegetation condition index (VCI) profiles of 16 location times year combinations (location_year) used to perform grain yield trials in the year 2012 and 2013 for the training population and in the year 2019 to assess the selection gain of the reciprocal recurrent genomic selection program. **(B)** PCA based on the mean VCI profiles of 16 location times year combinations. The locations were indicated as ADE, Adenstedt; BOE, Boehnshausen; HAD, Hadmersleben; HA*R*, Harzhof; HOH, Hohenheim; MIN, Mintraching; SEL, Seligenstadt; SOS, Sossmar; WOH, Wohlde.

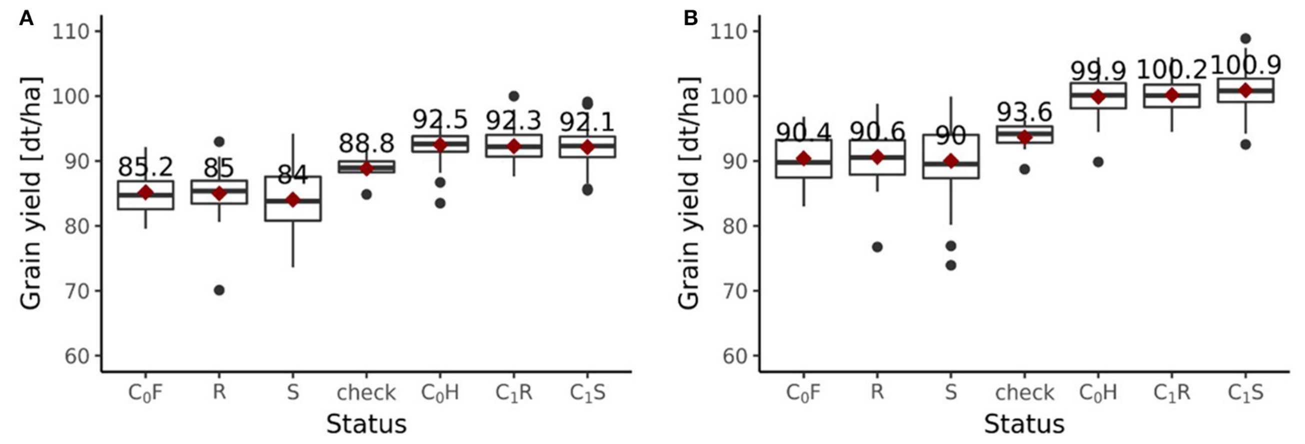


FIGURE 6 | Grain yield performance depending on the status of genotypes evaluated in the 2019 experiment. **(A)** Performances of the fractions from the breeding population with all six environments of 2018/2019 included. **(B)** Performances of the fractions from the breeding population with only 4 environments of 2018/2019 included. Status indicates the affiliation of each group of genotypes to a specific fraction within the breeding program. Female parent lines from the C_0 cycle are indicated as C_0F, female parent lines from the randomly selected fraction of the C_1 cycle are indicated as R, female parent lines from the genomic-selected fraction of the C_1 cycle are indicated as S, checks are indicated as "check," hybrids from the C_0 cycle are indicated as C_0H, hybrids from the randomly selected fraction of the C_1 cycle are indicated as C_1R, hybrids from the genomic-selected fraction of the C_1 cycle are indicated as C_1S.

which illustrates the great opportunity to accelerate classical RRS programs. The RRGS program focused exclusively on the female pool and can be viewed as a special case of RRGS in which only the allele frequencies in the pool of female parent lines have been shifted with respect to the frequencies of favorable alleles in the pool of male parent lines.

This situation implies consequences for the determination of selection directions, especially in the case of overdominance, $k > 1$, with $k = \frac{d}{a}$, where d denotes the dominance effect and a denotes the additive effect. If overdominance is present at a given locus, RRGS aims to fix different alleles in the pool of female parental lines and in the pool of male parental lines, thus guarantees the desired complementarity among the two heterotic groups. For loci with $k > 1$, at which the pool of male parent lines has a fixed allele, RRGS will result in the fixation of the complementary allele in the pool of female parent lines. If the allele is not fixed in the pool of the male lines, and no selection is applied to the pool of male parental lines, complementarity among the heterotic groups cannot be achieved.

If $0 < k \leq 1$, i.e., in the presence of partial dominance, RRGS aims to ultimately fix the favorable allele in both heterotic groups.

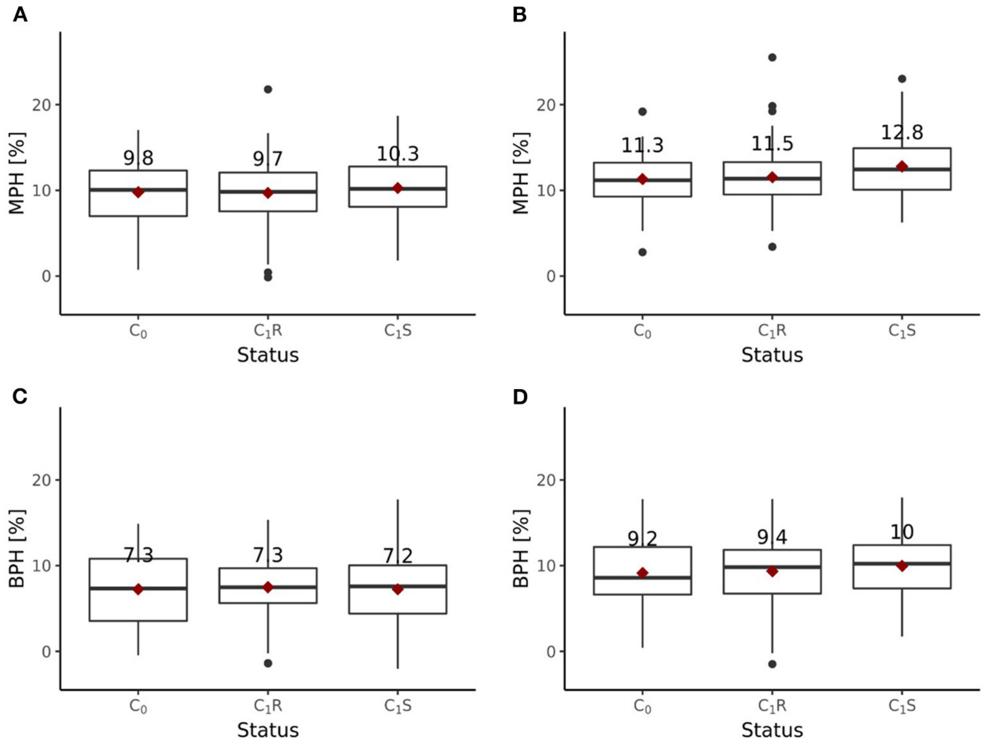


FIGURE 7 | Midparent heterosis (MPH) and better parent heterosis (BPH) for hybrids generated in the reciprocal recurrent selection program. MPH [%] estimated based on trials performed **(A)** at 6 locations and **(B)** 4 locations, excluding 2 stress environments. BPH [%] estimated based on trials performed **(C)** at 6 locations and **(D)** 4 locations, excluding 2 stress environments.

In the case where the male heterotic group is not fixed for the favorable allele, the optimal configuration cannot be achieved if the male heterotic group is not subject to selection.

For loci that exhibit negative dominance, i.e., $k < 0$, the desired selection direction is to fix the favorable allele in both heterotic groups. Complications arise when the unfavorable allele is present in the male heterotic group. Furthermore, if $k < -1$, i.e., negative overdominance is present, RRGS is directed toward fixation of the favorable allele only if the frequency, p, of the favorable allele is above the threshold $p > (k+1)/2k$ (Rembe et al., 2019).

In the present breeding program, the male heterotic group was kept constant between the C_0 and the C_1 cycle. As described above, this approach would not be expedient to reach the ideal allelic configurations between the two heterotic groups. However, the applied selection scheme is capable to evaluate the effectiveness of a selection that is conducted with respect to the allele frequencies within both heterotic groups. Therefore, the experimental design can serve as a model case for an RRGS breeding program.

The results of the field trials indicate that heterosis increased through RRGS (**Figure 7**). The selected fraction of the C_1S hybrids showed significantly higher midparent heterosis than the C_0 hybrids, but no significantly different better parent heterosis. In contrast, the C_1R hybrids did not show increased midparent or better parent heterosis compared to the C_0 hybrids. These findings highlight that the implemented selection models, which focused on additive and dominance effects, had an impact.

To evaluate the success of the RRGS program in more detail, the expected response to selection was compared to the observed response to selection. The expected response considering genomic selection at the F_2 and $F_{5:6}$ levels was $R_{\text{exp}} = 2.6\ dt\ ha^{-1}$, which was much lower than the observed response considering all six environments ($R_{obs\_6E} = -0.4\ dt\ ha^{-1}$) or the four environments ($R_{obs\_4E} = 1.0\ dt\ ha^{-1}$). The difference between R_{obs_6E} and R_{obs_E} clearly suggests that different growing conditions in the environments impacted the assessment of the response to selection. But even R_{obs_4E} was 2.6 times smaller than the expected response of selection R_{exp}, indicating that the implemented RRGS breeding program falls short of expectations. This observation can be mainly attributed to a high amount of genotype-by-year interactions between the 2011/2012, 2012/2013, and 2018/2019 experiments as highlighted in the detailed analyses of the interaction between genotypes and years (**Figures 4**, **5**). Multi-year testing could be an option to reduce the risk of unsuitable selection decisions.

So far, there are no experimental studies that have evaluated the effectiveness of an RRGS breeding program in cereals. In an RGS breeding program in wheat for the less complex trait grain fructans compared to grain yield, significant genotype-by-environment interactions were observed with little effects on prediction accuracies (Veenstra et al., 2020). In contrast, in an RRS program in tropical maize focusing on grain yield, Kolawole et al. (2018) also observed that genotype-by-environment interactions negatively affected the observed response to selection.

As an alternative approach to estimate the expected response of selection, realized prediction ability was examined as the correlation between predicted average hybrid performances and the observed average hybrid performance of the 30 randomly drawn female parent lines from the C_1 cycle. When all six environments of the season 2018/2019 were included in the analysis, a realized prediction ability of 0.13 was observed. Excluding environments with stressful growing conditions for the 2018/2019 data set resulted in a realized prediction ability of 0.27. These realized prediction abilities of the 2018/2019 growing season are substantially lower than the prediction abilities estimated by cross validations based on the data of the HYWHEAT experiment conducted in the 2011/2012 and 2012/2013 growing seasons (Zhao et al., 2015). This can only partly be explained by the small sample size of 30 randomly drawn female parent lines from the C_1 cycle used to estimate the prediction abilities. Moreover, it is unlikely that the low realized prediction abilities have been caused through recombination. More likely, the lower realized prediction abilities are due to interaction effects between genotypes, locations, and years.

When the prediction abilities estimated based on the 30 randomly drawn female parent lines from the C_1 cycle are used to estimate the expected response to selection, the value decreases to $R_{exp\_6E} = 0.09\ dt\ ha^{-1}$ and $R_{exp\_4E} = 1.22\ dt\ ha^{-1}$, depending on whether stressful environments are included or not. In this case, R_{obs_4E} was only 1.22 times smaller than the expected response of selection R_{exp}. Consequently, it is pivotal to obtain genome-wide prediction models that are not biased due to interaction effects between genotypes, locations, and years. One promising approach to achieve this, is to account for interaction effects between genotypes and environments by implementing environmental cofactors into genome-wide prediction models (de los Campos et al., 2020). This facilitates to reduce the adverse effects due to interactions between genotypes and environments and to develop more sustainable genome-wide prediction models. In addition, aggregation of available medium size genomic and phenotypic data across different projects and perhaps even breeding programs into large data sets can help substantially to reduce confounding effects of genotype-environment interactions (Zhao et al., 2021). These adjustments seem urgently needed to further leverage the potential of RRGS.

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 at: https://doi.org/10.1093/database/baw033.

AUTHOR CONTRIBUTIONS

JR, EE, EK, CL, PT, and YZ conceived and designed the study. EE, VK, JS, PB, PV, NPh, NH, SK, NPf, and MG acquired and contributed data. MR processed the data, performed the analyses, and analyzed the results. YZ supervised the data analyses. MR, JR, and YZ interpreted the results and wrote the manuscript. EE, PT, VK, JS, PB, PV, EK, NPh, SK, NPf, CL, NH, and MG provided input. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by BMEL through BLE within the ZUCHTWERT project (Grant ID: 2814604113) and the HYFLOR project (Grant ID: 2818401A18).

ACKNOWLEDGMENTS

We thank GFPi and proWeizen for project coordination.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.703419/full#supplementary-material

REFERENCES

Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., et al. (2021). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theor. Appl. Genet.* 134, 279–294. doi: 10.1007/s00122-020-03696-9

Becker, W. A. (1975). *Manual of Quantitative Genetics*. 3rd ed. Pullman, WA: Academic Enterprises

Bernal-Vasquez, A.-M., Utz, H. F., and Peter, H. (2016). Outlier detection methods for generalized lattices : a case study on the transition from ANOVA to REML. *Theor. Appl. Genet.* 129, 787–804. doi: 10.1007/s00122-016-2666-6

Braun, H. J., Rajaram, S., and Van Ginkel, M. (1996). CIMMYT's approach to breeding for wide adaptation. *Euphytica* 92, 175–83. doi: 10.1007/BF00022843

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., and Thompson, R. (2017). *ASReml-R Reference Manual Version 4*. ASReml-R Reference Manual.

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

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, 1–17. doi: 10.1186/s12864-015-1866-9

Doney, D. L., and Theurer, J. C. (1978). Reciprocal recurrent selection in sugarbeet. *Field Crops Res.* 1, 173–181. doi: 10.1016/0378-4290(78)90020-5

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package RrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Eyherabide, G. H., and Hallauer, A. R. (1991). Reciprocal full-sib recurrent selection in maize: II. Contributions of additive, dominance, and genetic drift effects. *Crop Sci.* 31, 1442–1448. doi: 10.2135/cropsci1991.0011183x003100060009x

Geiger, H. H., and Miedaner, T. (2015). *Hybrid Rye and Heterosis*. John Wiley & Sons, Ltd.

Gowda, M., Zhao, Y., Würschum, T., Longin, C. F., Miedaner, T., Ebmeyer, E., et al. (2014). Relatedness severely impacts accuracy of marker-assisted selection for disease resistance in hybrid wheat. *Heredity* 112, 552–561. doi: 10.1038/hdy.2013.139

Hallauer, A. R., Russell, W. A., and Lamkey, K. R. (1988). *Corn and Corn Improvement*. In, edited by G. F. Sprague and J. W. Dudley, 3rd ed. Madison: Agronomy Pubilcations. Available online at: http://lib.dr.iastate.edu/agron_pubs/259

Hallauer, A. R., Carena, M. J., and Filho Miranda, J. B. (2010). *Quantitative Genetics in Maize Breeding*. 3rd ed. New York, NY: Springer-Verlag. doi: 10.1007/978-1-4419-0766-0

Hecker, R. J. (1985). Reciprocal recurrent selection for the development of improved sugarbeet hybrids. *J. Sugarbeet Res.* 23 47–58. doi: 10.5274/jsbr.23.1.47

Huang, M., Tang, Q., Yuan, H., and Zou, Y. (2017). Yield potential and stability in super hybrid rice and its production strategies. *J. Integr. Agric.* 16, 1009–1017. doi: 10.1016/S2095-3119(16)61535-6

Jiang, Y., Schmidt, R. H., Zhao, Y., and Reif, J. C. (2017). A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. *Nat. Genet.* 49, 22–24. doi: 10.1038/ng.3974

Kempe, K., Rubtsova, M., and Gils, M. (2014). Split-gene system for hybrid wheat seed production. *Proc. Natl. Acad. Sci. U.S.A.* 111, 9097–9102. doi: 10.1073/pnas.1402836111

Kinghorn, B. P., Hickey, J. M., and Werf, J. H. J. (2010). "Reciprocal Recurrent Genomic Selection (RRGS) for total genetic merit in crossbred individuals," in *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production* (Leipzig).

Kolawole, A. O., Menkir, A., Blay, E., Ofori, K., and Kling, J. G. (2018). Genetic advance in grain yield and other traits in two tropical maize composites developed via reciprocal recurrent selection. *Crop Sci.* 58, 2360–2369. doi: 10.2135/cropsci2018.02.0099

Li, J., Schulz, B., and Stich, B. (2010). Population structure and genetic diversity in elite sugar beet germplasm investigated with SSR markers. *Euphytica* 175, 35–42. doi: 10.1007/s10681-010-0161-8

Liu, F., Jiang, Y., Zhao, Y., Schulthess, A. W., and Reif, J. C. (2020a). Haplotype-based genome-wide association increases the predictability of leaf rust (puccinia triticina) resistance in wheat. *J. Exp. Bot.* 71, 6958–6968. doi: 10.1093/jxb/eraa387

Liu, F., Zhao, Y., Beier, S., Jiang, Y., Thorwarth, P., H., et al. (2020b). Exome association analysis sheds light onto leaf rust (puccinia triticina) resistance genes currently used in wheat breeding (*Triticum Aestivum* L.). *Plant Biotechnol. J.* 18, 1396–1408. doi: 10.1111/pbi.13303

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

Longin, C., Friedrich, H., Gowda, M., Mühleisen, J., Ebmeyer, E. (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., Friedrich, H., and Reif, J. C. (2012). Redesigning the exploitation of wheat genetic resources. *Trends Plant Sci.* 19, 631–636. doi: 10.1016/j.tplants.2014.06.012

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27, 209–220.

Melonek, J., Duarte, J., Martin, J., Beuf, L., Murigneux, A., Varenne, P., et al. (2021). The genetic basis of cytoplasmic male sterility and fertility restoration in wheat. *Nat. Commun.* 12:1036. doi: 10.1038/s41467-021-21225-0

Mühleisen, J., Maurer, H. P., Stiewe, G., Bury, P., and Reif, J. C. (2013). Hybrid breeding in barley. *Crop Sci.* 53, 819–824. doi: 10.2135/cropsci2012.07.0411

Mühleisen, J., Piepho, H.-P., Maurer, H. P., Longin, C. F. H., and Reif, C. J. (2014). Yield stability of hybrids versus lines in wheat, barley, and triticale. *Theor. Appl. Genet.* 127, 309–316. doi: 10.1007/s00122-013-2219-1

Perez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Reif, J. C., Zhao, Y., Würschum, T., Gowda, M., and Hahn, V. (2013). Genomic prediction of sunflower hybrid performance. *Plant Breed.* 132, 107–114. doi: 10.1111/pbr.12007

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.* 132:3244. doi: 10.1007/s00122-018-3244-x

Santantonio, N., Atanda, S. A., Beyene, Y., Varshney, R. K., Olsen, M., Jones, E., et al. (2020). Strategies for effective use of genomic information in crop breeding programs serving Africa and South Asia. *Front. Plant Sci.* 11:353. doi: 10.3389/fpls.2020.00353

Schulthess, A. W., Zhao, Y., Longin, C. F. H., and Reif, J. C. (2018). Advantages and limitations of multiple-trait genomic prediction for fusarium head blight severity in hybrid wheat (*Triticum Aestivum* L.). *Theor. Appl. Genet.* 131, 685–701. doi: 10.1007/s00122-017-3029-7

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

Souza, C. L., Barrios, S. C. L., and Moro, G. V. (2010). Performance of maize single-crosses developed from populations improved by a modified reciprocal recurrent selection. *Sci. Agric.* 67, 198–205. doi: 10.1590/s0103-90162010000200011

Tardin, F. D., Pereira, M. G., Gabriel, A. P. C., Amaral Júnior, A. T., and Souza Filho, G. A. (2007). Selection index and molecular markers in reciprocal recurrent selection in maize. *Cropp Breed. Appl. Biotechnol.* 7, 225–233. doi: 10.12702/1984-7033.v07n03a01

Thorwarth, P., Liu, G., Ebmeyer, E., Schacht, J., Schachschneider, R., Kazman, E., et al. (2019). Dissecting the genetics underlying the relationship between protein content and grain yield in a large hybrid wheat population. *Theor. Appl. Genet.* 132, 489–500. doi: 10.1007/s00122-018-3236-x

Thorwarth, P., Piepho, H. P., Zhao, Y., Ebmeyer, E., Schacht, J., Schachschneider, R., et al. (2018). Higher grain yield and higher grain protein deviation underline the potential of hybrid wheat for a sustainable agriculture. *Plant Breed.* 137, 326–337. doi: 10.1111/pbr.12588

Troyer, A. F. (1999). Background of U.S. Hybrid Corn. *Crop Sci.* 39, 601–626. doi: 10.2135/cropsci2004.3700

United Nations Office for Outer Space Affairs (2021). *Step by Step: Recommended Practice Drought Monitoring Including a Cloud Mask (R) - Example Central America | UN-SPIDER Knowledge Portal*. Vienna: Recommended Practice: Drought Monitoring Using the Vegetation Condition Index (VCI) (2021).

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

Veenstra, L. D., Poland, J., Jannink, J. L., and Sorrells, M. E. (2020). Recurrent genomic selection for wheat grain fructans. *Crop Sci.* 60, 1499–1512. doi: 10.1002/csc2.20130

Watson, A., Ghosh, S., Williams, M. J., Cuddy, W. S., Simmonds, J., Rey, M. D., et al. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nat. Plants* 4, 23–29. doi: 10.1038/s41477-017-0083-8

Zhao, Y., Li, Z., Liu, G., Jiang, Y., Maurer, H. P., Würschum, T., et al. (2015). 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., Thorwarth, P., Jiang, Y., Philipp, N., Schulthess, A.W., Gils, M., et al. (2021). Unlocking big data doubled the accuracy in predicting the grain yield in hybrid wheat. *Sci. Adv.* 7. doi: 10.1126/sciadv.abf9106

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Sci.* 53, 802–810. doi: 10.2135/cropsci2012.08.0463

Zink, M., Samaniego, L., Kumar, R., Thober, S., Mai, J., Schafer, D., et al. (2016). The German drought monitor. *Environ. Res. Lett.* 11:74002. doi: 10.1088/1748-9326/11/7/074002

Conflict of Interest: EE and NPf were employed by KWS LOCHOW GmbH. PT, VK, and SK were employed by KWS Saat SE & Co. KGaA. JS, PB, and PV were employed by Limagrain GmbH. EK and NPh was employed by Syngenta Seeds GmbH. NH was employed by Nordsaat Saatzucht GmbH. MG was employed by Nordsaat Saatzucht GmbH and is presently employed by Strube Research GmbH & Co. KG.

The remaining 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.

ORIGINAL RESEARCH
published: 06 October 2021
doi: 10.3389/fpls.2021.735285

Improving Selection Efficiency of Crop Breeding With Genomic Prediction Aided Sparse Phenotyping

Sang He[1,2]*, *Yong Jiang*[3], *Rebecca Thistlethwaite*[4], *Matthew J. Hayden*[1,5], *Richard Trethowan*[4,6] *and Hans D. Daetwyler*[1,5]*

[1]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC, Australia, [2]CAAS-IRRI Joint Laboratory for Genomics-Assisted Germplasm Enhancement, Agricultural Genomics Institute in Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China, [3]Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany, [4]School of Life and Environmental Sciences, Plant Breeding Institute, Sydney Institute of Agriculture, The University of Sydney, Narrabri, NSW, Australia, [5]School of Applied Systems Biology, La Trobe University, Bundoora, VIC, Australia, [6]School of Life and Environmental Sciences, Plant Breeding Institute, Sydney Institute of Agriculture, The University of Sydney, Cobbitty, NSW, Australia

OPEN ACCESS

Edited by:
Valentin Wimmer,
KWS Saat, Germany

Reviewed by:
Julio Isidro Sanchez,
Universidad Politécnica de Madrid, Spain
Surinder Banga,
Punjab Agricultural University, India

****Correspondence:***
Hans D. Daetwyler
hans.daetwyler@agriculture.vic.gov.au
Sang He
hesang@caas.cn

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *02 July 2021*
Accepted: *14 September 2021*
Published: *06 October 2021*

Citation:
He S, Jiang Y, Thistlethwaite R, Hayden MJ, Trethowan R and Daetwyler HD (2021) Improving Selection Efficiency of Crop Breeding With Genomic Prediction Aided Sparse Phenotyping. Front. Plant Sci. 12:735285. doi: 10.3389/fpls.2021.735285

Increasing the number of environments for phenotyping of crop lines in earlier stages of breeding programs can improve selection accuracy. However, this is often not feasible due to cost. In our study, we investigated a sparse phenotyping method that does not test all entries in all environments, but instead capitalizes on genomic prediction to predict missing phenotypes in additional environments without extra phenotyping expenditure. The breeders' main interest – response to selection – was directly simulated to evaluate the effectiveness of the sparse genomic phenotyping method in a wheat and a rice data set. Whether sparse phenotyping resulted in more selection response depended on the correlations of phenotypes between environments. The sparse phenotyping method consistently showed statistically significant higher responses to selection, compared to complete phenotyping, when the majority of completely phenotyped environments were negatively (wheat) or lowly positively (rice) correlated and any extension environment was highly positively correlated with any of the completely phenotyped environments. When all environments were positively correlated (wheat) or any highly positively correlated environments existed (wheat and rice), sparse phenotyping did not improved response. Our results indicate that genomics-based sparse phenotyping can improve selection response in the middle stages of crop breeding programs.

Keywords: sparse phenotyping, genomic prediction, multi-environment trials, response to selection, correlations between environments

INTRODUCTION

Genomic selection is a promising tool to assist plant breeding by accelerating selection gain per unit time (Endelman et al., 2014; Slater et al., 2016; Crossa et al., 2017; Voss-Fels et al., 2019). In crop breeding programs, there is a consensus that genomic selection should be applied in the early stages as phenotyping intensity during this period is low, especially for grain yield and hard-to-measure traits (Endelman et al., 2014; He et al., 2016). However, this genomic selection strategy depends on an independent and robust reference population, normally consisting of historical data collected across several years (Dawson et al., 2013; Rutkoski et al., 2015; Jarquin et al., 2016).

Another way to deploy genomic selection in breeding is through phenotype imputation (Hori et al., 2016), which does not require an independent reference population. In the middle stages of breeding programs (e.g., sometimes referred to as stages one or two), crop lines are regularly phenotyped in only a few environments. Increasing the number of testing environments during these stages with genomic selection could markedly boost selection accuracy, compared to the advanced stages where most selection candidates are intensively tested in many environments (He et al., 2016). However, budget and seed availability constraints make complete phenotyping of all selection candidates in many environments impractical earlier in the breeding program. Nevertheless, the phenotype imputation scheme proposed by Hori et al. (2016) suggests that lines do not need to be tested in each environment, i.e., sparse phenotyping. Instead, the phenotype of lines in untested environments is reliably predicted using methods such as multi-environment genomic prediction approaches based on the remaining observations in tested environments. Consequently, a multi-environment trial (MET) with more testing environments could improve overall selection accuracy.

Traditionally, the correlation between the best linear unbiased estimation (BLUE) of genetic value and the genomic estimated genetic value (GEGV) is used to evaluate genomic prediction accuracy (Heslot et al., 2012; Rutkoski et al., 2015; He et al., 2016; Jarquin et al., 2016). BLUEs are assumed to be the best benchmark of GEGV because they are derived directly from *per se* performance, which is trusted by plant breeders. However, the true genetic value is unknown and whether BLUE or GEGV is closer to the true genetic value is difficult to establish. Thus, rather than prediction accuracy, the focus could be on the actual breeders' interest, e.g., the response to selection, which can be inferred from a simulation-based approach (Piepho and Möhring, 2007) to directly evaluate the effectiveness of genomic selection. To our knowledge, no study has applied this approach to assess the effectiveness of genomic selection.

Our study utilized an Australian pre-breeding wheat population and a publicly available rice pureline population, both with complete and orthogonal phenotypic records of grain yield across 3 years and two sowing times, to investigate the potential of genomics-assisted sparse phenotyping to improve selection response in the context of multi-environment trials. We also investigate the relationship among environments and how this affects the effectiveness of the proposed genomics-assisted sparse phenotyping method.

MATERIALS AND METHODS

Wheat Data Set

The wheat grain yield data set used in this study originated from the data set used in He et al. (2019), which consisted of five individual data sets including 1,351 genotypes. The genotypes were evaluated from year 2012 to 2017 with two times of sowing (TOS) per year at Narrabri in north-western New South Wales, Australia. The randomized complete block design with two replicates was applied to measure five agronomic traits incl. Grain yield, plant height, protein content, screenings percentage, and thousand kernel weight. The experiments in the current study were based on 189 lines consistently tested from year 2015 to 2017 at two TOS per year. These lines composed an orthogonal data set with a dimension of 189 lines and six environments.

Phenotypic analysis was implemented for each data set to derive the repeatability estimate per environment (year–TOS combination) and best linear unbiased estimates (BLUEs) per line in each environment, as described in He et al. (2019). Specifically, the phenotypic data of each environment were analyzed using a mixed linear model. The field design relevant effects such as range, row, and replicates as well as residual effect were all designated as random effects which followed an identical and independent normal distribution. Genetic effects were in tandem treated as fix and random to derive the best linear unbiased estimates (BLUE) and repeatability of each environment. Another mixed linear model based on BLUE of the 189 genotypes in each environment was fitted to estimate the heritability of grain yield, which was formulated as $\hat{\mathbf{y}} = \mathbf{1_n}\mu + \mathbf{Z_r r} + \mathbf{Z_l l} + \varepsilon$, where n is the number of BLUE values, $\hat{\mathbf{y}}$ is n-dimensional vector of genotype BLUEs across environments, μ is the common intercept, $\mathbf{1}_\mathrm{n}$ is a n-dimensional vector of ones, **r** is the vector of environment effects, **l** is the vector of genetic effects of genotypes, **Zr** and **Zl** are incidence matrices for **r** and **l**, and **ε** is the random residual. Effects **r**, **l**, and **ε** were fitted as random effects following identical and independent normal distributions. The heritability of grain yield was estimated using formula: $1 - \frac{\bar{c}}{2\sigma_l^2}$, where $\bar{c}$ is the mean variance of a difference between two best linear unbiased predictions (BLUP) of genetic effects of genotypes (Cullis et al., 2006).

The genotypic data of the 189 lines used in this study were drawn from the genotypic data of 1,351 wheat lines fingerprinted with 41,666 90 K single nucleotide polymorphisms (SNP) in He et al. (2019). As the number of genotypes was reduced, SNPs were refiltered by removing those with a minor allele frequency of less than 0.05, which left 32,800 SNP for subsequent analyses. The genetic diversity of the 189 genotypes was inspected based on a cluster analysis using Rogers' distance (Roger, 1972) estimated by the 32,800 SNP. The correlation between environments was estimated by Pearson correlation coefficient between the BLUEs of the 189 genotypes in different environments.

Rice Data Set

The publicly available rice data set (Spindel et al., 2015) included 358 rice lines phenotyped for six agronomic traits across 4 years and two seasons, i.e., eight environments (year–season combinations). As in wheat, phenotypic analyses included estimation of repeatability per environment and BLUEs per line. Based on the BLUEs, we selected six environments with the greatest range in correlations between environments out of the total eight environments to evaluate the effectiveness of the sparse phenotyping method. Finally, 160 lines were available with orthogonal yield phenotypic data in all six environments. Genotyping-by-sequencing (GBS) genotypes for 108,024 SNPs were quality controlled as follows. Low-quality SNPs with MAF less than 0.05 and call rate less than 0.9 were removed. Eventually, 46,232 SNPs were available for the 160 used lines. The correlation between environments was estimated by Pearson correlation coefficient between the BLUEs of the 160 genotypic lines in different environments.

Multi-Environment Genomic Prediction Model

A multi-environment genomic prediction model explicitly describing genotype-by-environment interactions was used:

$$\mathbf{y} = \mathbf{1_{mn}}\mu + \mathbf{Z_v v} + \mathbf{Z_g g} + \mathbf{gv} + \mathbf{e}$$

where m is the number of environments, n is the number of genotypes, $\mathbf{y}$ is a $m \times n$ vector of BLUEs of genotypes in each environment, μ is the common intercept, $\mathbf{v}$ is the m-dimensional vector of environment main effect, $\mathbf{g}$ is the n-dimensional vector of additive genetic main effect of genotypes, $\mathbf{gv}$ is the $m \times n$ vector of genotype-by-environment interaction effects, $\mathbf{e}$ is the random residual, $\mathbf{Zv}$ is the incidence matrices for $\mathbf{v}$, and $\mathbf{Zg}$ is the incidence matrices for $\mathbf{g}$. We assumed $\mathbf{v} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}\sigma_{\mathrm{v}}^2\right)$, $\mathbf{g} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{G}\sigma_{\mathrm{g}}^2\right)$, $\mathbf{gv} \sim \mathrm{N}\left(0, \left[\mathbf{Z_g G Z_g'}\right] \odot \mathbf{Z_v Z_v'}\sigma_{\mathrm{gv}}^2\right)$, and $\mathbf{e} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}\sigma_{\mathrm{e}}^2\right)$, where $\odot$ is the Hadamard product of matrices, σ_{g}^2, σ_{gv}^2, and σ_{e}^2 are their variance components, respectively, for genotype, genotype-by-environment interaction effects, and random residual. $\mathbf{G}$ is the genomic relationship matrix proposed by VanRaden (2008) constructed based on SNP genotypic profiles. The genomic prediction model was run in R (R Core Team, 2016) using the BGLR package (de los Campos and Pérez-Rodríguez, 2016). Iteration times were fixed to 30,000, and the first 5,000 times were set as burn-in.

Sparse Phenotyping Method

We compared the selection response of the complete phenotyping trial in fewer environments with a sparse genomic phenotyping method in additional environments. In this sense, all possible combinations of three environments out of the total six environments were used as the complete phenotyping trials, which retained total phenotypic values (BLUEs per environment). Phenotypic values in combinations of four, five, and six environments (there is just one combination using all six environments) were proportionally masked to create the sparse phenotyping trials. The percentage of phenotypic values retained in the 4-, 5-, and 6-environment combinations was 75, 60, and 50%, respectively, which made the phenotyping intensity in all 3-, 4-, 5-, and 6-environment combinations equivalent. Thus, the number of BLUEs and the amount of phenotype data collected was the same in all scenarios. There were 20 different combinations of three environments out of the total six environments. Each 3-environment combination was extended to three 4- or 5-environment combinations by including one or two environments from the remaining three environments. According to the phenotyping proportions (75, 60, and 50%) of 4-, 5-, and 6-environment combinations, phenotypic values in each 4-, 5-, and 6-environment combination were randomly masked one hundred times according to the cross-validation strategy two (CV2) in He et al. (2019). Specifically in this study, each genotype has six environment-specific BLUEs. We first attempted to randomly mask one BLUE of genotypes in the 4-, 5-, and 6-environment combinations to make the phenotyping proportions the same as the 3-environment complete phenotyping trial. If masking one BLUE was insufficient to meet the required phenotyping proportion, another BLUE of genotypes was masked until the required phenotyping proportion was reached.

Response to Selection

The genomic prediction model, also known as a mixed linear model, can be used to directly estimate the response to selection through a simulation-based approach following Piepho and Möhring (2007). Briefly, the multi-environment genomic prediction model was fitted using phenotypic records of complete phenotyping trial (3-environment combination) and phenotypic records of sparse phenotyping trials (4-, 5-, and 6-environment combinations). We were mainly interested in the relationship between the true genetic main effect $\mathbf{g}$ and its best linear unbiased prediction (BLUP) $\hat{\mathbf{g}}$, because the selection was based on the BLUP, while the response of selection was determined by the true values. In fact, the joint distribution of $\mathbf{g}$ and $\hat{\mathbf{g}}$ is multivariate normal and the corresponding variance–covariance matrix $\boldsymbol{\Omega} = \mathrm{var}\begin{pmatrix}\mathbf{g} \\ \hat{\mathbf{g}}\end{pmatrix}$ can be derived from the mixed model equations. Then, $\boldsymbol{\Omega}$ was eigendecomposed as $\boldsymbol{\Omega} = \mathbf{D\Lambda D'} = \boldsymbol{\Gamma\Gamma'}$, where $\mathbf{D}$ is the matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues, $\boldsymbol{\Gamma} = \mathbf{D}\sqrt{\boldsymbol{\Lambda}}$. The vector combining the true and predicted genetic main effects $\mathbf{w} = \begin{pmatrix}\mathbf{g} \\ \hat{\mathbf{g}}\end{pmatrix}$ could be simulated by $\mathbf{w} = \boldsymbol{\Gamma}\mathbf{z}$, where $\mathbf{z}$ is a 2n-dimensional vector of independent standard normal deviates because $\mathrm{var}(\mathbf{w}) = \mathrm{var}(\boldsymbol{\Gamma}\mathbf{z}) = \boldsymbol{\Gamma}\,\mathrm{var}(\mathbf{z})\boldsymbol{\Gamma}' = \boldsymbol{\Gamma\Gamma'} = \boldsymbol{\Omega}$ as desired.

For each 3-environment complete phenotyping trial, the responses to selection under varying selection ratios (corresponding to different selection intensities) ranging from 10 to 90% with a gap of 10% were simulated 10,000 times. In each simulation run, the vector $\mathbf{w}$ combining the true and predicted genetic main effects was simulated and a subset of genotypes ($\mathrm{S_q}$) with top p% ($p = 10$–90) of $\hat{\mathbf{g}}$ was selected. The response to selection of the simulation run ($\mathrm{q^{th}}$) was calculated as $\mathrm{R_q} = \frac{\sum_{i \in S_q} g_i}{\#(\mathrm{S_q})}$, where $\#(\mathrm{S_q})$ is the size of $\mathrm{S_q}$. For each selection ratio (10–90%), the average value of response to selection of the 10,000 runs was finally used as the achieved responses to selections of the complete phenotyping trial, i.e., $\mathrm{R} = \frac{\sum_{q=1}^{10000} \mathrm{R_q}}{10000}$. The responses to selections of each extended

4-, 5-, and 6-environment sparse genomic phenotyping trial scenario were simulated in the same manner based on only unmasked phenotypic values. The effectiveness of genomic selection was determined by comparing the achieved selection response between each complete phenotyping trial and its extended different sparse phenotyping trials. The difference between the achieved response of the complete phenotyping scenarios and responses from one hundred replicates of the corresponding extended sparse phenotyping scenarios (with random phenotype masking) under each selection ratio (10–90%) was statistically tested with Student's *t* tests.

RESULTS

Phenotypic Data and Population Structure

For the wheat data set, the overall heritability of grain yield was 0.38 and repeatability of each environment was above 0.4, indicating that the phenotypic data were of high quality (**Figure 1A**). The distribution of BLUEs in different environments was asymptotically normal (**Figure 1B**). Several large families were identified by clustering analysis and linkages existed across families (**Supplementary Figure 1**). The Rogers' distance values between any pair of genotypes ranged from 0.01 to 0.53. For the rice data set, the overall heritability was 0.83 and repeatability of each environment was over 0.4 (**Supplementary Figure 2A**). The distribution of BLUEs across different environments was near normal (**Supplementary Figure 2B**).

Correlations Between Environments

In the wheat data set, pairwise correlations ranged from −0.35 to 0.84 among the six environments (**Figure 2**). Among the 3-environment combinations, five combinations showed all positive pairwise correlations. Each 3-environment combination displayed at least one positive pairwise correlation (**Supplementary Table 1**). Inspecting the pairwise correlations within the twenty 3-environment combinations, four groupings became clear: (1) one pair of environments had high positive correlation 0.84, i.e., combinations 1–4; (2) environments where all pairwise correlations were positive, i.e., combinations 5, 11, and 19; (3) one pair of environments had negative correlations, i.e., combinations 6–7, 12–13, and 17–18; and (4) two pairs of environments had negative correlations, i.e., combinations 8–10, 14–16, and 20 (**Supplementary Table 1**).

In the rice data set, correlations of pairs of environments varied from 0.05 to 0.67 (**Supplementary Figure 3**). Among the 3-environment combinations, in one combination all correlations were below 0.18 and four combinations had one highly positive correlation of 0.67 (**Supplementary Figure 3**). Based on the pairwise correlations within the twenty 3-environment combinations, there were four distinct groupings: (1) one pair of environments with high positive correlation 0.67, i.e., combinations 10, 16, 19, 20; (2) all pairwise correlations moderately positive above 0.18, i.e., combinations 12, 13, 17, 18; (3) one pair of environments lowly positively correlated below 0.18, i.e., combinations 3, 4, 6–9, 11, 14, 15; and (4) more than one pair of environments lowly positively correlated below 0.18, i.e., 1, 2, 5 (**Supplementary Figure 3**).

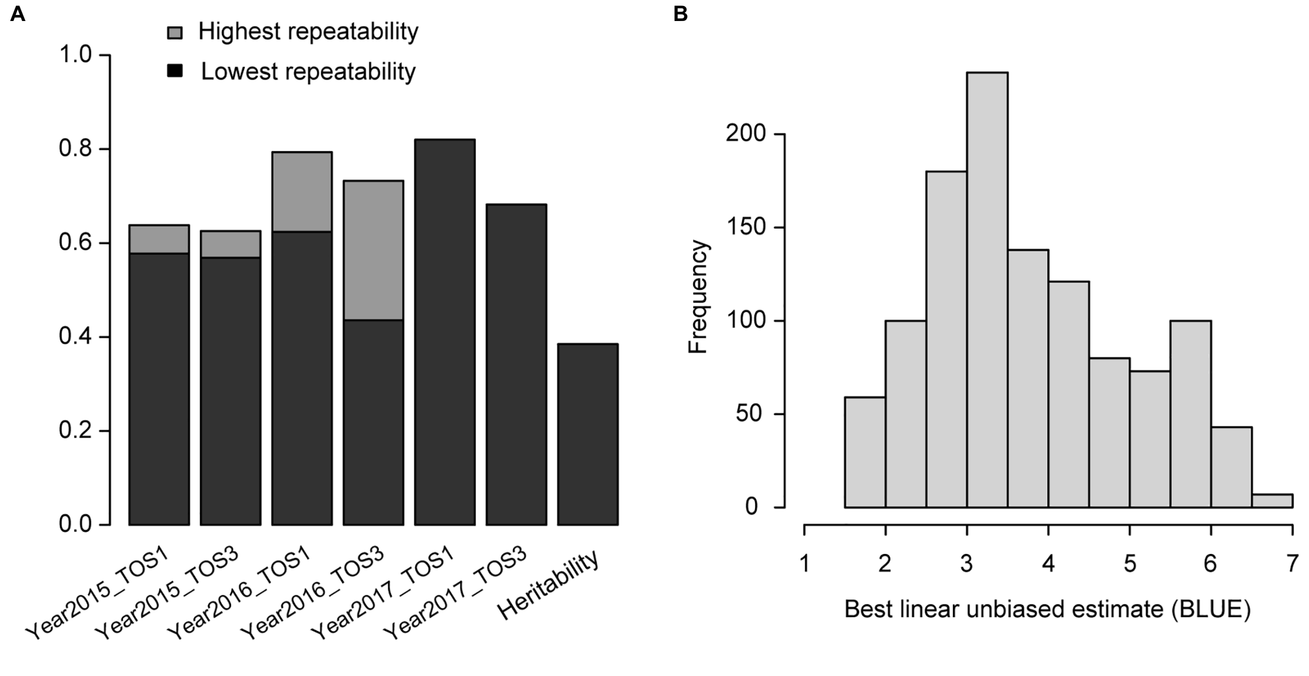


FIGURE 1 | Wheat – **(A)** heritability of grain yield and repeatability in each environment. The highest and lowest repeatability of specific environments evaluated in different data sets are shown in two grayscales; **(B)** distribution of best linear unbiased estimate (BLUE) of genotypes in different environments.

Simulated Response to Selection

For the wheat data set, twenty-one 4-environment combinations with sparse phenotyping applied had statistically significant higher responses to selection, compared to their equivalent 3-environment combination with complete phenotyping under each selection ratio, i.e., 10–90% (**Figure 3**). Most of the combinations contained one negative correlation between the three base environments with complete phenotypic records and one highly positive correlation (0.84) between the extension environment and the base environments (**Figure 3**). For the 5- and 6-environment combinations, there were twenty-three and seven sparse combinations showing higher response, respectively (**Figures 4, 5**). One negative correlation between the base environments and one highly positive correlation between expansion environment and base environments were also observed in the 5- and 6-environment combinations (**Figures 4, 5**). Comparison of the responses of all 3-environment combinations and their extended 4-, 5-, and 6-environment combinations identified five 3-environment combinations where the sparse phenotyping combinations did not result in a significantly higher response than the corresponding full 3-environment scenarios (combinations 1–4, 19; **Supplementary Table 2**). For most 3-environment complete phenotyping combinations, the responses achieved by the extended 4-environment sparse phenotyping scenarios were the highest compared to the 5- and 6-environment combinations (**Figure 6**).

For the rice data set, twenty-five 4-environment combinations sparse phenotyping scenarios showed statistically significant higher responses to selection than their corresponding 3-environment complete phenotyping combination under each selection ratio, i.e., 10–90% (**Supplementary Figure 4**). Most of these included two lowly positive correlations (<0.18) within the three complete phenotyping environments and/or one highly positive correlation (0.67) between the extended environment and one complete phenotyping environment (**Supplementary Figure 4**). For the 5- and 6-environment combinations, there were twenty-one and seven combinations, respectively, displaying higher response (**Supplementary Figures 5, 6**). Again, one highly positive correlation between the expansion environment and base environments and at least two lowly positive correlations within the base environments were observed in the 5- and 6-environment combinations (**Supplementary Figures 5, 6**). The 3-environment combinations with one highly positive correlation, i.e., group 1, showed no improved response from sparse phenotyping (**Supplementary Figures 4–6**). The responses of 4-environment sparse combination with one extended environment tended to be higher than those of 5- and 6-environment sparse combinations (**Supplementary Figure 7**).

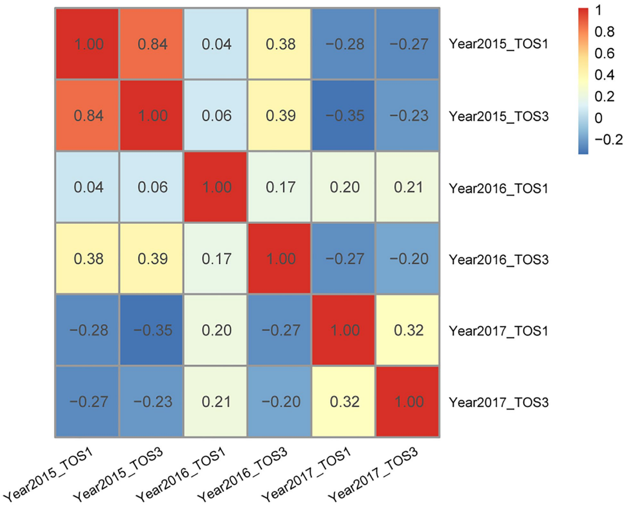


FIGURE 2 | Wheat – pairwise correlation between environments.

DISCUSSION

Our study investigated the potential of a genomics-assisted sparse phenotyping method *via* simulated selection responses based on a wheat and a rice data set. Results of both data sets showed that the sparse phenotyping can lead to a similar or greater response and provides information on genotype performance in more environments, compared to fully replicated trials. As the level of phenotyping (i.e., the number of observations) was the same in all scenarios, the advantage of sparse phenotyping was achieved with a similar budget. While families existed in the populations, our sparse phenotyping method tested each genotype in at least one environment. Consequently, as all genotypes were included in the reference set, the families did not introduce bias due to relatedness discrepancy to genomic prediction in the different phenotype masking scenarios.

Inclusion of Environment Correlation in Genomic Prediction Model Reduces the Benefit of Genomics-Assisted Sparse Phenotyping

In our study, a basic multi-environment genomic prediction model considering environments independent was used to simulate response to selection. Nevertheless, a sophisticated model that accommodates correlation between environments seems more reasonable in theory and more suited to be implemented. Jarquin et al. (2014) and Saint Pierre et al. (2016) demonstrated using environmental descriptors such as weather data to describe environmental relationship could improve genomic prediction accuracy. However, such environmental data are not always available. Martini et al. (2020) proposed to straightforwardly use phenotypic correlation of overlapped genotypes in different environments to specify the environmental relationship matrix. Thus, we also tested the effectiveness of the model in the wheat data set using correlation between BLUEs of unmasked genotypes in both environments to compile the environmental relationship matrix. Results showed that the sophisticated model including environmental correlation reduced the number of cases where

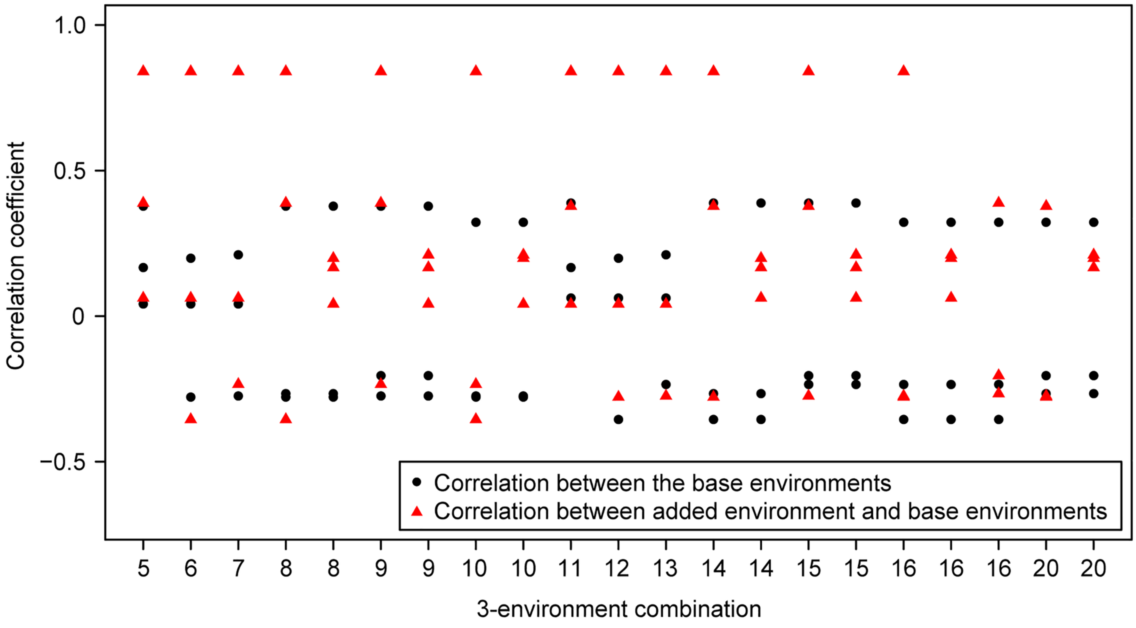


FIGURE 3 | Wheat – 3-environment combinations with complete phenotypic values showing statistically significant ($p < 0.05$) lower response to selection than their extended 4-environment combinations using genomics-assisted sparse phenotyping. Labels of horizontal axis are the scenario numbers of 3-environment combinations. Black dots represent correlation coefficients between the three base environments with complete phenotypic values. Red triangles indicate correlation coefficients between the added environment and base environments.

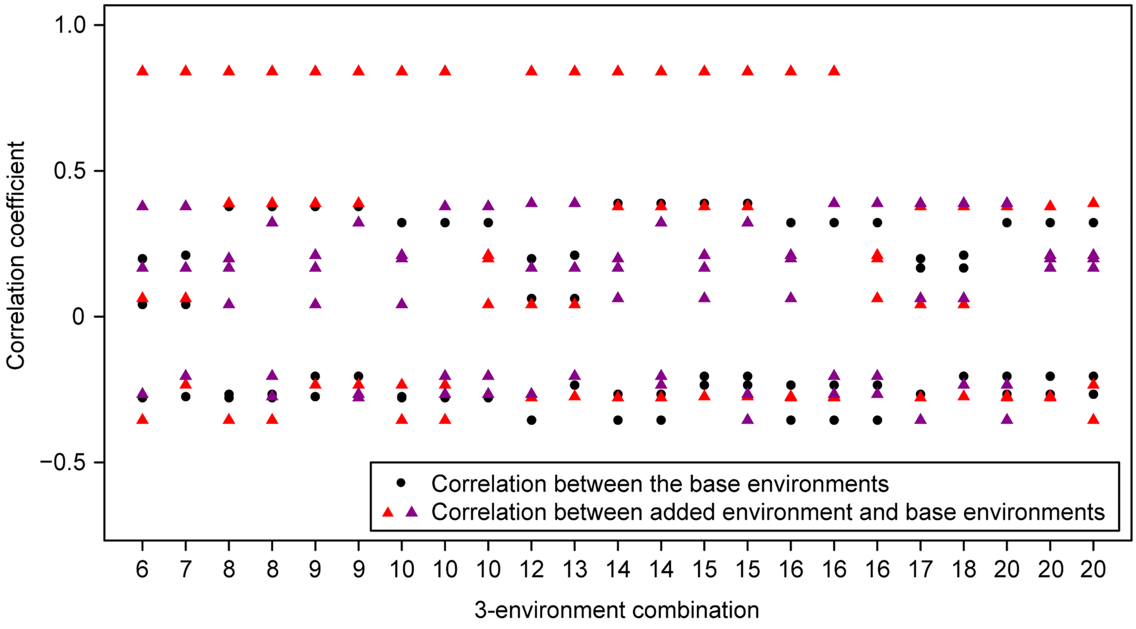


FIGURE 4 | Wheat – 3-environment combinations with complete phenotypic values showing statistically significant ($p < 0.05$) lower response to selection than their extended 5-environment combinations using genomics-assisted sparse phenotyping. Labels of horizontal axis are the scenario numbers of 3-environment combinations. Black dots represent correlation coefficients between the three base environments with complete phenotypic values. Triangles with different colors indicate correlation coefficients between separate added environments, i.e., the first or second added environment, and base environments.

the sparse phenotyping method displayed significantly higher response than complete phenotyping, as compared to the basic model (**Supplementary Figures 8–10**). This may be attributed to the number of genotypes used in our study being insufficient to reliably estimate the environmental relationship matrix (Martini et al., 2020). For the sparse phenotyping scenarios, the number of genotypes that can be used to estimate environmental relationship matrix, i.e., unmasked genotypes in both environments, would decrease even more. Particularly, when a total six environments were used, there was only one combination in which the sparse phenotyping performed significantly better (**Supplementary Figure 10**). This is because when the number of expansion environments increased, the number of unmasked genotypes with phenotypes in all environments reduced (**Supplementary Figure 11**), leading to a reduction in the reliability of correlation estimates. Alternatively, a more sophisticated model with unstructured environment covariances was also fitted (Burgueño et al., 2012). However, the phenotypic variance–covariance matrix was not always invertible when the sparse phenotyping pattern changed. Based on these results, we recommend to use the basic multi-environment genomic prediction model to compare the effectiveness of sparse and complete phenotyping strategies unless there are adequate common genotypes in different environments available to reliably estimate the environmental relationship matrix.

Effectiveness of Sparse Phenotyping Could Be Further Improved by Selective Phenotyping

Our study used a simple stochastic masking design to simulate the sparse phenotyping patterns on the basis that each genotype was tested in at least one environment. However, a more sophisticated selective phenotyping design could help improve the effectiveness of sparse phenotyping (Heslot and Feoktistov, 2020; Jarquin et al., 2020). Jarquin et al. (2020) proposed to completely phenotype a small proportion of genotypes in all environments to facilitate the estimation of environmental variance. As a result, substantial savings of phenotyping cost can be achieved while a high prediction accuracy was maintained. Heslot and Feoktistov (2020) demonstrated that precisely selecting a subset of genotypes for phenotyping based on relatedness could optimize the estimation of marker effect and tremendously increase prediction accuracy compared to randomly selecting a subset with equal size. This suggests that the unit of selection could shift to alleles being sufficiently replicated across environments. Therefore, instead of phenotyping each line in at least one environment, selecting a subset of lines would capitalize on genetic relationship and adding emphasis by testing some individuals in more environments to boost the overall phenotyping intensity could in turn further improve the effectiveness of sparse phenotyping. In this sense, further studies are needed to substantiate the merit of selective phenotyping design on promoting simulated response to selection of sparse phenotyping.

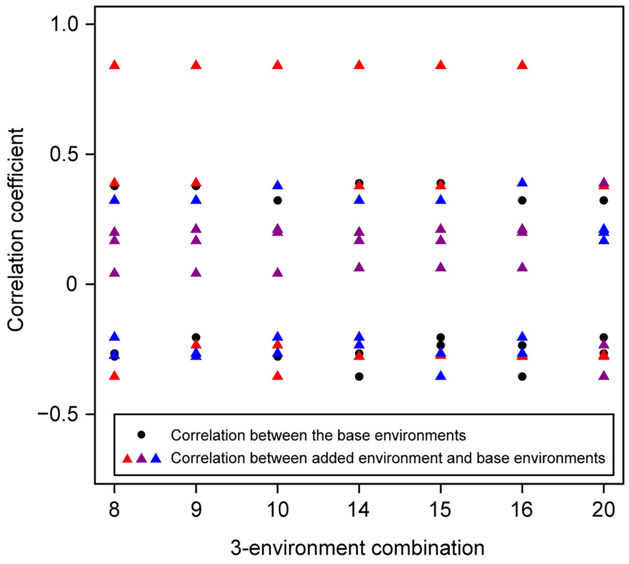


FIGURE 5 | Wheat – 3-environment combinations with complete phenotypic values showing statistically significant ($p < 0.05$) lower response to selection than using total six environments with genomics-assisted sparse phenotyping. Labels of horizontal axis are the scenario numbers of 3-environment combinations. Black dots represent correlation coefficients between the three base environments with complete phenotypic values. Triangles with different colors indicate correlation coefficients between separate added environments, i.e., the first, second, or third added environment, and base environments.

The Benefit of Sparse Phenotyping Can Be Anticipated From Correlations Between Environments

The correlations between environments in the wheat data set included high (e.g., 0.84), moderate (e.g., 0.32 and 0.38), low (e.g., 0.04 and 0.06), and negative (e.g., −0.28 and −0.35), which is representative of the types of environments encountered in plant breeding. These four groupings of 3-environment combinations are illustrated in **Table 1** and can be used to understand when sparse phenotyping can be beneficial.

Group 1 had a highly positive correlation (0.84) between environments and the sparse phenotyping method did not result in additional selection response, regardless of the number of expansion environments added (**Table 1**; **Figures 3–5**).

In group 2, all pairwise correlations were positive and when the extended environment was highly positively correlated (0.84) with any of the complete phenotyping environments, sparse phenotyping was always superior (**Table 1**; **Figure 3**; **Supplementary Tables 1**, **2**). However, this superiority was not maintained when additional environment(s) were included that were only poorly correlated with the complete phenotyping environments (**Figures 4, 5**; **Supplementary Tables 1**, **2**). As there was no expansion environment with a high positive correlation (0.84) with the complete phenotyping environments in combinations 1–4, it was not possible to determine whether adding such a highly positively correlated expansion environment would be beneficial or not. It is therefore possible the efficacy of sparse phenotyping is actually very similar in groups 1 and 2.

Group 3 had two pairs of environments with a positive correlation and one pair with a negative correlation. Here, the sparse phenotyping method consistently resulted in an additional selection response when the expansion environment was highly positively correlated (0.84) or even when several expansion

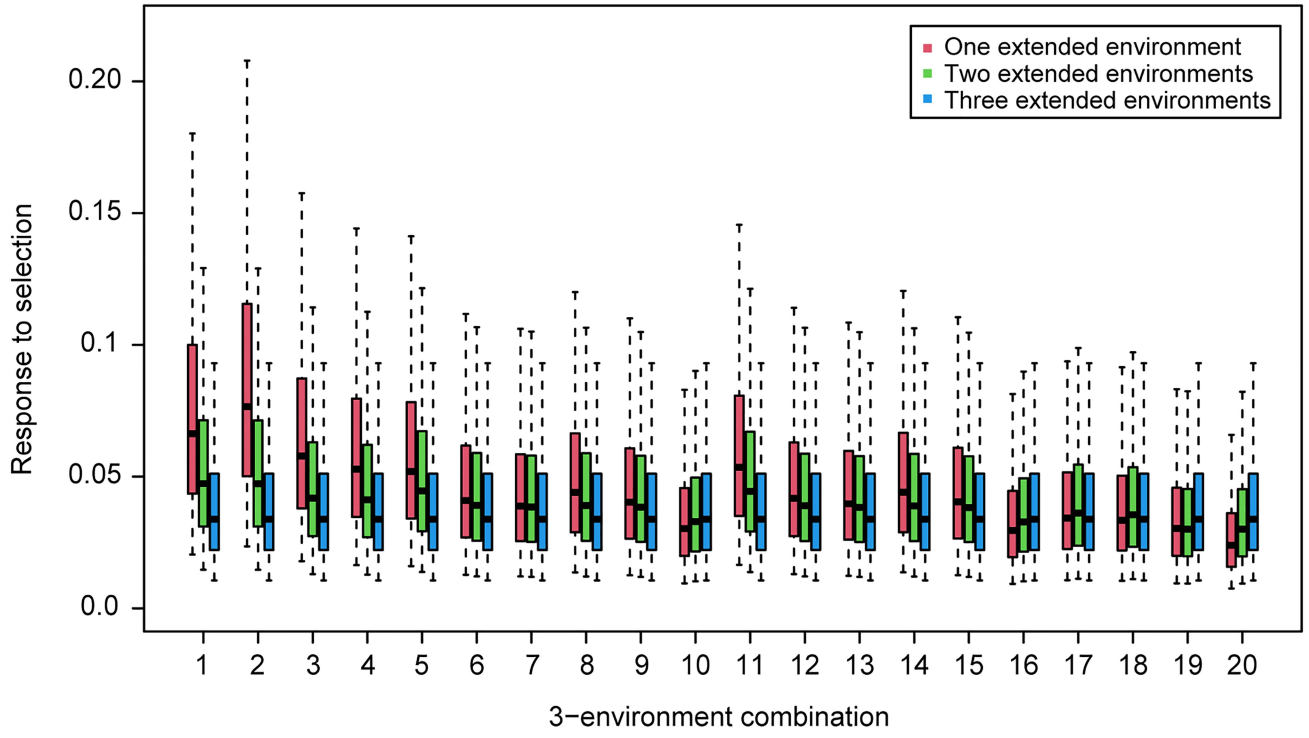


FIGURE 6 | Wheat – responses to selection of 4-environment (one extended environment), 5-environment (two extended environments) and 6-environment (three extended environments) sparse phenotyping combinations belonging to each 3-environment complete phenotyping combination. Labels of horizontal axis are the scenario numbers of 3-environment combinations.

environments were moderately positively correlated with the complete phenotyping environments (**Table 1**; **Figures 3, 4**; **Supplementary Tables 1**, **2**). This suggests that the robustness of group 3 is less than groups 1 and 2, and the superiority of including two expansion environments in group 3 depends on the relationship between the two expansion environments. In combination 17–18, no expansion environment was highly positively correlated with any of the complete phenotyping environments. However, two expansion environments were highly correlated (0.84), i.e., Year2015_TOS1 and Year2015_TOS3, and each was moderately positively correlated with one of the complete phenotyping environments, which made sparse phenotyping superior (**Figure 4**; **Supplementary Table 2**). In contrast, their *per se* 4-environment sparse phenotyping scenario did not show superiority (**Figure 3**; **Supplementary Table 2**).

For group 4, where one pair of environments had a positive correlation and two pairs a negative correlation, i.e., combinations 8–10, 14–16, and 20, sparse phenotyping resulted in a greater response when one expansion environment was highly correlated (0.84) or all expansion environments had moderate positive correlations with the complete phenotyping environments (**Table 1**; **Figures 3–5**; **Supplementary Tables 1**, **2**). In some cases, such as combination 16 and 20, even one extended environment with a moderate positive correlation with the complete phenotyping environments was superior (**Table 1**; **Figure 3**). This suggests that when environments are dissimilar, the sparse phenotyping method is particularly useful; a finding corroborated by the largest number of superior 5- and 6-environment combinations in group 4 (**Figures 4, 5**).

The relationship between correlations of environments and the benefit of sparse phenotyping was confirmed in the rice data set even though the range of correlations between environments was not as great as that observed in wheat.

Breeders are advised to consider the expected phenotypic correlation between environments when deciding whether genomics-assisted sparse phenotyping is of value. For instance, inspecting the correlations between environments observed in the wheat data set shown in **Table 1**, when the environments projected for complete phenotyping contain a highly positive correlation, the sparse phenotyping method does not increase selection response. For any other combination of complete phenotyping environments, adding one expansion environment that is positively highly correlated with any of the complete phenotyping environments will always be beneficial. When most complete phenotyping environments are negatively correlated, including more (≤3) expansion environments also consistently improved the response as long as one positive highly correlated expansion environment was added. It is worth noting that while adding one highly positively correlated expansion environment was of benefit, breeders could choose this environment for complete phenotyping if some prior knowledge was available, which would revert the combination to group 1. Nevertheless, adding positive correlation sparse phenotyping scenarios was generally of benefit (group 4,

TABLE 1 | Wheat – grouping of 3-environment combinations according to their utility of genomics-assisted sparse phenotyping over complete phenotyping.

Group	Complete phenotyping group correlation characteristics	Complete phenotyping three-environment combinations	Genomic sparse phenotyping better?		
			Plus 1 sparse environment	Plus 2 sparse environments	Plus 3 sparse environments
1	One highly positive correlation	1, 2, 3, 4	No	No	No
2	All correlations positive	5, 11, 19	Yes, when additional environment was positively highly correlated with the complete phenotyping environment	No	No
3	One negative correlation	6, 7, 12, 13, 17, 18	Yes, when additional environment was positively highly correlated with the complete phenotyping environment	Yes, when additional environments were positively highly or moderately correlated with the complete phenotyping environment, where the two moderately correlated environments need to be highly correlated	No
4	Two negative correlations	8, 9, 10, 14, 15, 16, 20	Yes, when additional environment was positively highly correlated with any or positively correlated with all complete phenotyping environments	Yes, when one additional environment was positively highly correlated with the complete phenotyping environment	Yes, when one additional environment was positively highly correlated with the complete phenotyping environment

Figure 3). However, in practice, breeders tend to choose environments that are distinct to select germplasm that are widely adapted.

It is also worth noting that the sparse phenotyping scenarios with less testing environments, e.g., one extended environment (4-environment combination) showed higher responses to selection than those with more environments, e.g., two and three extended environments (5- and 6-environment combinations; **Figure 6; Supplementary Figure 7**), which in part contradicts the experience on regular complete phenotyping that more testing environments imply higher selection accuracy and response to selection. Therefore, breeders may want to use one expansion environment when applying the sparse phenotyping approach as it would lead to a higher response. This would also facilitate the selection of extended environments as sparse phenotyping with more than one extended environment needs consideration of correlations between extended environments, which complicates the efficacy of the sparse phenotyping method.

Finally, although the budgets of the sparse phenotyping method with different number of expansion environments are theoretically identical, the actual cost would rise if the number of environments was increased, regardless of size. Hence, breeders should assess the practicality of the genomics-assisted sparse phenotyping approach based on both the relationship between testing environments and complexity of breeding program deployment.

CONCLUSION

Our study demonstrated that a genomics-assisted sparse phenotyping method can improve selection response for crop breeding, especially at the middle stages of a breeding program when multi-environment trials are not feasible due to cost. The sparse phenotyping approach was optimal when most of the complete phenotyping environments were negatively or lowly positively correlated and at least one of the extension environments was positively highly correlated with any of the complete phenotyping environment.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: wheat data is available upon request for non-commercial purposes. Requests to access these datasets should be directed to HD, hans.daetwyler@agriculture.vic.gov.au.

AUTHOR CONTRIBUTIONS

SH, HD, and YJ designed the study. SH conducted genomic prediction analyses and response simulations. RTr and RTh developed the plant populations and collected the phenotypes. MH oversaw genotyping. SH and HD wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study is funded by the Grain Research Development Corporation (GRDC, US00081), the University of Sydney and Agriculture Victoria.

ACKNOWLEDGMENTS

We thank staff and students at the Plant Breeding Institute at Narrabri and AgriBio for phenotyping and genotyping, respectively.

REFERENCES

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

Crossa, J., Perezrodriguez, P., Cuevas, J., Montesinoslopez, O. A., Jarquin, D., Campos, G. D. L., 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. *J. Agric. Biol. Environ. Stat.* 11, 381–393. doi: 10.1198/108571106X154443

Dawson, J. C., Endelman, J. B., Heslot, N., Crossa, J., Poland, J., Dreisigacker, S., et al. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crop Res.* 154, 12–22. doi: 10.1016/j.fcr.2013.07.020

de los Campos, G., and Pérez-Rodríguez, P. (2016). BGLR: Bayesian Generalized Linear Regression. R package version 1.

Endelman, J. B., Atlin, G. N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M. E., et al. (2014). Optimal Design of Preliminary Yield Trials with genome-wide markers. *Crop Sci.* 54, 48–59. doi: 10.2135/cropsci2013.03.0154

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

He, S., Thistlethwaite, R., Forrest, K., Shi, F., Hayden, M. J., Trethowan, R., et al. (2019). Extension of a haplotype-based genomic prediction model to manage multi-environment wheat data using environmental covariates. *Theor. Appl. Genet.* 132, 3143–3154. doi: 10.1007/s00122-019-03413-1

Heslot, N., and Feoktistov, V. (2020). Optimization of selective phenotyping and population design for genomic prediction. *J. Agric. Biol. Environ. Stat.* 25, 579–600. doi: 10.1007/s13253-020-00415-1

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

Hori, T., Montcho, D., Agbangla, C., Ebana, K., Futakuchi, K., and Iwata, H. (2016). Multi-task Gaussian process for imputing missing data in multi-trait and multi-environment trials. *Theor. Appl. Genet.* 129, 2101–2115. doi: 10.1007/s00122-016-2760-9

Jarquin, 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., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., et al. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3* 10:2725. doi: 10.1534/g3.120.401349

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

Martini, J. W., Crossa, J., Toledo, F. H., and Cuevas, J. (2020). On Hadamard and Kronecker products in covariance structures for genotype×environment interaction. *Plant Genome* 13:e20033. doi: 10.1002/tpg2.20033

Piepho, H. P., and Möhring, J. (2007). Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* 177, 1881–1888. doi: 10.1534/genetics.107.074229

R Core Team (2016). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/

Roger, J. (1972). *Measure of Genetic Similarity and Genetic Distance. Studies in Genetics VII. Vol. 7213.* Texas: University of Texas publication, 145–153.

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:plantgenome2014.2009.0046. doi: 10.3835/plantgenome2014.09.0046

Saint Pierre, C., Burgueño, J., Crossa, J., Dávila, G. F., López, P. F., Moya, E. S., et al. (2016). Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. *Sci. Rep.* 6, 1–11. doi: 10.1038/srep27312

Slater, A. T., Cogan, N. O., Forster, J. W., Hayes, B. J., and Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in Autotetraploid potato. *Plant Genome* 9, 1–15. doi: 10.3835/plantgenome2016.02.0021

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, 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:e1004982. doi: 10.1371/journal.pgen.1005350

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

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

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.735285/full#supplementary-material

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 Plant Science

ORIGINAL RESEARCH
published: 08 October 2021
doi: 10.3389/fpls.2021.720123

Increased Predictive Accuracy of Multi-Environment Genomic Prediction Model for Yield and Related Traits in Spring Wheat (*Triticum aestivum* L.)

Vipin Tomar[1,2,3*], *Daljit Singh*[4†], *Guriqbal Singh Dhillon*[5], *Yong Suk Chung*[6], *Jesse Poland*[4], *Ravi Prakash Singh*[7], *Arun Kumar Joshi*[1,3,7], *Yogesh Gautam*[1], *Budhi Sagar Tiwari*[2] and *Uttam Kumar*[1,3,7*]

[1] Borlaug Institute for South Asia, Ludhiana, India, [2] Department of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India, [3] International Maize and Wheat Improvement Center, New Delhi, India, [4] Department of Plant Pathology, Kansas State University, Manhattan, KS, United States, [5] Department of Biotechnology, Thapar Institute of Engineering & Technology, Patiala, India, [6] Department of Plant Resources and Environment, Jeju National University, Jeju-si, South Korea, [7] Global Wheat Program, International Maize and Wheat Improvement Center, Texcoco, Mexico

Edited by:
*Valentin Wimmer,
KWS Saat, Germany*

Reviewed by:
*Pedro José Martínez-García,
Spanish National Research
Council, Spain
Mian Abdur Rehman Arif,
Nuclear Institute for Agriculture and
Biology, Pakistan*

***Correspondence:**
*Vipin Tomar
viomics@gmail.com
Uttam Kumar
u.kumar@cgiar.org*

†Present address:
*Daljit Singh,
The Climate Corporation, Bayer Crop
Science, Creve Coeur, MO,
United States*

Specialty section:
*This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science*

Received: *03 June 2021*
Accepted: *03 September 2021*
Published: *08 October 2021*

Citation:
*Tomar V, Singh D, Dhillon GS,
Chung YS, Poland J, Singh RP,
Joshi AK, Gautam Y, Tiwari BS and
Kumar U (2021) Increased Predictive
Accuracy of Multi-Environment
Genomic Prediction Model for Yield
and Related Traits in Spring Wheat
(Triticum aestivum L.).
Front. Plant Sci. 12:720123.
doi: 10.3389/fpls.2021.720123*

Genomic selection (GS) has the potential to improve the selection gain for complex traits in crop breeding programs from resource-poor countries. The GS model performance in multi-environment (ME) trials was assessed for 141 advanced breeding lines under four field environments *via* cross-predictions. We compared prediction accuracy (PA) of two GS models with or without accounting for the environmental variation on four quantitative traits of significant importance, i.e., grain yield (GRYLD), thousand-grain weight, days to heading, and days to maturity, under North and Central Indian conditions. For each trait, we generated PA using the following two different ME cross-validation (CV) schemes representing actual breeding scenarios: (1) predicting untested lines in tested environments through the ME model (ME_CV1) and (2) predicting tested lines in untested environments through the ME model (ME_CV2). The ME predictions were compared with the baseline single-environment (SE) GS model (SE_CV1) representing a breeding scenario, where relationships and interactions are not leveraged across environments. Our results suggested that the ME models provide a clear advantage over SE models in terms of robust trait predictions. Both ME models provided 2–3 times higher prediction accuracies for all four traits across the four tested environments, highlighting the importance of accounting environmental variance in GS models. While the improvement in PA from SE to ME models was significant, the CV1 and CV2 schemes did not show any clear differences within ME, indicating the ME model was able to predict the untested environments and lines equally well. Overall, our results provide an important insight into the impact of environmental variation on GS in smaller breeding programs where these programs can potentially increase the rate of genetic gain by leveraging the ME wheat breeding trials.

Keywords: single-environment, multi-environments, genotyping by sequencing, genomic selection (GS), genomics predictions, best linear unbiased predictions, wheat

INTRODUCTION

Wheat (*Triticum aestivum* L.) is an essential cereal to secure global food security (Curtis and Halford, 2014). Significant efforts are needed to accelerate high-yielding varieties to fulfill future global wheat demand by 2050 (Hellin et al., 2012). Hence, the enhancement of grain yield (GRYLD) is a foremost target for wheat breeders. GRYLD is a complex trait administered by many small-effect loci with significant loci × loci interactions (Arzani and Ashraf, 2017; Sehgal et al., 2017). Moreover, the GRYLD trait is associated with strong genotype × environment interaction, which makes its genetic enhancement a difficult work.

Genomic selection (GS) integrates genome-wide dense markers and, as presented by Meuwissen et al. (2001), is a different marker-assisted selection approach. GS proves to be a powerful tool to improve the selection accuracy and prediction for quantitative traits in crop breeding (Crossa et al., 2017). GS utilizes a large set of, usually unidentified markers, spread over the whole genome in the same way as every quantitative trait locus (QTL) is in linkage disequilibrium (LD). GS is particularly beneficial for traits that cannot be evaluated on a few plants and for traits that are hard to estimate. It is still a vital issue for plant breeders to upsurge the accuracy of genomic prediction for selecting the advanced breeding lines.

The GS has been widely used in wheat breeding to predict various traits, such as yield, disease resistance, grain weight, heading, iron and zinc contents, end-use quality, and physiological traits (Charmet et al., 2014; Velu et al., 2016; Hayes et al., 2017; Juliana et al., 2017a,b; Norman et al., 2017; Lozada et al., 2019; Tomar et al., 2021). As such, GS embraces the prospects for the genomic enhancement of qualitative and quantitative traits. Many available GS models have been tested on various breeding and trait scenarios. Earlier numerous comparative studies of the GS model predictions in wheat showed that Random Forest and Reproducing Kernel Hilbert Space models performed better for traits of interest. However, any single GS model could not outperform other models (Pérez-Rodríguez et al., 2012; Charmet et al., 2014). Earlier studies have stated that many interconnected factors impact the overall model performance (Jannink et al., 2010; Heslot et al., 2012), such as heritability, population structure, statistical models, i.e., parametric and nonparametric models, cross-validation (CV) approaches, the genetics of traits, training population size and composition, marker density, and LD among markers and QTLs (Jannink et al., 2010; Pérez-Rodríguez et al., 2012; Crossa et al., 2017; Norman et al., 2018; Lozada et al., 2019).

The GS delivers the promise to accelerate genetic gain by increasing precision in selecting and shortening the breeding cycles. However, GS is a relatively new and advanced method for smaller and low-resource South Asian wheat breeding programs. Previously, substantial progress has been made in testing and validating various models for GRYLD and related traits in wheat in South Asia, albeit on larger breeding populations (De los Campos et al., 2009; Crossa et al., 2010, 2011, 2016; Heffner et al., 2011; Burgueño et al., 2012; Pérez-Rodríguez et al., 2012; Rutkoski et al., 2015; Juliana et al., 2017a,b, 2019; González-Camacho et al., 2018). These studies have highlighted the impact of environment and genotype × environment on the GS model performance. Therefore, to optimize the genetic gain from GS, the group of field-testing environments can be leveraged.

In this study, high-yielding, advanced wheat breeding lines from The International Maize and Wheat Improvement Center (CIMMYT) were evaluated for two consecutive wheat seasons (2017 and 2018) to adapt to the diverse environments of North and Central India. To evaluate the performance of multi-environment (ME) GS models on our specific set of selection environments, we tested different GS CV schemes mimicking the breeding schemes where untested lines and environmental performance are highly valuable to achieve the desired selection gains. This study is highly relevant particularly in the South Asian context where trial sizes are relatively small and broadly adapted wheat lines are sought after.

MATERIALS AND METHODS

Plant Material

A set of 141 South Asian spring wheat lines (*T. aestivum* L.) were selected from the International Yield Trial of CIMMYT International Nurseries (elite germplasm). These lines constitute a diverse association panel. The seeds of 141 genotypes were obtained from the Germplasm Resource Unit, CIMMYT (Mexico). Detailed information with a pedigree for each genotype is given in **Supplementary Table 1**.

Field Trials and Phenotypic Evaluation

The panel of selected lines was evaluated in field trials at the Borlaug Institute for South Asia (India) at Jabalpur (JBL) (23°14′00.6N and 80°04′40.7E) and Ludhiana (LDH) (30°59′28.74N and 75°44′10.87E), locations during the crop season for 2 years (2017 and 2018), genotypes were phenotyped and evaluated across all trials for four traits [days to maturity (DAYSMT), days to heading (DTHD), GRYLD, and thousand-grain weight (TGW)] (**Supplementary Table 2**). The experiment was conducted in two replications following the Randomized Block Design (RBD). The normal agronomic practice was followed for trial management. The row-to-row distance was maintained at 20 cm.

Genotyping-by-Sequencing and SNP Filtering

Genomic DNA was extracted from the fresh leaves of seedling wheat using the modified cetyltrimethylammonium bromide (CTAB) method (Dreisigacker et al., 2016). Genotyping-by-sequencing (GBS) was performed in Illumina HiSeq 2500 using a protocol suggested by Poland et al. (2012). Single nucleotide polymorphism (SNP) calling was performed using TASSEL version 5.2.43 (Bradbury et al., 2007) using the TASSEL-GBSv2 pipeline. Using Beagle version 4.1, the missing data were imputed with default settings. After quality control (filter criteria: sample call rate > 0.8, Minor allele frequency (MAF) ≥ 0.05, SNP call rate > 0.7), 14,563 polymorphic SNPs and 141 genotypes were selected for the follow-up analysis (**Supplementary Table 3**). Among polymorphic SNP markers, 40.66, 50.66, and 8.68% were from the A, B, and D genomes, respectively. With a genomic coverage of 13.9 GB and 14,563 markers across the genome, the average marker density was one marker per 0.95 Mb. The highest marker density with one marker per 0.54 Mb of chromosome 2B and the lowest marker density with one marker per 6.854 Mb at chromosome 4D were observed.

Statistical Analysis of Phenotypes

Each location-year combination is treated as a distinct environment for analysis purposes. Broad-sense heritability for each trait/environment combination was estimated at the plot level, and raw phenotypic values were adjusted to derive the best linear unbiased predictions (BLUPs) (**Supplementary Table 4**) for each trait at each environment using META-R (Alvarado et al., 2020) by using the following formula:

$$
\begin{aligned}
Y_{ik} &= \mu + Rep_i + Gen_k + \epsilon_{ik}(\text{within environments})\\
Y_{ijk} &= \mu + Env_i + Rep_j\,(Env_i) + Gen_k + Env_i \times Gen_k\\
&\quad + \epsilon_{ijk}(\text{across environments})
\end{aligned}
$$

where Y_{ik} is the trait of interest, μ is the mean effect, Rep_i is the effect of the ith replicate, Gen_k is the effect of the kth genotype, ϵ_{ik} is the error associated with the ith replication and the kth genotype, which is assumed to be normally and independently distributed, with mean 0 and homoscedastic variance. For across environments, Y_{ijk} is the trait response and the ith environment, $Rep_j(Env_i)$ is the effect of jth Rep in the ith environment, and $Env_i \times Gen_k$ is the environment × genotype interaction. The resulting analysis produced the adjusted trait phenotypic values in the form of BLUPs within and across environments. The BLUPs model considers genotypes as random effects, minimizing the effect of screening time and other environmental effects.

In addition, the components of the phenotypic variance of a given trait at an individual environment and across environments were also extracted to calculate the broad-sense heritability using the formula as follows:

$$
\begin{aligned}
H^2 &= \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{nReps}}(\text{within environments})\\
H^2 &= \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{ge}^2}{nEnvs} + \frac{\sigma_e^2}{(nEnvs \times nReps)}}\ (\text{across environments})
\end{aligned}
$$

where σ_g^2 and σ_e^2 are the genotype and error variance components, respectively, σ_{ge}^2 is genotype × environment interaction variance, nEnvs is the number of environments, and nReps is the number of replicates. All effects are considered random for calculating the BLUPs (**Supplementary Table 4**) and the broad-sense heritability. The BLUPs phenotypic distributions of GRYLD and other traits at each environment were plotted to check normality assumptions. Phenotypic and genetic correlations were calculated for each trait and environment combination in R software version 4.0.2. (R Core Team, 2019) using FactoMineR version 2.4 (Lê et al., 2008) and factoextra version 1.0.7 (Kassambara and Mundt, 2020).

Baseline Single-Environment (SE) Genomic BLUP Model (GBLUP), CV Schemes, and Predictive Ability

The baseline SE genomic prediction analysis was implemented in the BWGS program (Charmet et al., 2020). BWGS performs a GBLUP analysis using a marker-based relationship matrix. CV delivers an unbiased evaluation for the performance of a GS model; therefore, a 5-fold CV approach was implemented for reducing the unwanted bias (Kohavi, 1995), where the genotypes (for each trait separately) were randomly split into five equal-sized folds. SE_CV1 model was fitted with missing phenotypic values for the tested individuals. Prediction accuracy (PA) was subsequently calculated as the correlation of predicted breeding values with the observed phenotypic values for the missing genotypes. This step was repeated for each environment and fold separately. The genomic PA was then calculated by iteratively assigning 1-fold as the validation set and the remaining folds as the training set. This five-fold validation process was repeated 50 times to randomly shuffle the lines in each fold. The accuracy of the genomic predictions was measured as the Pearson's correlation between the predicted and actual trait BLUPs.

A mixed model of the simplified form was fitted for genomic predictions as follows:

$$y = Xb + Zg + e$$

where y is a vector of adjusted phenotypes, X is a design matrix relating the fixed effects to each genotype, b is a vector of fixed effects, Z is a design matrix connecting records to genetic values, g is a vector of additive genetic effects for a genotype, and e is a vector of random normal deviates with variance δ_e^2.

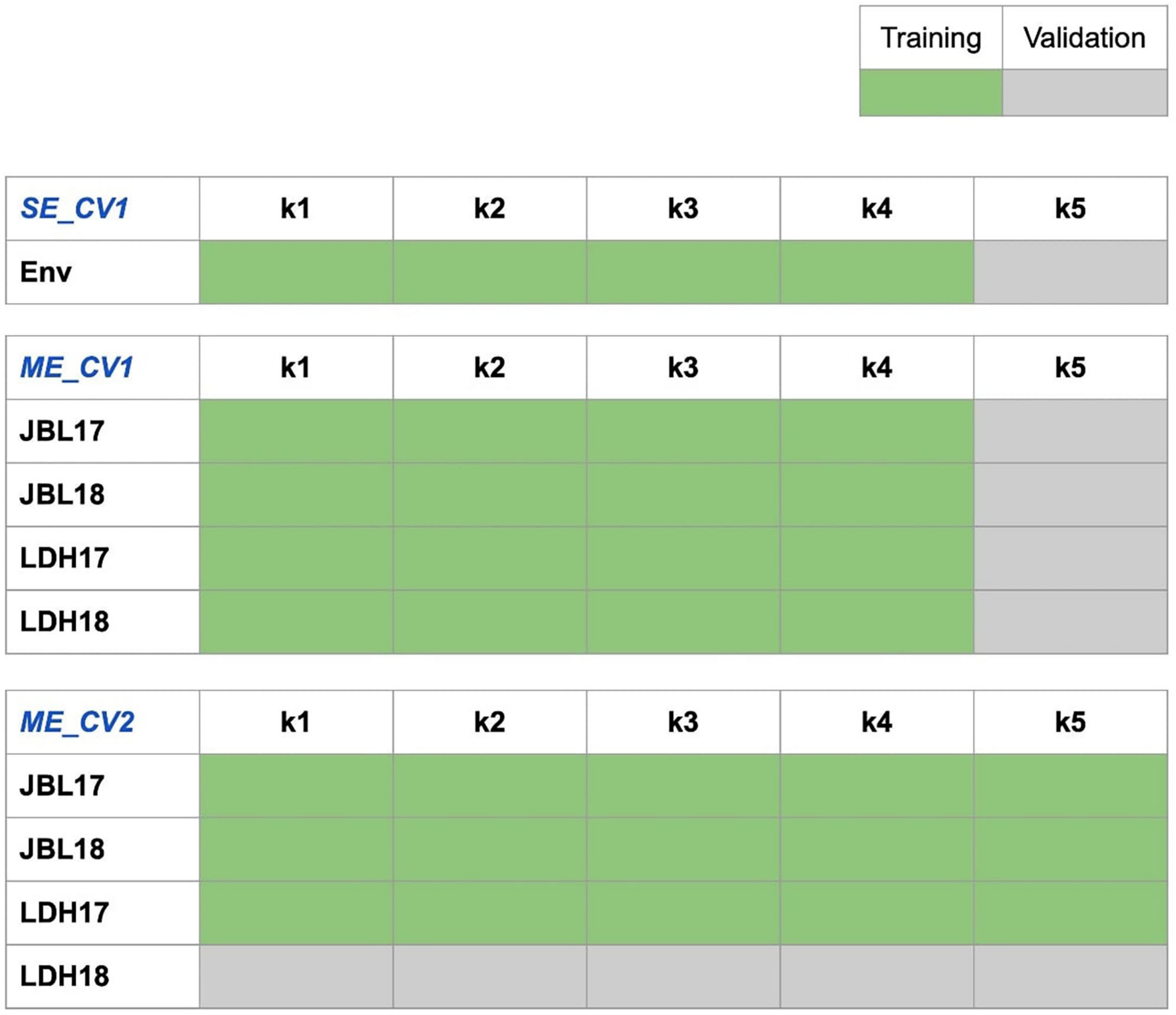
Advanced ME GBLUP Model, CV Schemes, and Predictive Ability

The advanced ME genomic prediction analysis was implemented in Solving Mixed Model Equations in the R (sommer) package (Covarrubias-Pazaran, 2016). Two types of ME_CV schemes representing actual breeding scenarios were implemented. The first scenario represents a use case where some genotypes are missing across all environments (ME_CV1). ME_CV1 was fitted by masking the phenotypic values of genotypes belonging to the test set across all environments. PA was calculated as the correlation of predicted and observed phenotypic values for the missing genotypes at each environment separately. In the second scenario, the entire trial or all genotypes are missing at one of the environments (ME_CV2). ME_CV2 was fitted by masking the phenotypic values of all lines in an SE. The trained model was then used to predict the breeding values of lines in the missing environment. PA was calculated as the correlation of predicted and observed phenotypic values of the tested lines. The CV schemes are illustrated in **Figure 1**.

In ME genomic predictions, the SE model was rewritten and implemented as follows:

$$y_{ij} = g_j + E_i + gE_{ij} + e_{ij}$$

where y_{ij} represents response of jth line in the ith environment ($i = 1, 2, \ldots\ldots i$, $j = 1, 2, \ldots\ldots\ j$; g_j is the effect of jth line with



Training	Validation
🟩	⬜

SE_CV1	k1	k2	k3	k4	k5
Env	🟩	🟩	🟩	🟩	⬜

ME_CV1	k1	k2	k3	k4	k5
JBL17	🟩	🟩	🟩	🟩	⬜
JBL18	🟩	🟩	🟩	🟩	⬜
LDH17	🟩	🟩	🟩	🟩	⬜
LDH18	🟩	🟩	🟩	🟩	⬜

ME_CV2	k1	k2	k3	k4	k5
JBL17	🟩	🟩	🟩	🟩	🟩
JBL18	🟩	🟩	🟩	🟩	🟩
LDH17	🟩	🟩	🟩	🟩	🟩
LDH18	⬜	⬜	⬜	⬜	⬜

FIGURE 1 | Prediction scheme for the single-environment (SE) and multi-environment (ME) genomic prediction models with two cross-validation schemes (CV1 and CV2) used in this study. SE_CV1 model: the SE prediction model with CV scheme 1 where a trait [e.g., grain yield (GRYLD)] is predicted at a time; we used 80% of individuals as the training set (phenotyped and genotyped, light green) and 20% of the individuals as the testing set (genotyped only, light gray with validation code for the trait to be predicted, yield as an example here). ME_CV1 model: the ME prediction model with CV scheme 1 for new un-phenotyped individuals; we used 80% of individuals as the training set (phenotyped for all traits and genotyped; light green) and 20% of the individuals as the validation set (genotyped but not phenotyped for any trait; light gray with validation code for the trait to be predicted, GRYLD as an example here). ME_CV2 model: the ME prediction model with CV scheme 2 where 100% of the information from other traits are available for the individuals to be predicted; we used 80% of individuals as the training set (phenotyped for all traits and genotyped; light green) and 20% of individuals as the validation set (phenotyped for associated traits but not for the targeted traits, and genotyped; light gray with predication code for the trait to be predicted, yield as an example here). Rectangles represent genotypes, and colors represent whether the phenotypic information was used (light green) or not (light gray with validation code for the trait to be predicted, GRYLD as an example) for the population. A similar scheme was applied for predicting days to heading (DTHD), days to maturity (DAYSMT), and thousand-grain weight (TGW).

$g = (g_{1 \ldots\ldots\ldots} g_j)$T∼N(0, δ_1^2Gg), δ_1^2 is the genomic variance, Gg is the genomic relationship matrix. E_i represents the effect of the ith environment. gE_{ij} is the random term that takes into account the interaction between the genomic effect of jth line and the ith environment with gE= ($g_1 \ldots\ldots\ldots g_j$)T∼N (0, $\delta_2^2 I_I \otimes$ G), where δ_2^2 is the interaction variance. Eij is a random residual effect of the jth line in the ith environment [N (0, δ_2^2)], where δ_2^2 is the residual variance.

RESULTS

Heritability, Correlations, and Trait Characterization

A range of variation was detected for GRYLD and other related traits across environments/years (LDH17 and LDH18 and JBL17 and JBL18). The highest averaged GRYLD over environments/years was observed at JBL18 (9.4 ton/ha),

TABLE 1 | Variability analysis of various yield-related agronomic traits for four environments at two locations.

Loc[#]	Env	Trait[##]	H^2	G Var	R Var	G Mean	LSD	CV	G Sig
JBL	JBL17	DTHD	0.84	10.98	4.04	81.96	3.65	2.45	0
		DAYSMT	0.86	5.86	1.89	124.82	2.52	1.10	0
		GRYLD	0.48	0.29	0.63	7.87	1.08	10.09	0.000151
		TGW	0.86	26.59	8.92	54.66	5.47	5.47	0
	JBL18	DTHD	0.78	12.79	7.30	79.26	4.71	3.41	0
		DAYSMT	0.71	4.89	3.96	124.67	3.32	1.60	9.08E-13
		GRYLD	0.47	0.15	0.34	8.76	0.79	6.67	0.000172
		TGW	0.80	12.53	6.34	46.22	4.45	5.45	0
LDH	LDH17	DTHD	0.96	12.61	1.19	94.85	2.11	1.15	0
		DAYSMT	0.74	4.79	3.29	148.73	3.09	1.22	6.88E-15
		GRYLD	0.74	0.21	0.15	7.06	0.66	5.55	2.73E-14
		TGW	0.81	15.42	7.03	45.48	4.73	5.83	0
	LDH18	DTHD	0.88	8.58	2.44	103.71	2.89	1.51	0
		DAYSMT	0.88	8.18	2.25	144.52	2.80	1.04	0
		GRYLD	0.62	0.16	0.20	7.26	0.69	6.11	1.92E-08
		TGW	0.83	14.66	6.13	44.30	4.47	5.59	0

[#] Loc, location; Env, Environment; H^2, heritability; G Var, genotypic variance; R Var, residual variance; LSD, least significant difference; CV, critical variance; G Sig, genotypic significance; LDH, Ludhiana; JBL, Jabalpur.
[##] DTHD, days to heading; DAYSMT, days to maturity; GRYLD, grain yield; TGW, thousand-grain weight.

TABLE 2 | Variability analysis of various yield-related agronomic traits for four environments at two locations.

Traits	H^2	G Var	G × E Var	R Var	G Mean	LSD	CV	n Rep	n Env	G Sig	G × E Sig
DTHD	0.90	8.94	2.29	3.74	89.94	2.69	2.15	2	4	8.93E-73	1.16E-18
DAYSMT	0.83	4.00	1.94	2.83	135.68	2.32	1.24	2	4	4.34E-44	2.01E-21
GRYLD	0.38	0.05	0.15	0.33	7.74	0.49	7.43	2	4	0.0003	3.69E-13
TGW	0.78	9.90	7.41	7.10	47.67	4.07	5.59	2	4	1.13E-33	4.23E-35

H^2, heritability; G Var, genotypic variance; R Var, residual variance; LSD, least significant difference; CV, critical variance; G Sig, genotypic significance; DTHD, days to heading; DAYSMT, days to maturity; GRYLD, grain yield; TGW, thousand-grain weight.

followed by JBL17 (8.7 ton/ha), LDH17 (8.2 ton/ha), and LDH18 (7.9 ton/ha). Similarly, the TGW trait also showed variation across environments. The highest averaged TGW over environments/years was observed at JBL17 (69 g), followed by JBL18 (59.5 g), LDH17 (58.4 g), and LDH18 (53.5 g). We observed significant G × E interaction for the GRYLD and DAYSMT in JBL18 and LDH17 (**Tables 1**, **2**). For all traits, the broad-sense heritability ranged from 0.47 to 0.96. The broad-sense heritability of DTHD was the highest (0.96) in LDH17, while GRYLD, the lowest (0.47) was in JBL18, and the highest (0.74) was in LDH17. TGW had the highest stability and relatively high heritability (0.80–0.86) across environments.

The phenological traits DTHD and DAYSMT displayed the strongest positive correlation (0.88), followed by a weak positive correlation TGW-GRYLD (0.15), while GRYLD and DTHD (−0.73) demonstrated negative correlations. The lowest correlation was observed between GRYLD and DAYSMT (−0.76) traits. The principal component analysis (PCA) of multivariate analysis enables the easier understanding of effects and networks among different traits and elucidates genotypic difference among a set of given traits, i.e., the first two PCs explained 92% of the total variation. The PC1 explained 70.3% of the total variance and PC2 explained 21.7% of the total variance (**Figure 2**).

Baseline SE Model: Performance of Untested Lines in the Same Environment

A GS scenario representing SE breeding programs was tested. In this model, the PAs of the GS models for each of the four traits were separately generated for all four tested environments. In other words, the environments were treated as independent. Overall, the PA of the SE model was significantly lower among the three tested GS scenarios (**Table 4**; **Figure 3**). PA was the highest for TGW (0.34) and the lowest for GRYLD (0.18) traits. A relatively low moderate PA ranging between 0.24 and 0.25 was observed for DAYSMT and DTHD traits. Among the tested environments, JBL18 had the lowest overall PA (0.01–0.02) compared to the rest of the three environments for DTHD and

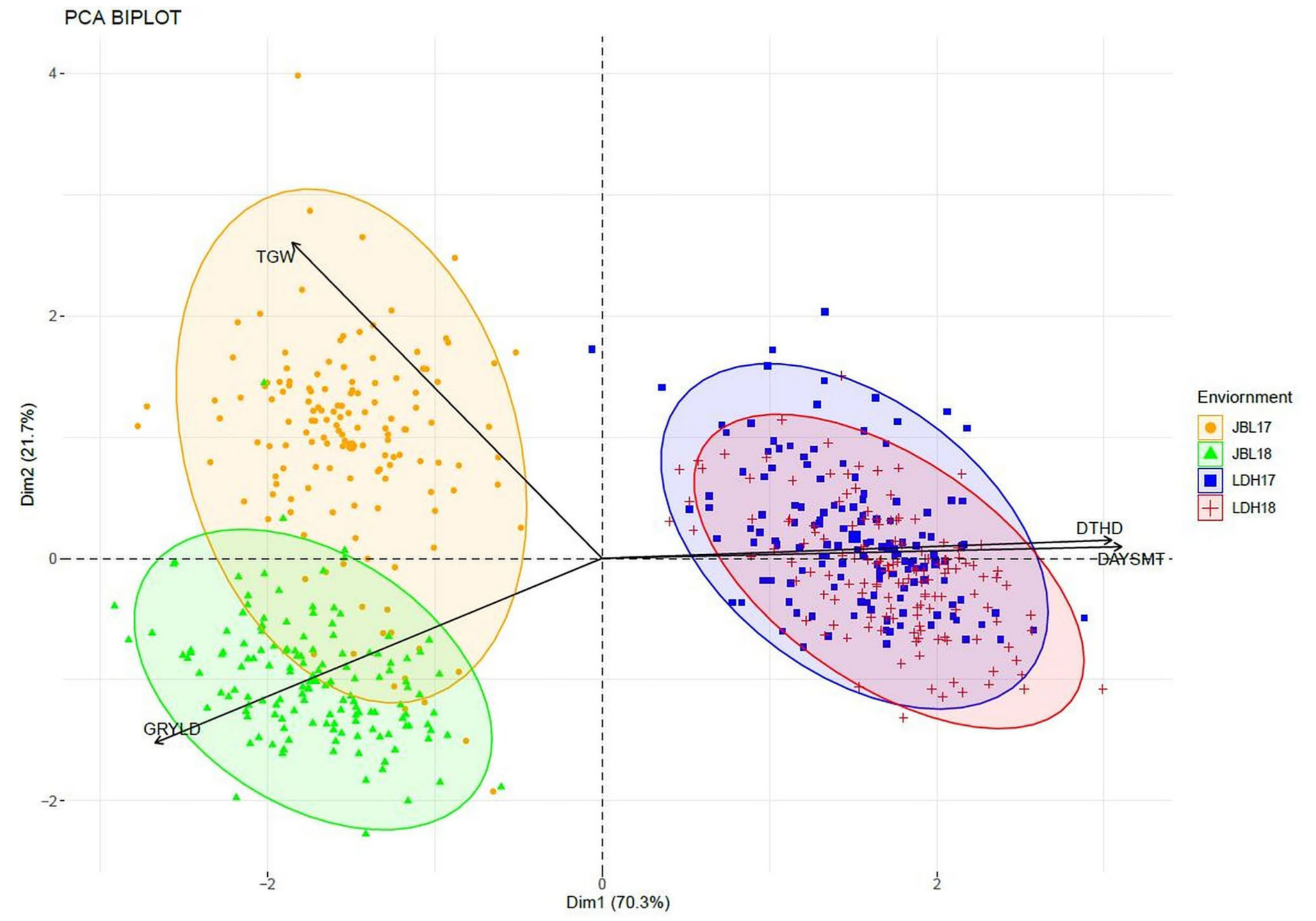


FIGURE 2 | The principal component analysis shows the correlation among GRYLD, TGW, DAYSMT, and DTHD in four environments (LDH17, LDH18, JBL17, and JBL18).

DAYSMT (0.25–0.40). TGW was the only trait where a highly consistent and moderate PA (0.32–0.35) across all environments was observed. PA for GRYLD was the highest for LDH18 (0.32) and the lowest for JBL17 (0.08).

Advanced ME Model: Performance of Tested Lines in Untested Environments and Untested Lines in Tested Environments

The inclusion of environmental information in ME models significantly improved the PA over SE models across all traits and environments (**Figure 3**). A very high and consistent PA ranging from 0.69 to 0.85 was observed for all traits and environments for both ME models (ME_CV1 and ME_CV2). The most considerable improvement in PA due to ME was observed for the GRYLD trait, where PA increased from 0.18 to 0.73 for SE and ME models (**Table 4**). Interestingly, identical trait rankings were also observed for two ME models, where the DTHD ranked the highest (0.85) and GRYLD ranked the lowest (0.69–0.73) among all four traits. While the ME models had identical trait rankings, the environments ranked slightly differently for the two models for all traits. For instance, both years (2017 and 2018) at the LDH location had higher overall PA compared to JBL for all traits.

DISCUSSION

Crop breeders regularly evaluate the performance of genotypes and collect multiple traits data in various environments. The genotype-based selection on phenotypic and GBS marker information using genomic prediction models is gradually acquiring acceptance in breeding with the initiation of economical next-generation sequencing (NGS) technologies (Poland and Rife, 2012). Limited study has been conducted using the multi-environment genomic prediction (ME-GP) methods due to the complexity and higher computing requirements (Oakey et al., 2016; Rincent et al., 2017; Montesinos-López et al., 2018; Roorkiwal et al., 2018; Bhandari et al., 2019; Tolhurst et al., 2019; Pandey et al., 2020).

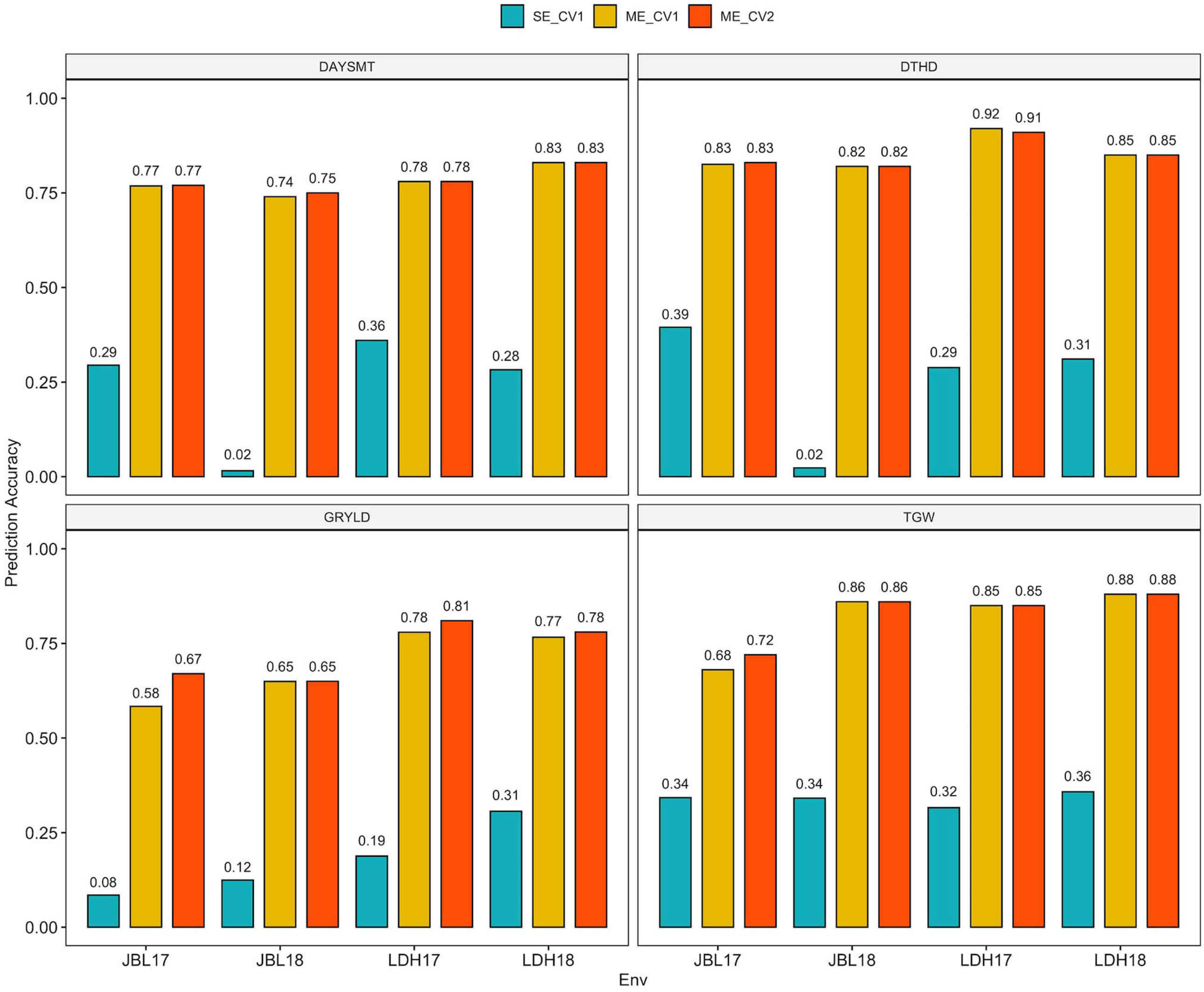


FIGURE 3 | Bar plots showing the prediction accuracy (PA) of DAYSMT, DTHD, GRYLD, and TGW using SE and ME models from individual experiments across locations (LDH17, LDH18, JBL17, and JBL18). SE_CV1 predicting SE at a time, ME_CV1 predicting new lines with genotypic information only, and ME_CV2 predicting partially phenotyped lines by using genotypic and phenotypic information from all traits from individuals in the training set, and genotypic and correlated phenotypic traits in the testing set.

TABLE 3 | Genetic and phenotypic correlations in agronomic important traits.

	Genetic correlations				Phenotypic correlations		
Traits	**DTHD**	**DAYSMT**	**GRYLD**	**Traits**	**DTHD**	**DAYSMT**	**GRYLD**
DAYSMT	0.94			DAYSMT	0.83		
GRYLD	−0.30	−0.29		GRYLD	−0.22	−0.08	
TGW	−0.33	−0.26	0.22	TGW	−0.29	−0.24	0.07

DTHD, days to heading; DAYSMT, days to maturity; GRYLD, grain yield; TGW, thousand-grain weight.

Trait Correlation and Characterization: A Vital Factor for Improving Accuracy in ME-GP

In this study, advanced breeding lines as part of the bread wheat program of CIMMYT were evaluated under irrigated conditions at two locations (JBL and LDH) for 2 years (2017 and 2018) (i.e., four environments). This study evaluated four traits (i.e., DTHD, DAYSMT, GRYLD, and TGW) for use in an ME trait GP model. GRYLD and related traits were positively correlated to each other in two sets (i.e., 1: DAYSMT and DTHD; and 2: GRYLD and TGW) (**Figure 4**). This positive correlation of GRYLD with TGW in this study points out that the GRYLD was mainly distinct by

TABLE 4 | Genomic prediction accuracies averaged across four environments for four traits and three modeling scenarios (a) single-environment CV1 (SE_CV1), (b) multi-environment CV1 (ME_CV1), and (c) multi-environment CV2 (ME_CV2).

Traits	Average prediction accuracy		
	SE_CV1	ME_CV1	ME_CV2
DAYSMT	0.24	0.78	0.78
DTHD	0.25	0.85	0.85
GRYLD	0.18	0.69	0.73
TGW	0.34	0.82	0.83

DTHD, days to heading; DAYSMT, days to maturity; GRYLD, grain yield; TGW, thousand-grain weight.

the TGW factor. The negative relationship between GRYLD and DTHD indicates that the early-headed genotypes play a vital role in the stability of advanced breeding line yield during grain filling and finally affecting the yield component (Sharma and Smith, 1986).

Yield and Related Trait Heritability Difference Among Environments

Our results showed that the heritability of the traits ranged from moderate (i.e., GRYLD) to high (i.e., DAYSMT, DTHD, and TGW). Among the four traits, the phenological traits (i.e., DTHD and DAYSMT) and TGW particularly showed high stable broad-sense heritability ranging from 0.71 to 0.96. It suggests the high quality of the phenotypic measurements and significant predictive potential of the traits. GRYLD, a highly quantitative and environmentally sensitive trait (Maphosa et al., 2014; Würschum et al., 2018), showed considerable fluctuation across environments with JBL environment having relatively lower heritability (0.47–0.48) compared to LDH (0.62–0.74). The variance explained by agronomic traits was significant (**Table 1**) and indicating a large G × E impact on GRYLD resulted in a lower heritability compared to other traits. Hence, lower heritability estimates for GRYLD were expected as numerous genes govern it. The low heritability and yield variances also could be the possible effect of the smaller plot size and lower sowing density (Rode et al., 2011; Sallam et al., 2015; Thorwarth et al., 2017; Bhatta et al., 2018) (**Tables 1**, **2**). The climate in these two environments is considerably different. While the growing season length is relatively shorter in JBL due to the high overall temperature, the LDH location has a moderately colder climate and longer growing season (Mondal et al., 2016). On the one hand, these highly variable environments do underscore a highly challenging phenotypic landscape; it also presents a significant opportunity to leverage the ME trial framework for trait improvement (Lillemo et al., 2005; Braun et al., 2010). The presence of significant genetic and environmental correlations (i.e., positive correlation in TGW and GRYLD, and DAYSMT and DTHD) in our experiments led us to hypothesize that the correlated traits and environmental relationships can be leveraged to improve the selection accuracy through marker-based ME-GS models (**Figure 4**). Therefore, we proceeded with applying the ME model to test this hypothesis on our selected set of lines (**Table 3**).

SE and ME Genomic Prediction Across Years and Sites and ME Model Utilities in Crop Breeding

While weak predictive capability continues to be a major issue in successfully applying GS (Crossa et al., 2013), numerous studies have demonstrated that GS could be beneficial for quantitative traits such as GRYLD with low heritability and also on how GS can be utilized in a breeding program by using even low to moderate GP in early generation selection (Belamkar et al., 2018; Lado et al., 2018; Michel et al., 2018). There are several aspects influencing the PA of GP models. Some of the crucial aspects associated with this study of ME were the genetic relationship between the testing and training sets, the size of the training set, heritability and trait architecture, and correlations among traits and environments (Asoro et al., 2011; Crossa et al., 2013; Heslot et al., 2013; Sallam et al., 2015; Zhang et al., 2015; Duangjit et al., 2016; Lado et al., 2016; Wang et al., 2016; Thorwarth et al., 2017; Akdemir and Isidro-Sánchez, 2019; Olatoye et al., 2020). Even though the size of the population was small in our study, the GP using correlated traits in the ME_CV1 and ME_CV2 schemes had higher PA, indicating that correlated traits up to some extent could balance the impact on the sizes of small population.

Models that leverage E and G × E components have been shown to improve the genomic prediction accuracies for highly quantitative traits such as phenology and GRYLD (Burgueño et al., 2012; Dias et al., 2018). To evaluate the potential of genomic predictions in highly productive but variable environments of JBL and LDH, we simulated three different genomic prediction scenarios representing actual breeding programs. A comparison of single and ME models showed a 2- to 3-fold improvement in model performance for all traits (**Table 4**; **Figure 3**). Among the four traits, GRYLD showed the highest (3.8X) absolute increase in PA from SE to ME models, highlighting the significance of ME modeling in GRYLD predictions. For the SE model, TGW had the most consistent PA across four environments (0.32–0.34), which was in agreement with the highly stable heritability and a lower fraction of G × E observed for this trait (**Table 2**; **Figure 3**). Interestingly, the PA of the two ME models (CV1 and CV2) showed no significant change, suggesting that the ME model was able to predict well the untested environments and lines equally. A model can be highly predictive of untested environments in scenarios where environments are highly correlated (Malosetti et al., 2016; Jarquín et al., 2017), which seems to be the case for our environments as reflected by the low G × E and high heritability (**Table 1**; **Figure 3**). Similarly, a remarkable improvement in the predictive performance of ME_CV1 can be partially attributed to the fact that our sampled set of lines came from the same breeding program and the sample size of 141 lines was relatively moderate. From the perspective of a breeding program, the strong performance of the two ME models suggests that our breeding program can increase the overall population size without losing any significant predictive power through sparse testing at these two environments (Cullis et al.,

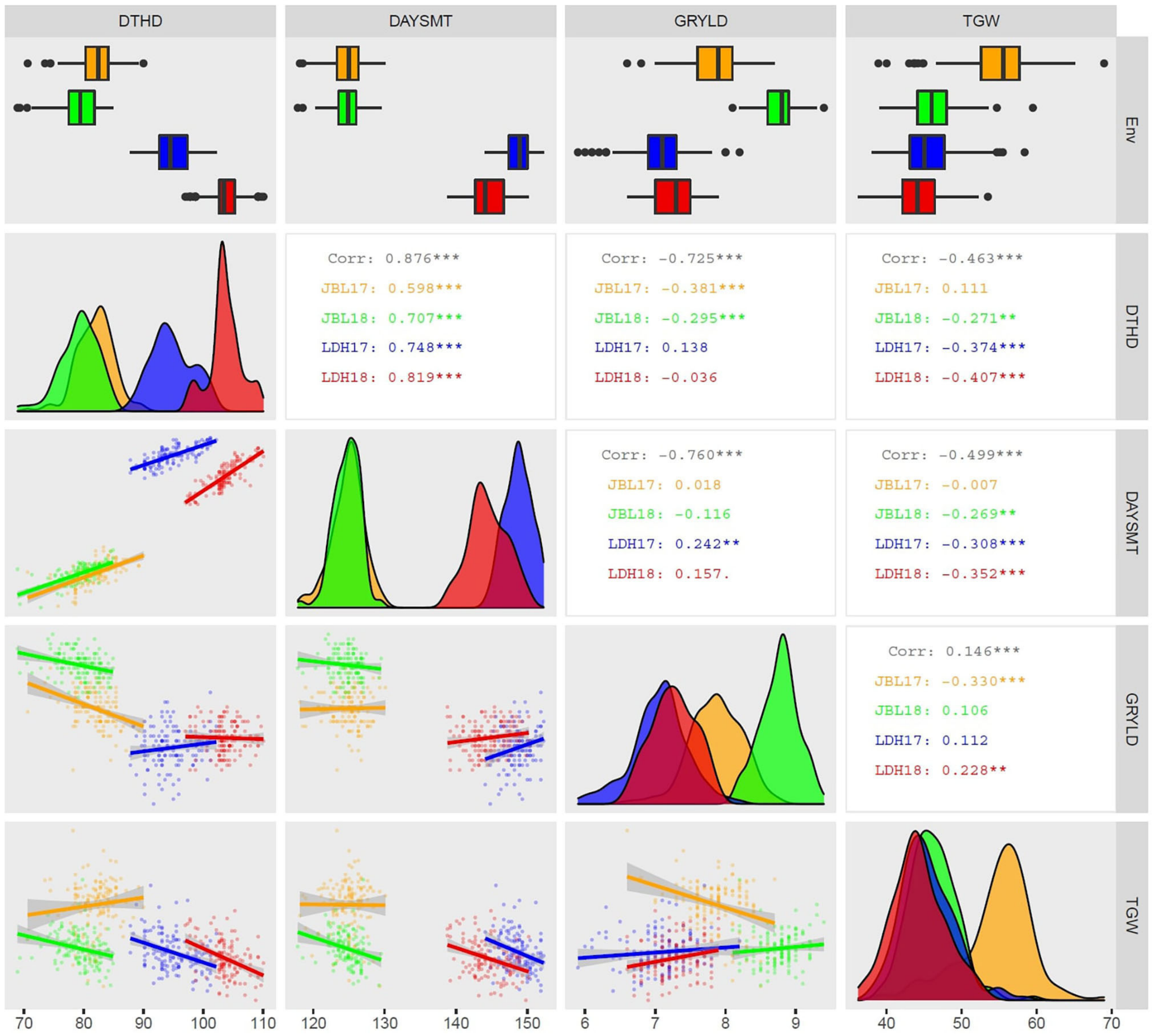


FIGURE 4 | Distributions, scatter plots, and correlations between agronomic traits using best linear unbiased predictions from combining and four experiments [Ludhiana (LDH)17, LDH18, Jabalpur (JBL)17, and JBL18]. The distribution of DTHD, DAYSMT, GRYLD, and TGW values is displayed on the diagonal with environments indicated by colors. The top row represents the distribution of traits as boxplots. The upper right triangle shows pairwise correlation values as overall correlation in black color while other colors are represented individually as explained earlier. The correlations among environments are displayed as scatter plots in the lower triangular area and as the Pearson's correlation coefficients in the upper triangular area. Numbers indicate a correlation that is significantly different from 0 at an alpha level of 0.05. DTHD, DAYSM, GRYLD, and TGW. ## level of significance; ***p < 0.001, **p < 0.01, *p < 0.1, and p <0.15.

2020; Jarquin et al., 2020). A high population size from the sparse testing framework here can deliver a high selection gain through increased selection intensity.

CONCLUSION

Breeding for quantitative traits is challenging due to the complex genetic architecture of traits that are highly affected by the complex G × E interactions in field trials. A suitable genomic prediction modeling strategy can potentially address this challenge through ME genomic prediction models. In this study, we evaluated genomic prediction accuracies of advanced spring wheat lines under four diverse environments in two wheat-growing regions in India. The ME-GS models showed significant improvement over SE models in terms of prediction accuracies. Our results suggest that ME can be leveraged to improve the breeding selection efficiency for major agronomic and phonological traits. Over the years, CIMMYT has established an extensive network of field-testing sites in South Asian countries including India, Pakistan, Bangladesh, and Nepal. Our results suggest that the wheat breeding programs in these countries can greatly benefit from GS through better modeling of environmental variance and sparse testing of a larger cohort of breeding lines. Future research efforts will be directed toward including high-throughput phenotyping traits such as plant height, Normalized Difference Vegetation Index (NDVI), and senescence into the genomic prediction framework to improve the selection efficiency of spring wheat in the South Asian breeding programs.

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**.

AUTHOR CONTRIBUTIONS

VT and DS drafted the manuscript. VT, DS, GD, and YC analyzed the data. UK, JP, and RS designed the field trials, conducted genotyping, and provided breeding lines. VT and YG collected field data. UK, BT, JP, RS, and AJ supervised the overall study. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the United States Agency for International Development (USAID), Feed the Future Innovation Lab for Applied Wheat Genomics (Cooperative Agreement No. AID-OAA-A-13-00051), and CGIAR Research Program on Wheat (CRP) Partner Grant to BISA (Grant Code: A5017.09.64).

ACKNOWLEDGMENTS

We would like to thank the field staff at the field sites of the Borlaug Institute of South Asia at Jabalpur and Ludhiana for their assistance with the data collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.720123/full#supplementary-material

Supplementary Figure 1 | Weather information of LDH17 and LDH18.

Supplementary Figure 2 | Weather information of JBL17 and JBL18.

Supplementary Table 1 | List of 141 genotypes with pedigree information used in this study.

Supplementary Table 2 | List of traits that were evaluated during this study in the field trials.

Supplementary Table 3 | GBS HapMap data used in this study.

Supplementary Table 4 | Best linear unbiased predictions (BLUPs) data used in this study.

REFERENCES

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

Alvarado, G., Rodríguez, F. M., Pacheco, A., Burgueño, J., Crossa, J., Vargas, M., et al. (2020). META-R: a software to analyze data from multi-environment plant breeding trials. *Crop J.* 8, 745–756. doi: 10.1016/j.cj.2020.03.010

Arzani, A., and Ashraf, M. (2017). Cultivated ancient wheats (*triticum spp.*): a potential source of health-beneficial food products. *Compr. Rev. Food Sci. Food Saf.* 16, 477–488. doi: 10.1111/1541-4337.12262

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., and Jannink, J. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite north american oats. *Plant Genom.* 4:007. doi: 10.3835/plantgenome2011.02.0007

Belamkar, V., Guttieri, M. J., Hussain, W., Jarquín, D., El-basyoni, I., Poland, J., et al. (2018). Genomic selection in preliminary yield trials in a winter wheat breeding program. *G3 Genes, Genomes, Genet.* 8, 2735–2747. doi: 10.1534/g3.118.200415

Bhandari, A., Bartholom,é, J., Cao-Hamadoun, T.-V., Kumari, N., Frouin, J., Kumar, A., et al. (2019). Selection of trait-specific markers and multi-environment models improve genomic predictive ability in rice. *PLoS ONE* 14:e0208871. doi: 10.1371/journal.pone.0208871

Bhatta, M., Morgounov, A., Belamkar, V., and Baenziger, P. S. (2018). Genome-Wide association study reveals novel genomic regions for grain yield and yield-related traits in drought-stressed synthetic hexaploid wheat. *Int. J. Mol. Sci.* 19:3011. doi: 10.3390/ijms19103011

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, 2633–2635. doi: 10.1093/bioinformatics/btm308

Braun, H. J., Atlin, G., and Payne, T. (2010). "Multi-location testing as a tool to identify plant response to global climate change," in *Climate Change and Crop Production*, ed M. P. Reynolds (CABI International), 115–138.

Burgueño, J., Campos, G., de los, 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

Charmet, G., Storlie, E., Oury, F. X., Laurent, V., Beghin, D., Chevarin, L., et al. (2014). Genome-wide prediction of three important traits in bread wheat. *Mol. Breed.* 34, 1843–1852. doi: 10.1007/s11032-014-0143-y

Charmet, G., Tran, L.-G., Auzanneau, J., Rincent, R., and Bouchet, S. (2020). BWGS: A R package for genomic selection and its application to a wheat breeding programme. *PLoS ONE* 15:e0222733. doi: 10.1371/journal.pone.0222733

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

Crossa, J., Campos, G., de los, 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., 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|Genetics* 6:1819. doi: 10.1534/g3.116.029637

Crossa, J., Pérez, P., Campos, G., de los, Mahuku, G., Dreisigacker, S., and Magorokosho, C. (2011). Genomic selection and prediction in plant breeding. *J. Crop Improv.* 25, 239–261. doi: 10.1080/15427528.2011.558767

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., et al. (2013). Genomic prediction in CIMMYT maize and wheat breeding programs. *Hered* 112, 48–60. doi: 10.1038/hdy.2013.16

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., Cocks, N. A., and Butler, D. G. (2020). The design of early-stage plant breeding trials using genetic relatedness. *J. Agric. Biol. Environ. Stat.* 25, 553–578. doi: 10.1007/s13253-020-00403-5

Curtis, T., and Halford, N. G. (2014). Food security: the challenge of increasing wheat yield and the importance of not compromising food safety. *Ann. Appl. Biol.* 164, 354–372. doi: 10.1111/aab.12108

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

Dias, K. O. D. G., Gezan, S. A., Guimarães, C. T., Nazarian, A., da Costa e 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. *Hered* 121, 24–37. doi: 10.1038/s41437-018-0053-6

Dreisigacker, S., Deepmala, S., Jaimez, R. A., Luna-Garrid, B., Muñoz-Zavala, S., Núñez-Ríos, C., et al. (2016). *CIMMYT Wheat Molecular Genetics: Laboratory Protocols and Applications to Wheat Breeding.* Mexico, DF: CIMMYT.

Duangjit, J., Causse, M., and Sauvage, C. (2016). Efficiency of genomic selection for tomato fruit quality. *Mol. Breed.* 36, 1–16. doi: 10.1007/s11032-016-0453-3

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* 11:170104. doi: 10.3835/plantgenome2017.11.0104

Hayes, B. J., Panozzo, J., Walker, C. K., Choy, A. L., Kant, S., Wong, D., et al. (2017). Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theor. Appl. Genet.* 130, 2505–2519. doi: 10.1007/s00122-017-2972-7

Heffner, E. L., Jannink, J.-L., Iwata, H., Souza, E., and Sorrells, M. E. (2011). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci.* 51, 2597–2606. doi: 10.2135/cropsci2011.05.0253

Hellin, J., Shiferaw, B., Cairns, J. E., Reynolds, M., Ortiz-Monasterio, I., Banziger, M., et al. (2012). Climate change and food security in the developing world: Potential of maize and wheat research to expand options for adaptation and mitigation. *J. Dev. Agric. Econ.* 4, 311–321. doi: 10.5897/JDAE11.112

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J. L. (2013). 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

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

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genomics.* 9, 166–177. doi: 10.1093/bfgp/elq001

Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., et al. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3 Genes| Genomes|Genet.* 10, 2725–2739. doi: 10.1534/g3.120.401349

Jarquín, D., Silva, C. L., da, 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* 10:0130. doi: 10.3835/plantgenome2016.12.0130

Juliana, P., Montesinos-López, O. A., Crossa, J., Mondal, S., González Pérez, L., Poland, J., et al. (2019). Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. *Theor. Appl. Genet.* 132, 177–194. doi: 10.1007/s00122-018-3206-3

Juliana, P., Singh, R. P., Singh, P. K., Crossa, J., Huerta-Espino, J., Lan, C., et al. (2017a). Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theor. Appl. Genet.* 130, 1415–1430. doi: 10.1007/s00122-017-2897-1

Juliana, P., Singh, R. P., Singh, P. K., Crossa, J., Rutkoski, J. E., Poland, J. A., et al. (2017b). Comparison of models and whole-genome profiling approaches for genomic-enabled prediction of *Septoria Tritici Blotch*, *Stagonospora Nodorum Blotch*, and Tan Spot resistance in wheat. *Plant Genome* 10:0082. doi: 10.3835/plantgenome2016.08.0082

Kassambara, A., and Mundt, F. (2020). *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. Available online at: https://cran.r-project.org/packagefactoextra (accessed May 05, 2020).

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of IJCAI'95* 2, 1137–1143.

Lado, B., Barrios, P. G., Quincke, M., Silva, P., and Gutiérrez, L. (2016). Modeling genotype × environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Sci.* 56, 2165–2179. doi: 10.2135/cropsci2015.04.0207

Lado, B., Vázquez, D., Quincke, M., Silva, P., Aguilar, I., and Gutiérrez, L. (2018). Resource allocation optimization with multi-trait genomic prediction for bread wheat (*Triticum aestivum* L.) baking quality. *Theor. Appl. Genet.* 131, 2719–2731. doi: 10.1007/s,00122-018-3186-3

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* 25, 1–18. doi: 10.18637/jss.v025.i01

Lillemo, M., Ginkel, M., van, Trethowan, R. M., Hernandez, E., and Crossa, J. (2005). Differential adaptation of CIMMYT bread wheat to global high temperature environments. *Crop Sci.* 45, 2443–2453. doi: 10.2135/cropsci2004.0663

Lozada, D. N., Mason, R. E., Sarinelli, J. M., and Brown-Guedira, G. (2019). Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. *BMC Genet.* 20, 1–12. doi: 10.1186/s12863-019-0785-1

Malosetti, M., Bustos-Korts, D., Boer, M. P., and Eeuwijk, F. A., van (2016). Predicting responses in multiple environments: issues in relation to genotype × environment interactions. *Crop Sci.* 56, 2210–2222. doi: 10.2135/cropsci2015.05.0311

Maphosa, L., Langridge, P., Taylor, H., Parent, B., Emebiri, L. C., Kuchel, H., et al. (2014). Genetic control of grain yield and grain physical characteristics in a bread wheat population grown under a range of environmental conditions. *Theor. Appl. Genet. 7* 127, 1607–1624. doi: 10.1007/s00122-014-2322-y

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. 4 1819–1829. doi: 10.1093/genetics/157.4.1819

Michel, S., Kummer, C., Gallee, M., Hellinger, J., Ametz, C., Akgöl, B., et al. (2018). Improving the baking quality of bread wheat by genomic selection in early generations. *Theor. Appl. Genet.* 131, 477–493. doi: 10.1007/s00122-017-2998-x

Mondal, S., Singh, R. P., Mason, E. R., Huerta-Espino, J., Autrique, E., and Joshi, A. K. (2016). Grain yield, adaptation and progress in breeding for early-maturing and heat-tolerant wheat lines in South Asia. *F. Crop. Res.* 192, 78–85. doi: 10.1016/j.fcr.2016.04.017

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3 Genes|Genomes|Genet.* 8, 3813–3828. doi: 10.1534/g3.118.200740

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|Genetics* 8, 2889–2899. doi: 10.1534/g3.118.200311

Norman, A., Taylor, J., Tanaka, E., Telfer, P., Edwards, J., Martinant, J.-P., et al. (2017). Increased genomic prediction accuracy in wheat breeding using a large Australian panel. *Theor. Appl. Genet.* 130, 2543–2555. doi: 10.1007/s00122-017-2975-4

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

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 Genes|Genomes|Genet.* 10, 2465–2476. doi: 10.1534/g3.120.401402

Pandey, M. K., Chaudhari, S., Jarquin, D., Janila, P., Crossa, J., Patil, S. C., et al. (2020). Genome-based trait prediction in multi- environment breeding trials in groundnut. *Theor. Appl. Genet.* 133, 3101–3117. doi: 10.1007/s00122-020-03658-1

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., and Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 Genes|Genomes|Genetics* 2, 1595–1605. doi: 10.1534/g3.112.003665

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme Genotyping-by-Sequencing approach. *PLoS ONE* 7:e32253. doi: 10.1371/journal.pone.0032253

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-Sequencing for plant breeding and genetics. *Plant Genome* 5:005. doi: 10.3835/plantgenome,2012.05.0005

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Found. Stat. Comput. Avaialble online at: https://www.R-project.org/

Rincent, R., Kuhn, E., Monod, H., Oury, F.-X., Rousset, M., Allard, V., et al. (2017). Optimization of multi-environment trials for genomic selection based on crop models. *Theor. Appl. Genet.* 130, 1735–1752. doi: 10.1007/s00122-017-29 22-4

Rode, J., Ahlemeyer, J., Friedt, W., and Ordon, F. (2011). Identification of marker-trait associations in the German winter barley breeding gene pool (*Hordeum vulgare* L.). *Mol. Breed.* 30, 831–843. doi: 10.1007/s11032-011-9667-6

Roorkiwal, M., Jarquin, D., Singh, M. K., Gaur, P. M., Bharadwaj, C., Rathore, A., et al. (2018). Genomic-enabled prediction models using multi-environment trials to estimate the effect of genotype × environment interaction on prediction accuracy in chickpea. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-018-30027-2

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* 8:74. doi: 10.3835/plantgenome2014.10.0074

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* 8:20. doi: 10.3835/plantgenome2014.05.0020

Sehgal, D., Autrique, E., Singh, R., Ellis, M., Singh, S., and Dreisigacker, S. (2017). Identification of genomic regions for grain yield and yield stability and their epistatic interactions. *Sci. Rep.* 7, 1–12. doi: 10.1038/srep41578

Sharma, R. C., and Smith, E. L. (1986). Selection for high and low harvest index in three winter wheat populations1. *Crop Sci.* 26, 1147–1150. doi: 10.2135/cropsci1986.0011183X002600060013x

Thorwarth, P., Ahlemeyer, J., Bochard, A.-M., Krumnacker, K., Blümel, H., Laubach, E., et al. (2017). Genomic prediction ability for yield-related traits in German winter barley elite material. *Theor. Appl. Genet.* 130, 1669–1683. doi: 10.1007/s00122-017-2917-1

Tolhurst, D. J., Mathews, K. L., Smith, A. B., and Cullis, B. R. (2019). Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model. *J. Anim. Breed. Genet.* 136, 279–300. doi: 10.1111/jbg.12404

Tomar, V., Dhillon, G. S., Singh, D., Singh, R. P., Poland, J., Chaudhary A. A., et al. (2021). Evaluations of genomic prediction and identification of new loci for resistance to stripe rust disease in wheat (Triticum aestivum L.) *Front. Genet.* 12:710485 (in press). doi: 10.3389/fgene.2021.710485

Velu, G., Crossa, J., Singh, R. P., Hao, Y., Dreisigacker, S., Perez-Rodriguez, P., et al. (2016). Genomic prediction for grain zinc and iron concentrations in spring wheat. *Theor. Appl. Genet.* 129, 1595–1605. doi: 10.1007/s00122-016-2726-y

Wang, X., Li, L., Yang, Z., Zheng, X., Yu, S., Xu, C., et al. (2016). Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Hered* 118, 302–310. doi: 10.1038/hdy.2016.87

Würschum, T., Leiser, W. L., Langer, S. M., Tucker, M. R., and Longin, C. F. H. (2018). Phenotypic and genetic analysis of spike and kernel characteristics in wheat reveals long-term genetic trends of grain yield components. *Theor. Appl. Genet.* 131, 2071–2084. doi: 10.1007/s00122-018-3133-3

Zhang, J., Song, Q., Cregan, P. B., and Jiang, G.-L. (2015). Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycine max). *Theor. Appl. Genet.* 129, 117–130. doi: 10.1007/s00122-015-2614-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.

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.

ORIGINAL RESEARCH
published: 11 November 2021
doi: 10.3389/fpls.2021.699589

Prediction of Maize Phenotypic Traits With Genomic and Environmental Predictors Using Gradient Boosting Frameworks

Cathy C. Westhues[1,2]*, *Gregory S. Mahone*[3], *Sofia da Silva*[3], *Patrick Thorwarth*[3], *Malthe Schmidt*[3], *Jan-Christoph Richter*[3], *Henner Simianer*[2,4] and *Timothy M. Beissinger*[1,2]

[1] Division of Plant Breeding Methodology, Department of Crop Sciences, University of Goettingen, Goettingen, Germany, [2] Center for Integrated Breeding Research, University of Goettingen, Goettingen, Germany, [3] Kleinwanzlebener Saatzucht (KWS) SAAT SE, Einbeck, Germany, [4] Animal Breeding and Genetics Group, Department of Animal Sciences, University of Goettingen, Goettingen, Germany

OPEN ACCESS

Edited by:
Fernando H. Toledo,
International Maize and Wheat
Improvement Center, Mexico

Reviewed by:
Diego Jarquin,
University of Nebraska-Lincoln,
United States
Germano Costa-Neto,
Cornell University, United States
Sergio Perez-Elizalde,
Colegio de Postgraduados
(COLPOS), Mexico

***Correspondence:**
Cathy C. Westhues
cathy.jubin@uni-goettingen.de

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: 23 April 2021
Accepted: 15 October 2021
Published: 11 November 2021

Citation:
Westhues CC, Mahone GS, da Silva S, Thorwarth P, Schmidt M, Richter J-C, Simianer H and Beissinger TM (2021) Prediction of Maize Phenotypic Traits With Genomic and Environmental Predictors Using Gradient Boosting Frameworks. Front. Plant Sci. 12:699589. doi: 10.3389/fpls.2021.699589

The development of crop varieties with stable performance in future environmental conditions represents a critical challenge in the context of climate change. Environmental data collected at the field level, such as soil and climatic information, can be relevant to improve predictive ability in genomic prediction models by describing more precisely genotype-by-environment interactions, which represent a key component of the phenotypic response for complex crop agronomic traits. Modern predictive modeling approaches can efficiently handle various data types and are able to capture complex nonlinear relationships in large datasets. In particular, machine learning techniques have gained substantial interest in recent years. Here we examined the predictive ability of machine learning-based models for two phenotypic traits in maize using data collected by the Maize Genomes to Fields (G2F) Initiative. The data we analyzed consisted of multi-environment trials (METs) dispersed across the United States and Canada from 2014 to 2017. An assortment of soil- and weather-related variables was derived and used in prediction models alongside genotypic data. Linear random effects models were compared to a linear regularized regression method (*elastic net*) and to two nonlinear gradient boosting methods based on decision tree algorithms (*XGBoost*, *LightGBM*). These models were evaluated under four prediction problems: (1) tested and new genotypes in a new year; (2) only unobserved genotypes in a new year; (3) tested and new genotypes in a new site; (4) only unobserved genotypes in a new site. Accuracy in forecasting grain yield performance of new genotypes in a new year was improved by up to 20% over the baseline model by including environmental predictors with gradient boosting methods. For plant height, an enhancement of predictive ability could neither be observed by using machine learning-based methods nor by using detailed environmental information. An investigation of key environmental factors using gradient boosting frameworks also revealed that temperature at flowering stage, frequency and amount of water received during the vegetative and grain filling stage, and soil organic matter content appeared as important predictors for grain yield in our panel of environments.

Keywords: machine learning, genotype-by-environment interactions, gradient boosting, maize, yield, genomic prediction, plant breeding

1. INTRODUCTION

The development of environmental sensing technologies, including local weather stations, soil and crop sensors has progressively enabled field-level climate data to be incorporated into the analysis of plant breeding experiments (Tardieu et al., 2017; Ersoz et al., 2020; Crossa et al., 2021). When used to enhance genomic prediction, climate data can be useful to estimate the differential response of genotypes to new environmental conditions, i.e., genotype-by-environment interactions (GxE), almost omnipresent in multi-environment trial (MET) experiments (Cooper and DeLacy, 1994; Chenu, 2015). In plant breeding, an environment generally refers to the set of growing conditions associated with a given location in a given year. Various statistical models, such as factorial regression methods, have been developed to model genotype sensitivity to continuous environmental covariates (ECs) (van Eeuwijk et al., 1996; Malosetti et al., 2004) or even to simple geographic coordinates (Costa-Neto et al., 2020b) capturing primarily genotype-by-location interaction effects explained by crop management or soil characteristics.

Before the emergence of environmental data in breeding, large whole-genome marker datasets, generated by high-throughput genotyping platforms, have progressively enabled the routine implementation of genomic prediction (GP) methods (Haley and Visscher, 1998; Meuwissen et al., 2001). GP allows to predict performance of untested genotypes based on their genetic similarity, estimated with marker data, with other phenotyped genotypes. GP has since been expanded to achieve predictions in a multi-environment context, for instance by implementing a multivariate GBLUP approach (Burgueño et al., 2012) to use genetic correlations among environments. Despite the overall success of genomic prediction, a lingering challenge has regularly been to incorporate interactions between high-dimensional genomic data and high-dimensional environmental data. A solution proposed by Jarquín et al. (2014) is to use reaction norm models, where markers and environmental effects are modeled using covariance structures. Interactions between markers and environmental covariates are computed with the Hadamard product which avoids the need to fit all first-order interaction terms. This extension of the GBLUP GxE mixed effects models has been applied on a large number of datasets in different species (Pérez-Rodríguez et al., 2015; Pérez-Rodríguez et al., 2017; Jarquín et al., 2017; Sukumaran et al., 2017, 2018; Monteverde et al., 2019; Rincent et al., 2019; De Los Campos et al., 2020). Several studies have also focused on the integration of crop growth models in genomic prediction to better model the differential impact of abiotic stress depending on the crop developmental stage (Heslot et al., 2014a; Rincent et al., 2017, 2019). Rincent et al. (2019) proposed a method to select the optimal subset of ECs from the output of a crop growth model on the basis of the correlation between the environmental covariance matrix, which is based on ECs, and the covariance matrix between GxE interactivity of environments obtained by AMMI decomposition. Overall, many studies have found that using quantitative environmental information in genomic prediction models in the form of additional covariates can result in an enhancement of prediction accuracies (Heslot et al., 2014b; Jarquín et al., 2014; Malosetti et al., 2016; Millet et al., 2019; Monteverde et al., 2019; Costa-Neto et al., 2020a) and a better characterization of the genotype-by-environment interaction effects (Rogers et al., 2021).

However, modeling interaction effects with nonlinear techniques is a crucial topic that has not been conclusively explored for genomic prediction in MET. In particular, machine learning techniques have gained attention over the last two decades due to their ability to handle nonlinear effects (Hastie et al., 2009) and to uncover higher-order interactions between predictor variables (Lampa et al., 2014; Behravan et al., 2018). With machine learning algorithms, the mapping function linking input variables to the outcome—i.e., a phenotypic trait—is learned from training data and no strong assumptions about its form need to be explicitly formulated beforehand. Hence, these methods represent relatively flexible frameworks for data-driven integration of different data types. Among these new techniques, ensembles of trees, such as methods based on bagging (e.g., random forests), or on boosting (e.g., gradient boosted trees) have become increasingly popular. Ensemble methods designate predictive modeling techniques which aggregate the predictions of a group of base learners, and thereby generally allow better predictions than by using only the single best learner (Friedman, 2001; Hastie et al., 2009; Géron, 2019). Broad applications of these approaches include human disease prediction (Fukuda et al., 2013; Romagnoni et al., 2019; Yu et al., 2019; Kopitar et al., 2020), bioinformatics (Yu et al., 2019), ecology (Moisen et al., 2006; Elith et al., 2008) and agricultural forecasting (Fukuda et al., 2013; Delerce et al., 2016; Jeong et al., 2016; Crane-Droesch, 2018; Shahhosseini et al., 2020). In the field of genomic prediction, ensemble methods have progressively been used, as they appear especially interesting for capturing non-additive effects such as epistasis or dominance effects, which can be important for predicting complex phenotypic traits (Ogutu et al., 2011; González-Recio et al., 2013; Azodi et al., 2019; Abdollahi-Arpanahi et al., 2020). Abdollahi-Arpanahi et al. (2020) concluded from results obtained on both a real animal and simulated datasets that gradient boosting was the best predictive modeling approach when the genetic architecture included non-additive effects. While these new predictive modeling approaches can also potentially enable superior prediction results, special attention must be paid to an appropriate optimization of hyperparameters during the training phase in order to prevent overfitting on new test data (Friedman, 2001; Hastie et al., 2009; Géron, 2019).

In addition, these new predictive modeling frameworks, coupled with large volumes of environmental data, can provide powerful data mining opportunities to identify critical environmental factors affecting economically important phenotypic traits in the field. Much research has already been done to examine the expected impact of climate change on the vulnerability of major staple food crops. Extreme weather events are expected to happen at a higher frequency in the future, characterized for instance by heat waves or prolonged drought periods according to various climate scenarios (Rahmstorf et al., 2012; Trnka et al., 2014). When occurring at crucial crop developmental stages, risks for important yield losses are augmented. Different studies on maize have for instance reported a physiological sensitivity to higher temperatures, heightened during the reproductive phase, which often results in grain yield reduction when a certain threshold is exceeded (Cicchino et al., 2010; Butler and Huybers, 2015; Lizaso et al., 2018). In addition, nonlinear effects of environmental covariates, especially of temperature and precipitation on maize plants, have also been regularly described in the literature (Schlenker and Roberts, 2009; Mushore et al., 2017). Therefore, machine learning techniques break new ground to get an extended comprehension of the effect—both in direction and magnitude—of environmental conditions in the context of breeding for abiotic stress resilience.

Motivated by previous studies emphasizing the benefit of nonlinear methods, we tested two machine learning ensemble methods, based on gradient boosted trees, which, to our knowledge, have never been examined for data-driven predictions and interpretation using MET experimental datasets from the Maize Genomes to Fields initiative. The Maize Genomes to Fields (G2F) initiative (www.genomes2fields.org) includes yearly evaluations of inbred and hybrid maize across a large range of climatically-distinct regions in North America. The project makes publicly available phenotypic and genotypic (genotyping-by-sequencing datasets relating to the inbred lines) information, as well as weather (field weather stations), agronomic practices and soil data (Falcon et al., 2020; McFarland et al., 2020). The large number of phenotypic observations, and the assortment of various data types makes the application of machine learning models here particularly relevant to evaluate their performance, as well as their usefulness to disentangle hidden relationships. Our objectives in this study were (1) to evaluate recent gradient boosting methods for prediction of two phenotypic traits (plant height and grain yield) across four different cross-validations, and compare them to traditional prediction models classically used for multi-environment trials; (2) to examine if the use of environmental information, in addition to genomic predictor variables, could lead to a gain of predictive ability of genotype performance based on these various prediction models; and (3) to better understand the influence of some environmental factors on maize grain yield using tools derived from the machine learning framework.

2. MATERIALS AND METHODS

2.1. Phenotypic Data Cleaning and Analysis

Phenotypic datasets (years 2014–2017) were downloaded from the official website of the Genomes to Fields project. The full dataset represents a large collection of trials located on the North-American continent run by different principal investigators and institutions, but the experimental design used for most of the hybrid trials was a randomized complete block design with two replications per environment. A total number of 71 trial experiments remained for further analysis (**Supplementary Figure 1**; **Supplementary Table 1**) after having eliminated environments with critical missing information, such as flowering time (**Supplementary Table 2**). Plots with low phenotypic quality, as interpreted by the researcher groups who collected field data, were removed before within-experiment analysis. Replicates within a same ID experiment but planted seven or more days apart were considered as different environments and treated as unreplicated environments, due to the difference in the weather conditions they experienced at their respective phenological stages.

Each environment (Year-Site combination) was independently analyzed to obtain best linear unbiased estimates (BLUEs) for each hybrid in each environment for grain yield, plant height and silking date. We performed this analysis with the *lme4* package (Bates et al., 2015) in R version 3.6.0 (R Core Team, 2019) based on the following model:

$$Y_{ij} = \mu + G_i + R_j + \varepsilon_{ij},$$

where Y_{ij} is the observed phenotypic response variable of the i-th hybrid genotype (G) in the j-th replicate (R), μ is the general mean, G_i is the effect of the i-th hybrid genotype, R_j is the effect of the j-th replicate and ε_{ij} is the error associated with the observation Y_{ij}. We treated genotype as a fixed effect and replicate as a random effect.

Phenotypic observations with absolute studentized conditional residuals greater than three were identified as potential outliers and removed from the dataset. The plant material and phenotypic datasets are described in more details in previous publications (AlKhalifah et al., 2018; McFarland et al., 2020) and on the project website (https://www.genomes2fields.org/home/). Ultimately, 18,325 and 16,951 phenotypic observations for grain yield and plant height, respectively, with available silking date, genotypic and environmental data, were used as target response variable in the prediction models.

2.2. Genotypic Data

Genotype-by-sequencing (GBS) data of inbred lines used in Genomes to Fields hybrid experiments were downloaded on CyVerse. SNPs with more than two observed alleles were removed before analysis. Taxa with less than 70% site coverage and more than 8% heterozygosity were discarded. Monomorphic markers were removed, as were those missing or heterozygous in more than 5% of the parental lines. These filtering analyses were performed with TASSEL 5 (Bradbury et al., 2007). After filtering, 246,818 SNPs remained for analysis. These were imputed using the software LinkImpute (Money et al., 2015). The genotype matrix was coded as the number of minor alleles at each locus (0, 1, or 2). Markers with minor allele frequency less than 2% and in high linkage Disequilibrium (LD) were further removed using the pruning function of Plink (Purcell et al., 2007) with a window of size 100 markers, a step of 5, and a LD threshold of 0.99. *In silico* genotypes of maize hybrids, for which phenotypic data had been analyzed, were constructed based on the processed genotypes of parental lines, and a final minor allele frequency filtering of 2% was applied. The final hybrid genotype dataset contained 107,399 SNPs characterizing 2,033 hybrids. Additional details regarding the genotype-by-sequencing procedure implemented by the Genomes to Fields project has been previously published (Gage et al., 2017).

2.3. Weather Data

All field experiment locations in the Genomes to Fields project had a WatchdogTM Model 2700 weather station (Spectrum Technologies Inc., East-Plainfield, Illinois, 60585, USA) on-site. Weather records were recorded every 30 min during the growing season. Measurements for air temperature (°C), relative humidity (%), rainfall (mm), solar radiation (W/m2) and wind speed (m/s) were specifically analyzed. In-field weather station measurements provide climatic information of a very localized scale in comparison to weather service stations. Therefore, we prioritized the use of weather-station data whenever data quality criteria were fulfilled and the proportion of missing data was reasonable. When quality criteria were not met, weather data was acquired from nearby weather service stations.

In the first step, we summarized the hourly or semi-hourly records for each climatic variable on a daily basis using various quality control criteria (consistent number of weather records per day; threshold tests; persistence tests, i.e., flagging observations with null variability during the day; internal consistency tests, i.e., verification of the relation between measured variables). These criteria were applied based on the recommendations from the official published guidelines on quality control procedures for data acquired from weather stations (Zahumenský, 2004; Estévez et al., 2011) and are detailed in **Supplementary Table 3**. Data from the field weather station were compared against weather data obtained from public climate summaries to check for possible large data divergences and to fill out missing values. Data from the Global Historical Climatology Network (GHCN) and from the Global Surface Summary of the Day (GSOD) were retrieved from the National Oceanic and Atmospheric Administration (NOAA) website to investigate American locations, while data for Canadian locations were downloaded from the Environment and Climate Change Canada (ECCC) website, based each time on a 70-kilometer radius from the geographic coordinates for each field experiment. In case data from the field weather station data were missing or assigned as erroneous, data from the closest publicly accessible weather station were used, if it was located less than 2 km from the field. If the distance to the nearest station was large, interpolation by spatio-temporal kriging or inverse distance weighting was performed using the R package *gstat* to impute the missing data (Pebesma, 2004; Gräler et al., 2016). For wind data, we only used results obtained from inverse distance weighting because of the consistency regarding the standard height measurement obtained from GSOD data. Similarly, in-field weather stations solar radiation data were characterized by a high percentage of missing values and inconsistencies; we used instead the R package nasapower (Sparks, 2018), which enables an easy access to NASA POWER surface solar radiation energy data. Some environments were irrigated: for those of which the precise amount was tracked during the growing season, these data were added to the final daily precipitation data.

Hence, the daily weather data consisted of the daily maximum, minimum and mean temperature (average of minimum and maximum daily temperatures), average wind speed, precipitation, humidity, incoming solar radiation. Based on these processed weather data, we were then able to calculate the daily growing degrees (Baskerville and Emin, 1969), the photothermal time (product between GDs and day length in hours, for each day, also referred as an environmental index; Li et al., 2018), the mean vapor pressure deficit, the reference evapotranspiration (ET_0) using FAO-56 Penman-Monteith method (Allen et al., 1998). These latter variables were computed because they incorporate crop physiological parameters which make them sometimes more relevant than the initial weather data.

2.4. Derivation of Environmental Variables per Hybrid Growth Stage

The next step was to obtain pertinent environmental predictors from daily weather summaries for the predictive modeling framework. The objective was to relate each hybrid phenotypic performance (e.g., yield) in a particular environment, individually characterized by its specific flowering dates, to the corresponding weather series during the growing season. To develop a unified framework across the different growing season lengths, which varied throughout locations and years, we used three critical maize growth stages, as was performed in previous similar work for other crops (Heslot et al., 2014b; Delerce et al., 2016; Gillberg et al., 2019; Monteverde et al., 2019). This approach was needed to account for the differential impact of weather-based variables according to the crop developmental stage. Each intermediate plant developmental stage could not be precisely determined since visual scoring for all stages is in practice highly time-consuming and expensive. However, the sowing date and the flowering date, i.e., when 50% of plants in a plot have visible silk, were recorded for each hybrid kept after phenotypic data analysis. Based on these known dates, three hybrid maize growth periods could be estimated: vegetative (from the planting date to 1 week before the 50% silking date); flowering (from 1 week before 50% silking date to 2 weeks after that date, which corresponds approximately to the end of the pollination period); and the grain filling stage (from the end of the flowering period to 65 days after, after which maturity should be reached). By definition, these three periods do not overlap. The typical duration of the grain filling stage varies according to the hybrid and the environment; nonetheless, based on literature and agronomic knowledge, the corn plant is normally at physiological maturity (R6) about 55–65 days after silking (Ritchie et al., 1993).

Based on these dates, 13 weather-based environmental predictor variables were computed for each phenological stage and therefore were both environment- and hybrid-specific (**Table 1**). We included stress covariates related to heat, as it is expected that an excess of heat can be detrimental, especially during the flowering stage, and results in a lower yield. To examine the presence of clusters of environments based on climatic similarity, a principal component analysis on the weather-based covariates using the R package factoextra (Kassambara and Mundt, 2017) was applied.

In addition to climatic variables, our framework accommodates four soil-based environmental variables: soil

TABLE 1 | Environmental predictor variables used in the prediction models.

Acronym	General description
P.V, P.F, P.G	Accumulated precipitation + irrigation (mm) by growth period
FreqP5.V, FreqP5.F, FreqP5.G	Frequency of days with more than 5 mm precipitation by growth period
MeanT.V, MeanT.F, MeanT.G	Average of daily mean temperature (°C) by growth period
MinT.V, MinT.F, MinT.G	Average of minimum daily temperature (°C) by growth period
MaxT.V, MaxT.F, MaxT.G	Average of maximum daily temperature (°C) by growth period
GDD.V, GDD.F, GDD.G	Cumulative growing degree days, Base 10°C (°C) by growth period
Photothermal.Time.V, Photothermal.Time.F, Photothermal.Time.G	Cumulative photothermal time (GDD x Day Length) by growth period
FreqMaxT30.V, FreqMaxT30.F, FreqMaxT30.G	Frequency of days with maximum temperature above 30°C by growth period
FreqMaxT35.V, FreqMaxT35.F, FreqMaxT35.G	Frequency of days with maximum temperature above 35°C by growth period
St30.V, St30.F, St30.G	Sum of the daily maximal temperatures above 30°C (°C)
CumSumET0.V, CumSumET0.F, CumSumET0.G	Accumulated reference evapotranspiration (mm), under standard conditions, according to the FA0-56 Penman-Monteith methodology for each growth period
CumDailyWaterBalance.V, CumDailyWaterBalance.F, CumDailyWaterBalance.G	Cumulative daily water balance, i.e., daily precipitation + irrigation - daily reference evapotranspiration (mm)
Sdrad.V, Sdrad.F, Sdrad.G	Accumulated incoming daily solar radiation (MJ m-2 day-1) by growth period
SandProp.SC	Sand composition (%)
Silt.Prop.SC	Silt composition (%)
ClayProp.SC	Clay composition (%)
OM.SC	Percentage of organic matter (%)

The suffixes refer to: V, vegetative period; F, flowering period; G, grain fill period; SC, soil covariate.

quality types (percentages of sand, silt, and clay composition) and percentage of soil organic matter. The majority of the soil information originates from the soil samples realized at each G2F field location; otherwise, when the location presented missing information, we defined an area of interest based on field geographical coordinates using the Web Soil Survey application for American locations, and the web mapping application Agricultural Information Atlas for Canadian locations, and retrieved the aforementioned data of interest. In the rest of the paper, the abbreviation "W" refers to the set of weather-based and soil-based environmental covariates. For the trait plant height, weather-based covariates from the grain filling stage were not used as explanatory variable for prediction, since this trait was usually measured shortly after flowering time.

2.5. Prediction Models Implemented

2.5.1. Linear Random Effects Models (LRE Models)

In multi-environment trial analysis and plant breeding experiments, linear random effects models, abbreviated to LRE models thereafter, are often used as genomic prediction models and were compared in this study with machine learning techniques, according to the models outlined in Jarquín et al. (2014). In particular, GxE can be modeled with a covariance function equal to the product of two random linear functions of markers and of environmental covariates, which is equivalent to a reaction norm model (Jarquín et al., 2014). An environment always refers to a Site x Year combination.

Main effects models

(1) *Model G + E: Marker + Environment Main Effects (baseline model)*

The response variable is modeled as the sum of an overall mean (μ), plus random deviations due to the environment E_i and to the genotypic random effect of the jth hybrid genotype g_j based on marker covariates (G-BLUP component), plus an error term ε_{ij}:

$$y_{ij} = \mu + E_i + g_j + \varepsilon_{ij}, \tag{1}$$

where $E_i \overset{IID}{\sim} N(0, \sigma_E^2)$, $\mathbf{g} \overset{IID}{\sim} N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ and $\varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$, and N(.,.) denotes a normally distributed random variable, IID stands for independent and identically distributed, and σ_E^2, σ_g^2 are the corresponding environmental and genomic variances, respectively.

g_j corresponds to a regression on marker covariates of the form $g_j = \sum_{m=1}^{p} x_{jm} b_m$, linear combination of p markers and their respective marker effects. Marker effects were regarded as IID draws from normal distributions of the form $b_m \overset{IID}{\sim} N(0, \sigma_b^2)$, $m = 1,...,p$. The vector **g=Xb** follows a multivariate normal density with null mean and covariance-matrix $Cov(\mathbf{g}) = \mathbf{G}\sigma_g^2$, where $G = \frac{XX'}{p}$ is the genomic relationship matrix, X representing the centered and standardized genotype matrix and p is the total number of markers.

(2) *Model G + S: Marker + Site Main Effects*

The present model allows to gain information from a site evaluated over several years, as it includes the site effect:

$$y_{kj} = \mu + S_k + g_j + \varepsilon_{kj} \tag{2}$$

Here y_{kj} corresponds to the phenotypic response of the jth genotype in the kth site with $S_k \overset{IID}{\sim} N(0, \sigma_S^2)$, $k = 1,...,K$.

(3) *Model G+E+W: Marker + Ennvironment + Environmental Covariates Main Effects*

This model incorporates additionally the main effect of the environmental covariates (including the longitude and latitude coordinates). We can model the environmental effects by a random regression on the ECs (**W**), that represents the environmental conditions experienced by each hybrid in each environment: $w_{ij} = \sum_{q=1}^{Q} W_{ijq}\gamma_q$, where W_{ijq} is the value of the qth EC evaluated in the ijth environment x hybrid combination, γ_q is the main effect of the corresponding EC, and Q is the total number of ECs. We considered the effects of the ECs as IID draws from normal densities, i.e., $\gamma_q \sim N(0, \sigma_\gamma^2)$. Consequently, the vector $\mathbf{w} = \mathbf{W}\boldsymbol{\gamma}$ follows a multivariate normal distribution with null mean and covariance matrix $\boldsymbol{\Omega}\sigma_w^2$, where $\boldsymbol{\Omega} \propto \mathbf{W}\mathbf{W}'$, and the matrix $\mathbf{W}$, which is centered and standardized, contains the values of the ECs. The model becomes then:

$$y_{ij} = \mu + E_i + g_j + w_{ij} + \varepsilon_{ij} \tag{3}$$

with $\mathbf{w} \sim N(\mathbf{0}, \Omega\sigma_w^2)$.

In this model, as explained in Jarquín et al. (2014), environmental effects are subdivided in two components, one that originates from the regression on numeric environmental variables, and one due to deviations from the Year-Site combination effect which cannot be accounted for by the ECs. Indeed, the environmental variables might not be able to fully explain the differences across environments. The modeling of the covariance matrices $\boldsymbol{\Omega}$ and $\mathbf{G}$ allows to borrow information between environments and between hybrid genotypes, respectively.

Models with interaction

(4) *Model G+E+GxE: main effects G+E with Genomic x Environment Interaction*

The model G+E was extended by including the interaction term between environments and markers (GxE):

$$y_{ij} = \mu + E_i + g_j + gE_{ij} + \varepsilon_{ij} \tag{4}$$

with $\mathbf{gE} \sim N(\mathbf{0}, [\mathbf{Z_g}\mathbf{G}\mathbf{Z_g'}] \circ [\mathbf{Z_E}\mathbf{Z_E'}]\sigma_{gE}^2)$, $\varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$, where $\mathbf{Z_g}$ and $\mathbf{Z_E}$ are the design matrices that connect the phenotype entries with hybrid genotypes and with environments, respectively; σ_{gE}^2 is the variance component of the gE_{ij} interaction term; and $\circ$ denotes the Hadamard product between two matrices.

(5) *Model G+S+GxS: main effects G+S with Genomic x Site Interaction*

Similar to the previous model, this model extends model G+S by including the interaction term between sites and markers (GxS):

$$y_{kj} = \mu + S_k + g_j + gS_{kj} + \varepsilon_{kj} \tag{5}$$

where $\mathbf{gS} \sim N(\mathbf{0}, [\mathbf{Z_g}\mathbf{G}\mathbf{Z_g'}] \circ [\mathbf{Z_S}\mathbf{Z_S'}]\sigma_{gS}^2)$, $\varepsilon_{kj} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$, where $\mathbf{Z_S}$ and σ_{gS}^2 are the design matrix for sites and the associated variance component for this interaction, respectively.

(6) *Model G+E+S+Y+GxS+GxY+GxE: main effects G+E+S+Y with Genomic x Environment Interaction, Genomic x Site Interaction and Genomic x Year Interaction*

This model corresponds to the most complete model using only basic GxE information (year and site information) about environments:

$$y_{jkm} = \mu + g_j + S_k + Y_m + E_{km} + gS_{jk} + gY_{jm} + gE_{jkm} + \varepsilon_{jkm} \tag{6}$$

where $\mathbf{gY} \sim N(\mathbf{0}, [\mathbf{Z_g}\mathbf{G}\mathbf{Z_g'}] \circ [\mathbf{Z_Y}\mathbf{Z_Y'}]\sigma_{gY}^2)$, $\varepsilon_{kj} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$, where $\mathbf{Z_Y}$ and σ_{gY}^2 are the design matrix for years and the associated variance component for this interaction, respectively.

(7) *Model G+E+W+GxW: main effects G+E+W with interactions between markers and environmental covariates*

The model G+E+W was extended by adding the interaction between genomic markers and environmental covariates. Jarquín et al. (2014) demonstrated that this interaction term induced by the reaction-norm model can be described by a covariance structure which corresponds, under standard assumptions, to the Hadamard product of two covariance structures: one characterizing the relationships between lines based on markers information (e.g., $\mathbf{G}$), and one describing the environmental resemblance based on ECs (e.g., $\boldsymbol{\Omega}$). The vector of random effects, denoted $\mathbf{gw}$ represents the interaction terms between markers and ECs, is assumed to follow a multivariate normal distribution with null mean and covariance structure $[\mathbf{Z_g}\mathbf{G}\mathbf{Z_g'}] \circ \Omega$. The model can be expressed as follows:

$$y_{ij} = \mu + E_i + g_j + w_{ij} + gw_{ij} + \varepsilon_{ij}, \tag{7}$$

with $\mathbf{gw} \sim N(\mathbf{0}, [\mathbf{Z_g}\mathbf{G}\mathbf{Z_g'}] \circ \Omega\sigma_{gw}^2)$.

(8) *Model G+E+W+GxW+GxE: main effects G+E+W with Genomic x Environment Interaction and Genomic x Environmental Covariates Interaction*

The interaction term gE_{ij} is incorporated in this model, because some GxE might not be completely captured by the interaction term gw_{ij}, and the model becomes:

$$y_{ij} = \mu + E_i + g_j + w_{ij} + gw_{ij} + gE_{ij} + \varepsilon_{ij} \tag{8}$$

Main and interactions effects included in the different models described above are summarized in **Supplementary Table 5**. Models using W, i.e., the matrix of environmental covariates, were tested with and without longitude and latitude data included. Additional combinations of main effects and interactions not detailed here were also evaluated and results are presented as **Supplementary Material**. These models were implemented in a Bayesian framework using the R package BGLR (Pérez and de Los Campos, 2014), for which the MCMC algorithm was run for 42,000 iterations and the first 2000 cycles were removed as burn-in with thinning equal to 5.

2.5.2. Machine Learning Based-Methods Used

The potential of machine learning models was explored using the following three algorithms: the linear regularized Elastic Net (Zou and Hastie, 2005), XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017). All the machine learning regression models were conducted in R version 3.6.1 (R Core Team, 2019) using the tidymodels framework (Kuhn and Wickham, 2020) and wrapper functions of treesnip (https://github.com/curso-r/treesnip/). Elastic net is a regularized linear regression method that has proven to be useful with datasets characterized by multicollinearity to identify the most relevant predictor variables as well as reducing the computing time (Zou and Hastie, 2005). It corresponds to a linear combination of two penalty terms: the lasso (L1 regularization), noted $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ and the ridge (L2 regularization), noted $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$. While the L2 penalty tends to contract the coefficients of highly correlated features toward each other, the L1 penalty supports a sparse solution, as many coefficients are zeroed. However, this method does not account for interactions between features.

Originally introduced by Friedman (2001), gradient boosting approach sequentially builds an ensemble of decision trees, with each new tree improving the predictions of the previous one by fitting on its residual errors. Two implementations of gradient boosting of decision trees (GBDT) for regression were used: Light Gradient Boosting Machine (LightGBM) and eXtreme Gradient Boosting (XGBoost). The two GBDT frameworks stand out from other similar boosting algorithms regarding their efficiency, which can be achieved by their common implementation of a histogram-based method for split finding, which groups continuous features into discrete bins. Hence, the algorithm does not iterate through all feature values, which is extremely time-consuming, but instead performs splitting on the bins. This speeds up training for very large datasets, as well as reducing memory usage. LightGBM, developed more recently, incorporates additional features, among others a downsampling during the training on basis of gradients. GBDT frameworks can handle well various types of data (binary, continuous data), and they are relatively robust to the effects of outliers among predictor variables (Hastie et al., 2009). Decision trees can capture, by construction, higher-order interactions between features, as well as nonlinear relationships between predictors and response variable (Friedman, 2001). Hence, interactions do not need to be explicitly provided as input data, since new splits are built conditional on preceding splits made on other predictors.

2.5.3. Data Pre-processing for Machine Learning-Based Models

For data processing, we used the R package recipes (Kuhn and Wickham, 2020). To reduce genomic data dimensionality, we did not input SNP data into our prediction models directly. Instead, we used the top 275 or 350 principal components (PCs) of SNP data, for the traits grain yield and plant height, respectively. This set of PCs was chosen after evaluation of the predictive ability using different sets of top PCs explaining a various proportion of the variance in the data. Covariates which had no variance were removed using the step_nzv function. Retained covariates were standardized to zero mean and unit variance. As for linear random effect models, we tested the influence on prediction of longitude and latitude data by including and removing them as predictor variables across the different cross-validation scenarios. The year was also included as an input variable as a predictor variable in some models to account for environmental variation not fully captured by environmental covariates. In that case, the factor variable was converted into four new variables corresponding to each level of the original predictor. To model the site effect in models without numerical environmental information, we used the simple geographic coordinates of each location instead of using its label. Indeed, in decision trees, the use of a categorical predictor with a high number of levels can lead to overfitting (Hastie et al., 2009).

2.5.4. Optimization of Hyperparameters and Hyperparameter Importance for Machine Learning-Based Models

Bayesian optimization using an iterative Gaussian process was used for hyperparameter tuning. It represents a much faster approach than grid search while allowing more flexibility in how the parameter space is covered. The Gaussian process builds a probability model based on an initial set of performance metrics obtained for various hyperparameter combinations during an initialization step, and predicts new tuning hyperparameters to test based on these previous results (Williams and Rasmussen, 2006; Snoek et al., 2012). Bayesian optimization incorporates prior assumptions on model parameter distribution and update it after each iteration, seeking to minimize the root mean square error (RMSE). Hyperparameter tuning was evaluated with 30 iterations under resampling based on a fivefold cross-validation (CV) with two repeats on the training set. **Supplementary Table 4** indicates the set of hyperparameters tuned for each method during this optimization step. This set of hyperparameters was then used to fit the whole training data and predict the test set, which was unused during the optimization of hyperparameters. The general procedure for this nested cross-validation is illustrated in **Figure 1**. Fine-tuning of hyperparameters is required in order to prevent overfitting and to achieve the best prediction accuracy and representation of the data.

In addition, we examined the role of each hyperparameter on the overall model performance. This analysis provide insights into the most important hyperparameters to primarily tune in order to yield accurate models. We focus here on the LightGBM algorithm and XGBoost. A method based on random forests and functional ANOVA (fANOVA) was proposed by Hutter et al. (2014) to quantify the marginal contribution of each hyperparameter and pairwise interaction effects. Briefly, we used the output table of performance metrics of each algorithm with different hyperparameter combinations, which was obtained during the optimization step. The metric (root mean square error) is then used as target variable while hyperparameters represent the explaining variables to fit a random forest algorithm. fANOVA is then applied to evaluate the importance of each hyperparameter used in the grid search.

2.5.5. Assessment of Prediction Accuracy for New Environments

In order to mimic real plant breeding problems, we considered four different cross-validation strategies aiming at predicting genotypes in environments that were never tested before, namely CV0-Year, CV0-Site, CV00-Year, and CV00-Site, described in Jarquín et al. (2017). The CV0 cross-validation scheme allows to borrow information in the training set about the performance of predicted genotypes in other tested environments, while the CV00 cross-validation scheme consists of the prediction of newly developed genotypes. This means that for implementation of the CV00 cross-validation, any observation from a genotype included in the test set (i.e., new environments) was removed from the training set. Predictions of untested genotypes can be achieved by exploiting information from marker data on

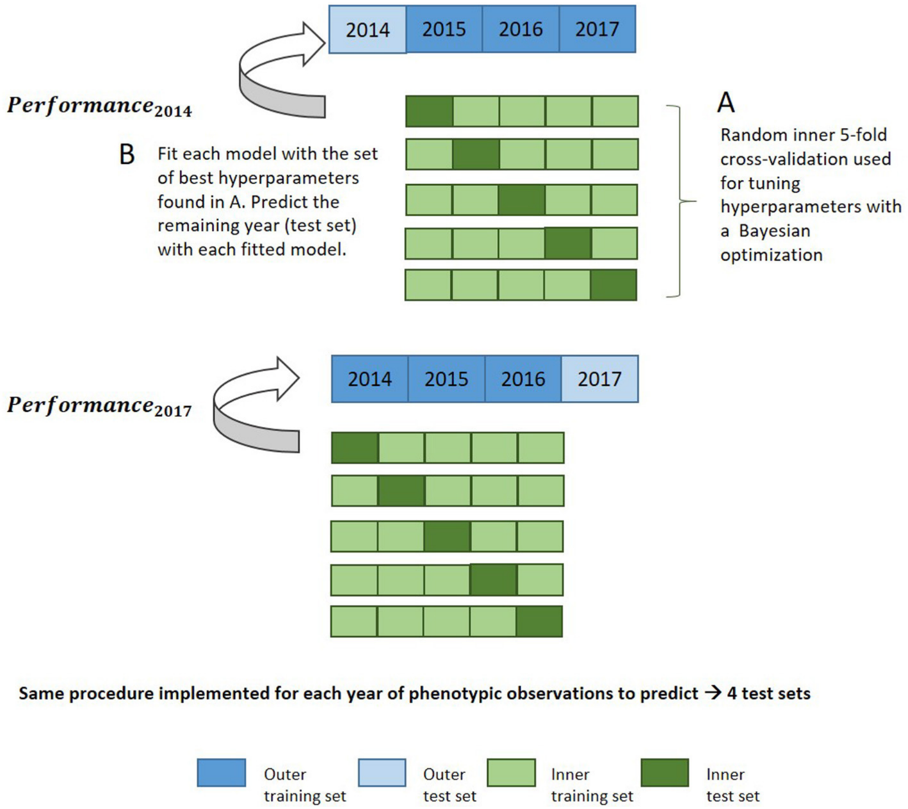

2014
2015
2016
2017
Performance2014
A
Random inner 5-fold cross-validation used for tuning hyperparameters with a Bayesian optimization
B
Fit each model with the set of best hyperparameters found in A. Predict the remaining year (test set) with each fitted model.
Performance2017
Same procedure implemented for each year of phenotypic observations to predict → 4 test sets
Outer training set
Outer test set
Inner training set
Inner test set

FIGURE 1 | Nested cross-validation diagram for evaluation of model performance in the leave-1-year-out CV scheme with a machine learning approach.

genetic similarities between genotypes from the training set and from the test set. Four scenarios in total were examined, which differ according to whether site or year were used to build the test set, and to the degree of relationship between training and test set: (1) CV0-Year, where phenotypic information about the performance of genotypes evaluated in the same year was masked; (2) CV00-Year, where phenotypic information about the performance of any genotypes present in the test set in other years was additionally masked; (3) CV0-Site, where phenotypic information about the performance of genotypes evaluated in the same site was masked and (4) CV00-Year, where phenotypic information about the performance of any genotypes present in the test set in other sites was additionally masked. In this procedure, the number of observations contained in each outer fold is not the same, due to the unbalanced character of the dataset. This approach reflects a common issue arising in multi-environment plant breeding trials, as all selection candidates cannot be grown in all environments. However, we can ensure a fair model comparison by having the same data splits across tested models.

Regarding evaluation metrics, we define the prediction accuracy as the Pearson correlation between the predicted and the observed performance in a given environment, i.e., correlations were computed on a trial basis.

In order to take into account the difference in sample sizes between environments, we evaluated the weighted average predictive ability across environments according to Tiezzi et al. (2017), for each combination of prediction model, predictor variables and trait, as following:

$$r_w = \frac{\sum_{j=1}^{J} \frac{r_j}{V(r_j)}}{\sum_{j=1}^{J} \frac{1}{V(r_j)}},$$

with r_j the Pearson's correlation between predicted and observed values at the j^{th} environment, $V(r_j) = \frac{1-r_j^2}{n_j-2}$ its sampling variance and n_j the total number of phenotypic observations in the j^{th} environment.

2.6. Variable Importance and Partial Dependence Plots for Grain Yield

We used the gain metric to quantify the feature importance in the XGBoost model fitted to the full dataset. This metric corresponds to the relative contribution of the variable to the ensemble model, calculated by considering each variable's contribution for each boosting iteration. A superior value of the gain for one feature compared to another feature means that this feature is more important to generate a‘prediction.

Overall partial dependence plots (PDPs) were computed using the R package DALEX (Biecek, 2018) using the four trained datasets from the CV0-Year scheme and the full dataset. PDPs are relevant to study how the predicted outcome of a machine learning model is partially influenced by a subset of explanatory variables of interest, by marginalizing over the values of all other variables.

The partial dependence profile of *f(X)* is defined as following by Friedman (2001):

$$f_S(X_S) = E_{X_C} f(X_S, X_C),$$

where the X_S represents the set of input predictor variables for which the effect on the prediction is analyzed, and X_C represent the complement set of other predictor variables used in the model. The following partial function can be used as an estimator:

$$\overline{f_S}(X_S) = \frac{1}{N}\sum_{i=1}^{N} f(X_S, x_{iC}),$$

where $x_{1C}, x_{2C}, ..., x_{NC}$ are the values of X_C observed in the training data. This means that we estimate this expected value as the average of the model predictions, over the joint distribution of variables in X_C, when the set of joint values in X_S is fixed. As emphasized by Hastie et al. (2009), partial dependence functions represent hence the influence of X_S on *f(X)*, after taking into account the average effects of the other variables X_C on *f(X)*.

2.7. Code Availability

A Github repository containing the various R scripts and Bash scripts used for phenotypic analysis, processing of weather data, spatio-temporal interpolation of missing weather data, and predictive modeling is available: https://github.com/cjubin/G2F_data.

3. RESULTS

3.1. Variability of Climatic Conditions in the Panel of Environments

Figure 2 reveals a partitioning of environments into clusters corresponding mostly to different US climate zones. It suggests that the sample of environments was broad enough to cover a large spectrum of environmental conditions across the North-American continent. The first two principal components explained more than 55% of total variation among environments on the basis of weather-based environmental covariates. The loading plot shows that MinT.F and GDD.F, FreqMaxT30.G, which are covariates related to temperature during flowering and grain filling stage, strongly influenced the first principal component (PC1). Environments from the South/Southeast (Arkansas, Texas, Georgia) showed positive PC1 and PC2 scores, which can be explained by a common humid subtropical climate, according to the Köppen climate type classification (Köppen and Geiger, 1930). One exception was one location in Texas (denoted 2014_TXH2), associated with more semi-arid climatic conditions. These results indicate that a closer geographical distance does not necessarily imply similar environmental conditions, based on climate types. For instance, environments from Delaware were closer to environments from the Midwest than Northeastern environments. Environments from the Midwest, associated with a humid continental climate, were situated mostly around the origin of the plot, and environments further north or in Canada exhibited the lowest temperatures among this set of sampled environments and presented a negative PC1 score.

3.2. Hyperparameter Importance for Gradient Boosting Approaches

Computing by fANOVA the marginal contribution of each tuned hyperparameter, using the performance data gathered during the hyperparameter optimization step on the different training sets, highlights large differences regarding their respective impact on model performance (**Supplementary Figure 3**). For the two gradient boosting algorithms, the learning rate (named eta in XGBoost) and the maximum depth of the tree were the most relevant algorithm parameters, as well as their interaction. The number of boosting iterations did not play a major role in model performance. We also found an advantage of using the hyperparameter feature_fraction and colsample_bytree, implemented in LightGBM and XGBoost, respectively, as it allowed an important reduction of the training time without having any observed negative effect on the accuracy of the predictions. It should be emphasized that we did not fully explore the influence of all possible hyperparameters implemented in these algorithms because of computational limitations, and therefore many of these were fixed during the hyperparameter optimization step.

3.3. Comparison of Model Performance Across Two Traits and Four Different CV Scenarios

CV0-Year

When the aim was to predict yield performance of already tested hybrids in new environments, the weighted average correlation of the baseline LRE model (G+E) was 0.356 (**Figure 3**; **Supplementary Table 6**). When the GxE term was added, the average correlation improved to 0.362. The model that included all interactions (G+E+W+GxW+GxE) was the best LRE model, while using only interactions between environmental covariates and genomic information (model G+E+W+GxW) slightly decreased the predictive ability of the baseline model to 0.347. In this prediction scenario, the two GBDT methods outperform all LRE models; model XGBoost-G+W+Y+Lon+Lat improved upon the baseline model by 18%. In addition, a small increase of predictive ability could be observed when environmental covariates were included as features for the machine learning-based frameworks. Furthermore, models that included geographical coordinates as predictor variables resulted in better prediction accuracies, and this revealed true across all prediction problems; therefore, **Figures 4**, **5** display results from LRE models using W as including longitude and latitude as predictor variables. For plant height, the baseline

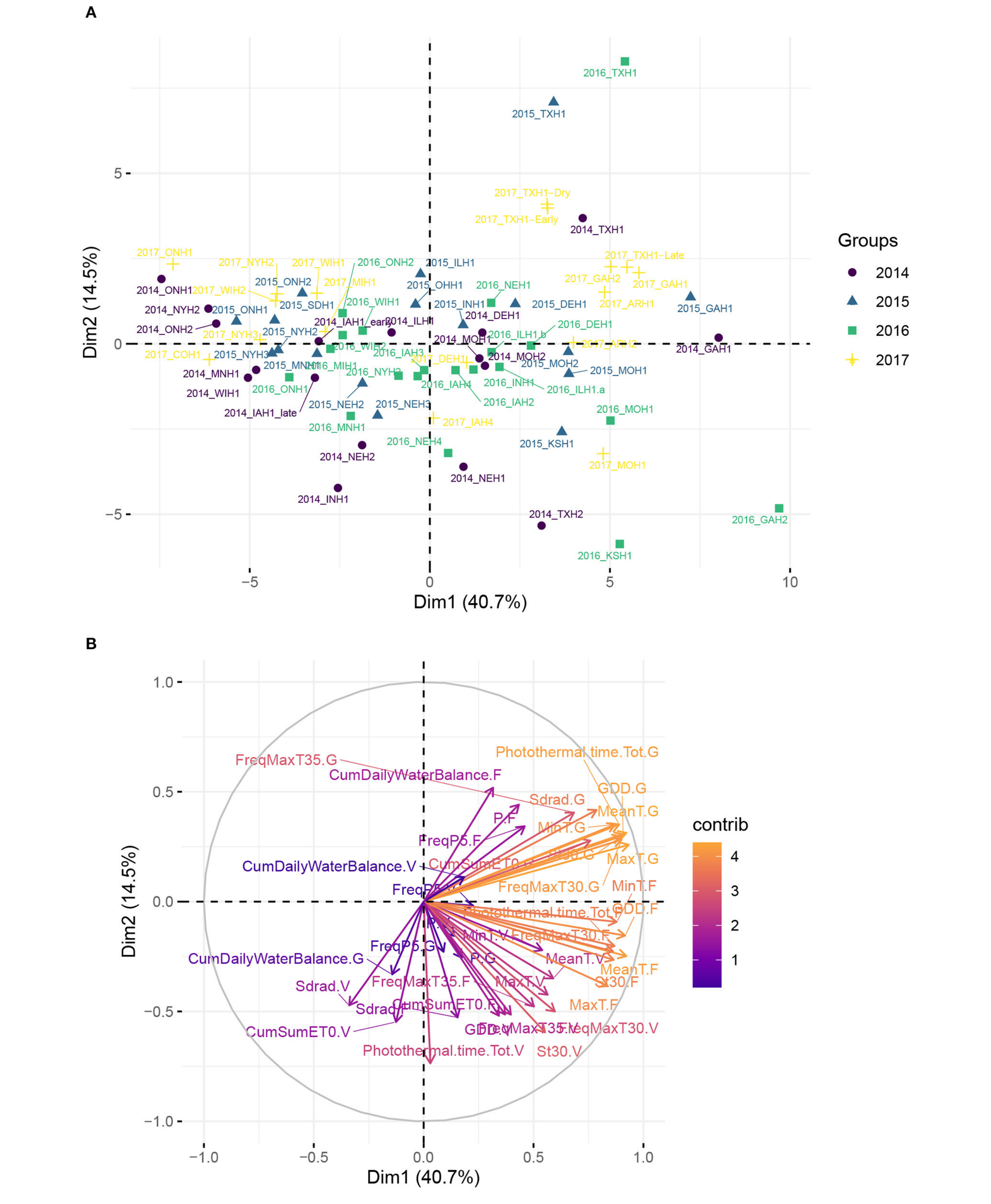


FIGURE 2 | Principal component analysis (PCA) plot of environmental data from the 71 environments, using the median flowering date as reference in each environment. **(A)** Maize trial experiments located in the US and in Canada used in analyses. Name of the locations and their geographical position are given in **Supplementary Table 1**. **(B)** Correlation plot of the weather-based covariates used in the PCA.

model performed best (**Figure 4**; **Supplementary Table 8**), and gradient boosting models incorporating environmental predictor variables performed consistently worse than models based only on genotypic data, geographical data and year information.

CV00-Year

CV00-Year produced lower average correlation coefficients for the two traits and for all models compared to CV0-Year, which illustrates that genomic prediction in multi-environment trials achieves better results when the training set includes information from the same genotypes evaluated in other environments. Regarding the trait grain yield (**Figure 3**; **Supplementary Table 6**), modeling the effect of sites instead of environments resulted in a small improvement of the predictive ability (4% better than the G+E model). Adding the GxE term to the LRE baseline model also positively affected the predictive ability (8% better than the G+E model). However, the LRE model with main site and genotype-by-site interaction effects (G+S+GxS) outperformed LRE models based on the modeling of year-location (E) effects. Overall the best predictive model for this trait was again the GBDT model XGBoost-G+W+Y+Lon+Lat, which displayed an average correlation of 0.301 (20% higher than the baseline model). GBDT models incorporating W performed between 6 and 13% better than GBDT models excluding W, which demonstrates the usefulness of environmental data for prediction of yield performance of new genotypes in an untested year. Among LRE models, the LRE model with all interactions and using enviromental data was the best model and resulted in an improvement of 17% over the baseline model. Regarding the trait plant height (**Supplementary Table 8**), the best predictive model was the baseline LRE model with an average weighted correlation of 0.604. Among LRE and GBDT models, models which did not include any environmental data performed better than those using these. An explanation for this lack of improvement with environmental data for plant height in this prediction problem can be that year and geographical position are appropriate and sufficient data to efficiently characterize environments for prediction of plant height, while using all environmental variables might generate noise here.

CV0-Site

The prediction of already tested genotypes in all environments associated with a common site revealed higher predictive abilities than with the CV0-Year prediction problem (**Figures 3**, **4**; **Supplementary Tables 7**, **9**). Indeed, based on our dataset, which covers many different sites across the US (see **Supplementary Figure 1**), the leave-one-site-out CV strategy generates large ratios across all training/test splits. This greater amount of data available to predict environments from one site can explain why this CV scheme obtained higher predictive abilities than the CV0-Year strategy. For the trait grain yield (**Figure 3**; **Supplementary Table 7**), the XGBoost-G+Lon+Lat+Y outperformed other models, showing an increase of 9% compared to the baseline LRE model. LightGBM models showed also better predictive abilities than LRE models. Only for LRE models did the use of environmental data yield a very small increase in predictive ability; the best result within this type of statistical approach was obtained by the model including all interactions (0.477, 3% higher than the baseline model). However, for the trait plant height (**Figure 4**; **Supplementary Table 9**), LRE models performed better than machine learning-based methods, with the model G+E+S+Y+GxS+GxY+GxE, which uses only basic information on environments, showing a mean correlation of 0.742. LightGBM and XGBoost methods with geographical and year information predicted reasonably well compared to the latter model (average r between 0.7 and 0.72), and again, the addition of environmental covariates decreased the predictive ability of GBDT models G+Lon+Lat+Y.

CV00-Site

As expected, the prediction of new genotypes in new sites resulted in lower mean correlations than CV0-Site for the two traits under study across predictive models. This highlights again the importance of the relationship between training and test sets. For the trait grain yield (**Figure 3**; **Supplementary Table 7**), the weighted average predictive ability of the reference model (G+E) was 0.248, and the model using sites instead of environment main effect was slightly better with a mean correlation of 0.265 (7% over G+E model). When the GxE term was added to the baseline model, the weighted average predictive ability was improved to 0.269 (8% over G+E model). It is worth to underline that models incorporating genotype-by-site effects performed even better (10% and 11% higher than the reference model). Modeling the interaction between ECs and genotypes and between environments and genotypes (model G+E+W+GxW+GxE) yielded an improvement of the baseline model by 19% (average r = 0.296), which was closely followed by the LightGBM and XGBoost models incorporating environmental covariates (between 11 and 16 % increase over the baseline model). As for the CV0-Year and CV00-Year CV schemes, the use of environmental data slightly increased the average predictive ability for grain yield. For the trait plant height (**Figure 4**; **Supplementary Table 9**), the baseline model with interactions by environment (G+E+GxE) outperformed other models. As for the previous prediction problems, environmental data decreased predictive abilities over all implemented models for the trait plant height.

When comparing the predictive abilities across traits, grain yield was the trait showing the lowest predictive ability across all CV schemes. Across all CV schemes, Elastic Net was the worst predictive modeling approach, which can be related to the absence of interactions between predictors in this model, if these are not explicitly provided as new features.

Figure 5; **Supplementary Tables 10**, **11** display the detailed within-environment correlation results for grain yield for two (CV0-Year and CV0-Site) cross-validation schemes. If a predicted environment is over the identity line, this means that there was an increment of the predictive ability by using environmental information. For CV0-Year, the machine learning-based model including environmental data outperformed the model only using geographical and year information in 44 of the 71 considered environments. For CV0-Site, however, the model with environmental features was better

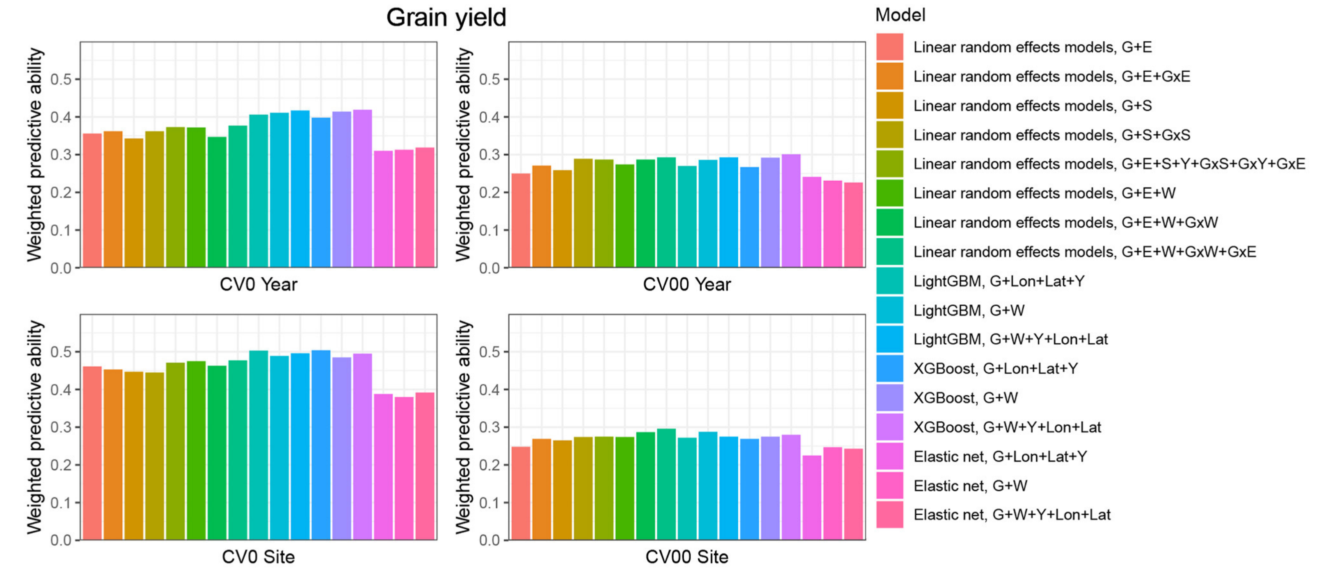


FIGURE 3 | Weighted average predictive ability across 71 environments obtained for four cross-validation schemes and 16 models for the trait grain yield. G, main effect of SNPs markers (genomic relationship matrix for LRE models; principal components derived from marker matrix for machine learning-based approaches); Y, year effect; S, site effect; GxS, genotype-by-site interaction; E, environment effect; GxY, genotype-by-year interaction; GxS, genotype-by-site interaction; GxE, genotype-by-environment interaction; GxW, interaction between W and SNPs; Lon, longitude; Lat, latitude; W, effect of weather- and soil-based covariates. For linear random effects models, results with models including longitude and latitude data in the matrix W are depicted here.

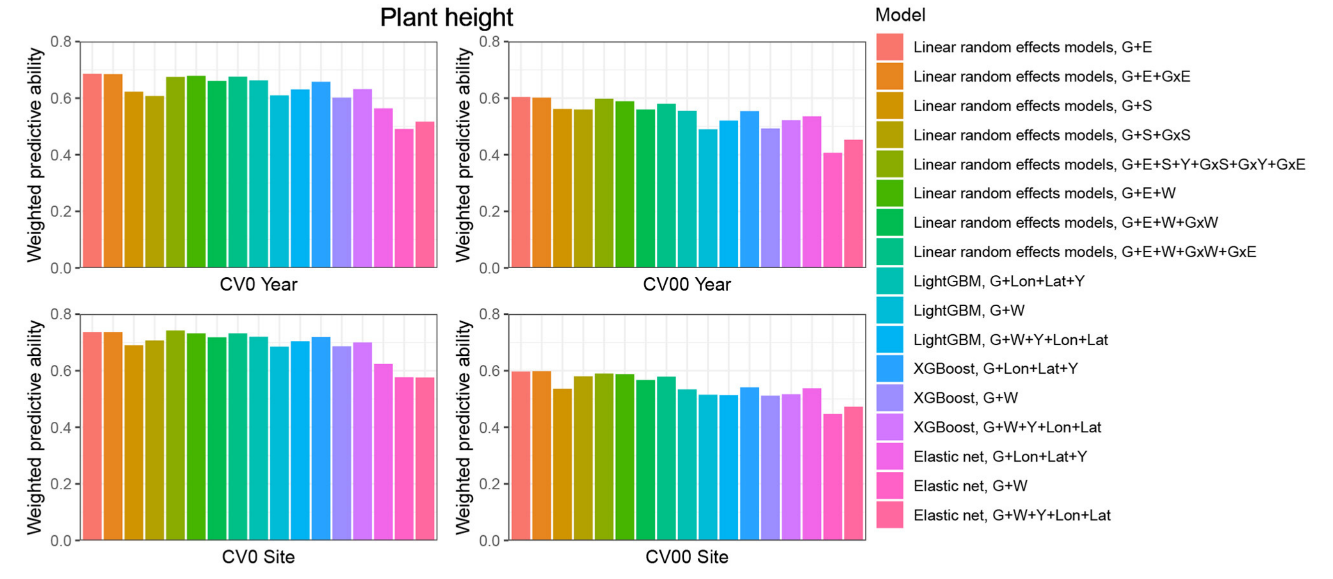


FIGURE 4 | Weighted average predictive ability across 71 environments obtained for four cross-validation schemes and 16 models for the trait plant height. G, main effect of SNPs markers (genomic relationship matrix for LRE models; principal components derived from marker matrix for machine learning-based approaches); Y, year effect; S, site effect; GxS, genotype-by-site interaction; E, environment effect; GxY, genotype-by-year interaction; GxS, genotype-by-site interaction; GxE, genotype-by-environment interaction; GxW, interaction between W and SNPs; Lon, longitude; Lat, latitude; W, effect of weather- and soil-based covariates. For linear random effects models, results with models including longitude and latitude data in the matrix W are depicted here.

than the less complex one in only 34 environments. This can be interpreted as a failure to explain a large part of the GxE by the computed ECs, and by a more efficient representation of environmental effects by simple geographic information.

3.4. Variable Importance

Regarding the trait grain yield, many of the identified top variables were related to temperature, such as the average minimum temperature during the flowering stage, or the

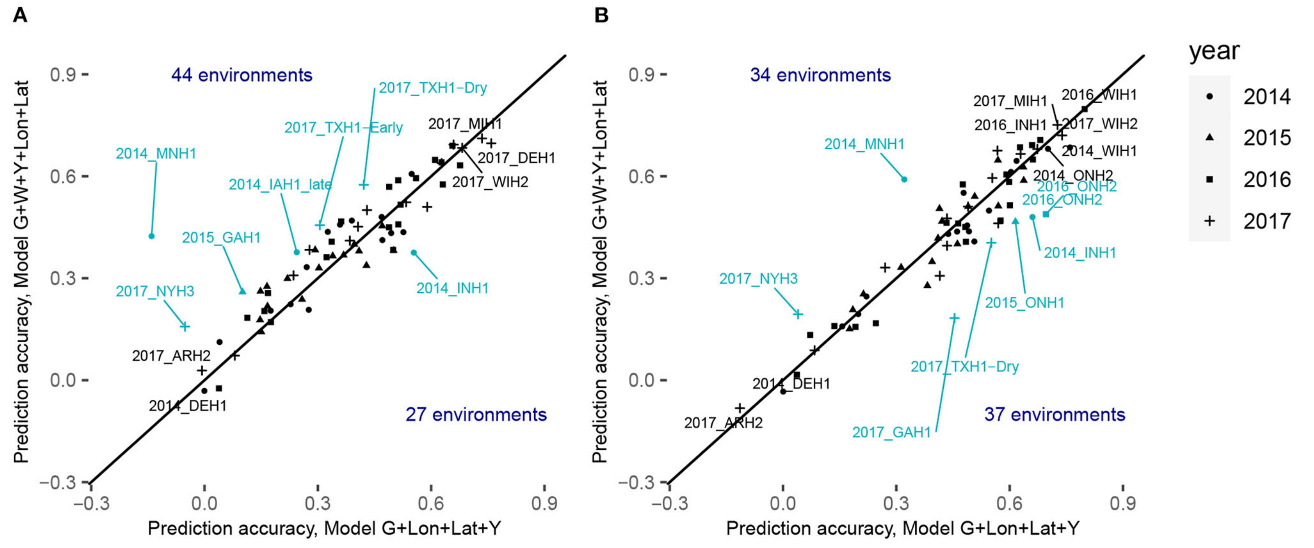


FIGURE 5 | Comparison of the within-environment predictive ability with different sets of predictors for the trait grain yield for XGBoost **(A)** with the CV0-Year scenario and **(B)** CV0-Site scenario. The x-axis corresponds to the within-environment correlation obtained with the model incorporating PCs derived from SNPs, year and geographical coordinates. The y-axis corresponds to the within-environment correlation obtained with the model incorporating PCs, year, W (i.e., weather- and soil-based covariates) and geographical coordinates. The line indicates the identity. Blue-colored points with a label indicate environments for which the absolute difference between the two predictive abilities was superior to 0.13. Black-colored points with a label indicate the least and the most accurately predicted environments.

frequency of days during which the maximum temperature was above 35°C (**Figure 6**). Organic soil matter concentration was the third most important feature, which demonstrates that fields with fertile soils were associated with higher yields. The amount of water received by the field (P.V) during the vegetative and grain filling stage was also a major feature for the model, as well as the frequency of days during the vegetative stage for which the amount of water was greater than 5 mm. Regarding the trait plant height, variables based on soil information played a major role for trait prediction, as they likely affect the crop shoot architecture. The amount of water received during the vegetative stage was also an important explanatory variable for plant height.

Partial dependence plots (**Figure 7**) show that minimum temperature at flowering stage was strongly impacting yield from approximately 20°C onwards. Maximum temperature during the vegetative stage had a detrimental effect on yield, suggesting that very elevated temperatures can impair a normal plant growth, eventually required to achieve optimal grain yield, although it tended to have a more gradual effect than minimum temperature at flowering stage. The relationship with yield of the total amount of precipitation during the vegetative stage was positive, before reaching a plateau. A high soil organic matter content yielded in superior yield predicted values.

4. DISCUSSION

Breeders, working on the development of climate resilient cultivars, risk making incorrect selection decisions if genotype-by-location and genotype-by-year interactions are not properly accounted for (Jarquín et al., 2017; De Los Campos et al., 2020). By incorporating environmental variables in our models, we assessed the value of these predictor variables for genomic prediction of complex phenotypes across four cross-validation scenarios. Gradient boosting frameworks based on decision trees have demonstrated high prediction performance for traits affected by non-additive effects (Abdollahi-Arpanahi et al., 2020), as well as model interpretability to extract important insights from the model's decision making process (Shahhosseini et al., 2020). Thus, a second objective was to evaluate these new prediction methods on the basis of prediction accuracies and for identification of the most relevant environmental variables.

4.1. Comparison of Prediction Methods Across the Two Traits

We observed that GBDT frameworks produced a slightly improved predictive ability for grain yield compared to the linear random effects models in three (CV0-Year, CV00-Year, and CV0-Site) out of the four CV schemes. However, no advantage was observed when GBDT was used to predict plant height. Overall, GBDT methods were competitive to LRE models, but we did not find any case where these machine learning-based methods considerably exceeded the predictive ability of LRE models. Previous studies have suggested that machine learning-based approaches can provide superior accuracy for prediction of phenotypic traits characterized by substantial non-additive effects. For instance, results from Zingaretti et al. (2020) in strawberries suggest that traits, exhibiting large epistatic effects, can be better predicted by convolutional neural networks (CNN), than by Bayesian penalized linear models. On the other hand, for

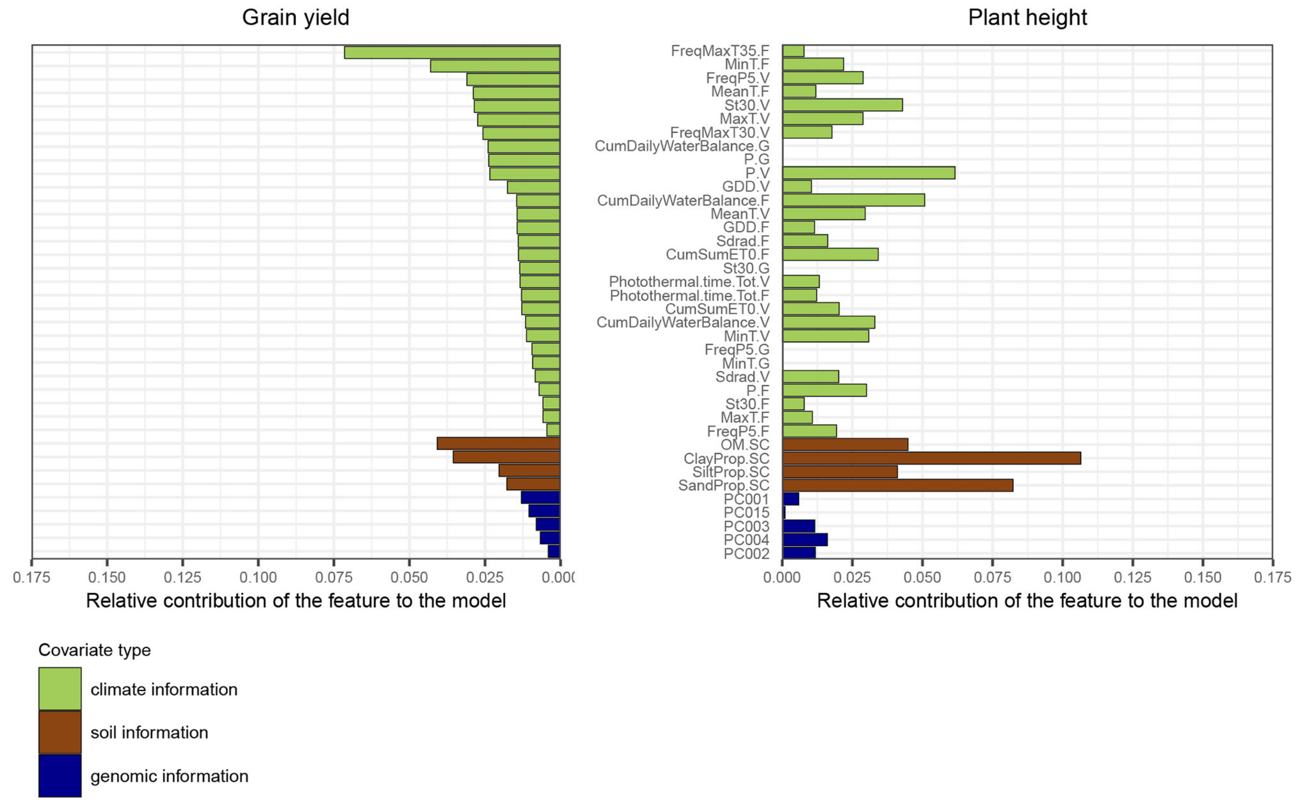


FIGURE 6 | Feature importance ranking based on the average relative gain per feature obtained with the model XGBoost-G+W, for the two traits grain yield and plant height. The metric was estimated using a model fitted on the full dataset. The gain represents the improvement in accuracy when using a feature for splitting, across all trees in the model. The order of features is based on feature performance within covariate class for the trait grain yield. The sum of all feature contributions is equal to 1. Weather-based variables from the grain filling stage were not used to predict plant height.

moderately to highly heritable traits, no real advantage of using machine learning-based methods was observed in their study. Bellot et al. (2018) pointed out that human height, a trait with a prevailing additive component and a polygenic architecture, was better predicted by linear methods than by CNNs. For other traits they examined in their study, a deep learning approach did not significantly outperform other methods in terms of prediction accuracy. Similar conclusions were drawn by Azodi et al. (2019) who reported an inconsistency of performance for non-linear machine learning-based algorithms in comparison with linear algorithms, according to the trait under study.

In our study, we incorporated not only genomic-based, but also environmental-based predictor variables. Yield component traits are controlled by numerous physiological processes under the influence of environmental factors, which can explain the large contribution of the GxE variance component for the phenotypic variance of grain yield, while for plant height, the proportion of variance explained by GxE is generally much lower than the proportion of variance related to genetic effects (Olivoto et al., 2017; Rogers et al., 2021). Nonlinear relationships between some environmental factors, such as temperature or rainfall amounts, and grain yield are well-known in the field of ecology and agriculture (Troy et al., 2015; Li et al., 2019). Hence, the slightly better prediction performance for grain yield with GBDT frameworks might originate from their ability to model nonlinear effects of environmental predictor variables, as observed with the partial dependence plots, as well as interactions with other predictor variables like genomic-based principal components. This asset was also described by Heslot et al. (2014b) when implementing soft rule fit (a modified ensemble method) capturing nonlinear interactions between markers and environmental stress covariates. Additional studies are required to validate this hypothesis using other phenotypic traits showing various genetic architectures. Moreover, it should be noted that we used only linear kernels in the reaction norm models to model genetic and environmental similarities. This means that we did not account for the specific combining ability (i.e., nonlinear genetic effects, due to dominance or epistasis, of specific hybrid combinations) which can influence the magnitude of yield heterosis in maize hybrids. Alternative approaches exist to model additive and dominant genetic effects, as well as environmental relatedness with nonlinear kernels (Bandeira e Sousa et al., 2017; Cuevas et al., 2018; Costa-Neto et al., 2020a). Bandeira e Sousa et al. (2017) and Cuevas et al. (2018) obtained better predictive abilities when using a Gaussian kernel rather than a linear GBLUP kernel with multi-environment G–E interactions models.

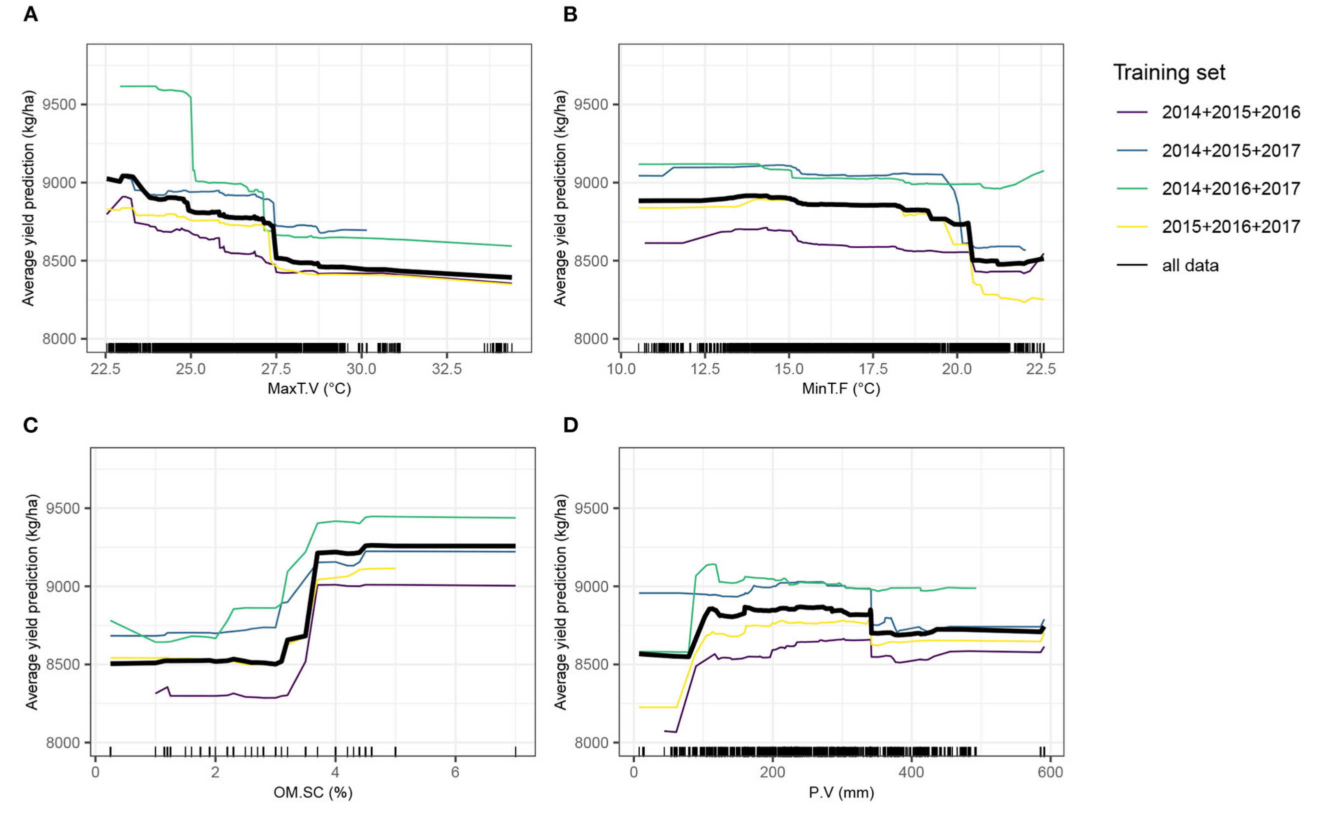


FIGURE 7 | Partial dependence plots (PDPs) showing the behavior of the expected value of predicted yield as a function of four top-ranked predictor variables. The Y-axis value of a PDP is calculated average of all model predictions obtained from the training dataset, when the value of the predictor variable is equal to X. The four training sets from the leave-1-year-out cross-validation scheme (CV0-Year) and the full dataset, separately trained with XGBoost, were used. Tick marks indicate individual observations. **(A)** MaxT.V, maximum temperature during the vegetative stage; **(B)** MinT.F, minimum temperature during the flowering stage; **(C)** OM.SC, percentage of soil organic matter; **(D)** P.V, Amount of precipitation and irrigation during the vegetative stage.

More recently, Costa-Neto et al. (2020a) implemented Gaussian and arc-cosine kernels-based approaches on both genomic and environmental datasets from a MET maize dataset, and noted an improvement in prediction accuracy using these methods across various cross-validation strategies. These results highlight the potential of nonlinear methods to better unravel nonlinear relationships existing in the input space.

4.2. Model Performance Under Various Prediction Problems

The four cross-validation schemes we evaluated represent challenging prediction problems. They seeked to assess the ability of the models to predict the effect of unknown combinations of environmental stresses on the studied phenotypic traits in a new year (CV0-Year and CV00-Year) or in a new site (CV0-Site and CV00-Site). Previously published work has revealed somewhat similar ranges of prediction accuracies for this trait in maize (Costa-Neto et al., 2020a; Jarquin et al., 2020). In winter wheat, Jarquín et al. (2017) and Sukumaran et al. (2017) reported the predictions of yield performance in future years (CV0-Year) as the most challenging prediction problem on the basis of results obtained for various cross-validation schemes, and results of Sukumaran et al. (2018) showed that modeling site effect instead of environment effect based on basic information about the environments (year and location) had a positive effect on predictive ability with CV0-Year, as we could also observe for CV0-Year, CV00-Year, and CV00-Site in our results. Indeed, this type of models allows to exploit information from the same site tested across several years. Another factor which is important to take into account in multi-year breeding data, as emphasized by Bernal-Vasquez et al. (2017), is the degree of genetic relatedness between the training and validation sets. Hence, CV00-Year and CV00-Site were more challenging prediction problems than CV0-Year and CV0-Site, respectively, and yielded lower weighted mean correlations across all models.

Regarding the usefulness of environmental information, the best model for grain yield based on mean predictive ability included these data for three (CV0-Year, CV00-Year, and CV00-Site) out of the four CV schemes. In addition, it must be taken into account that much less phenotypic observations were masked for CV0-Site (1/28, about 3.6% on average, with some sites being present more often than others across years in our dataset) than for CV0-Year (1/4, about 25% as the dataset is unbalanced). Hence, we can consider CV0-Year and

CV00-Year as more challenging prediction problems than CV0-Site and CV00-Site in our study. The improvement due to the incorporation of environmental data was however less remarkable and less consistent across CV schemes than expected, which was in contrast with previous results. Monteverde et al. (2019) also implemented a leave-1-year-out scenario, with one unique location present in the dataset, and the best prediction accuracies for grain yield were always reached by the models integrating environmental predictors alongside genomic predictors. Findings from Costa-Neto et al. (2020a) also show a significant increase of prediction accuracy with the linear GB kernel incorporating environmental data in a CV0 scheme, but the authors additionally modeled dominant genetic effects, which were not accounted for in our study. On the other hand, Jarquin et al. (2020) also used the same Genomes to Fields dataset and reported a lack of enhancement when using a model that solely incorporated interactions between genotype and environmental covariates (i.e., without using the environment label). The best predictive models for the CV0 and CV00 schemes, that they implemented, included both genotype-by-environment and genotype-by-EC interactions, similarly to our results (**Supplementary Tables 6–9**). In agreement with the reasons invoked by the authors of this study, we argue that environmental data are especially relevant for predictions when a larger number of environments is used, e.g., by testing sites within a limited geographical range with relatively similar environmental conditions across multiple years. This was for example achieved in the study of De Los Campos et al. (2020), where 16 sites located in France were tested over 16 years. A reasonable hypothesis is that historical weather data obtained across multiple years for a specific geographical area can lend the model reliable information on the effect of year-to-year climatic variation on phenotypic performance, in addition to site-based factors (soil and geographical position). A finding supporting this hypothesis is that the environments, which showed the best prediction accuracies with an environmental model, corresponded generally to the sites which were repeated across years, like Madison (WI) or College Station (TX) (**Supplementary Tables 10**, **11**). Interestingly, 2014_TXH2, a location for which data were only included for a single year, showed a moderate prediction accuracy with the XGBoost model without environmental information in CV0-Year ($r = 0.28$; **Supplementary Table 10**), which was superior to the model with environmental covariates ($r = 0.21$ with all environmental covariates included). We can suppose that the inclusion of environmental information, when predicting a new environment with properties that are very different from environments covered by the training set, is not useful to enhance the predictive ability of the model using basic predictors, such as the year factor and geographic coordinates. Extreme weather events can make some environments very unpredictable. 2017_ARH1 and 2017_ARH2 exhibited a very low prediction accuracy for grain yield (< 0 for 2017_ARH2) in both CV0-Year and CV0-Site (**Supplementary Table 11**), which is likely to be related to the effect of the tropical storm Harvey at the end of August 2017, which caused substantial lodging due to wind and excessive rainfall affecting the yield, and was reported by collaborators in the metadata.

4.3. Incorporation of Weather-Covariates in the Predictive Models

The use of environmental information yielded a small gain in average prediction accuracy for many models tested on grain yield, but did not lead to any improvement for plant height. For this latter trait, the large influence of soil-based variables, illustrated by the variable importance ranking (**Figure 6**), can also possibly explain why prediction models using only geographical coordinates outperformed more elaborate models. For this trait, latitude and longitude data might indirectly capture information which is site-specific and repeatable across years, e.g., related to the quality of soil. For instance, environments from the Corn Belt, which were present in our dataset, usually exhibited fertile soils with much higher organic soil matter content than environments located in other US regions. Costa-Neto et al. (2020b) highlighted that simple geographic-related information, such as longitude and latitude data, can also efficiently represent environmental patterns that are specific to a site (for instance related to soil characteristics), and hence capture well genotype-by-site interaction while using only two variables.

In general, the lack of real enhancement of predictive ability may result from the way we incorporated developmental stages into our models, as we defined only three main developmental stages (i.e., vegetative, flowering and grain filling stages). Trial data often lack a rigorous collection of phenological data due to phenotyping costs. A possible solution to predict plant developmental stages can be to use crop models, such as APSIM (Holzworth et al., 2014) or SiriusQuality (Keating et al., 2003), as done in related studies (Heslot et al., 2014b; Rincent et al., 2017, 2019; Bustos-Korts et al., 2019). In our case, we did not implement a crop model since we aimed at estimating the flowering stage at the hybrid level as accurately as possible, as it is known to be a critical period for the determination of yield-related components. Therefore, we based our environmental characterization on available field data (sowing date and silking date scored) in order to derive environmental covariates for three main developmental stages, similarly to Monteverde et al. (2019) in rice. The reported variability among crop growth models (CGM) in simulating temperature response can complicate the task of choosing the most appropriate one (Bassu et al., 2014). In addition, the task of integrating genetic variation for earliness in crop growth models can also be rather challenging, with the risk that the predicted developmental crop stages might not appropriately reflect the plant developmental stages observed in the field if the model does not properly account for genotype-specific parameters (Rincent et al., 2019). Technow et al. (2015) developed a complex framework combining both CGM and whole-genome prediction, where the CGM is used to predict grain yield as a function of several physiological traits and of weather and management data. Genotype-specific physiological parameters were estimated in this study by running a Bayesian algorithm which models them as linear functions of the effects of genomic features. It would be of high interest to apply CGM models on this dataset by taking advantage of the flowering time data that are available. We should also mention that other types of input data could be incorporated in future analyses, such as the type of field management, the field disease pressure, preceding crop, or the presence of external treatments (organic, nitrogen fertilizers).

4.4. Prerequisites to Use Machine Learning-Based Models and Their Usefulness to Understand Significant Environmental Factors

Specific techniques should be employed to ensure an efficient application of machine learning-based models. These can provide better results when expert knowledge is incorporated (Kagawa et al., 2017; Roe et al., 2020; Brock et al., 2021). Here, we restricted weather information to the duration of the growing season, transformed some raw weather information into new variables (evapotranspiration) and built stress indices besides typical climate covariates based on previous biological knowledge (e.g., detrimental temperature thresholds for maize (Greaves, 1996; Schlenker and Roberts, 2009; Lobell et al., 2014; Zhu et al., 2019; Mimić et al., 2020). Prior understanding of the role of input features can help mitigate the risk of using irrelevant information in the model. As expected, the correlation matrix between environmental covariates (**Supplementary Figure 2**) showed that numerous predictor variables were highly correlated with each other, especially those related to temperature and heat stress. We did not perform feature selection based on the Pearson correlation coefficients between environmental covariates, because of the risk of dropping highly predictive variables, since the metric ignores the relationship to the output variable. In addition, methods based on decision trees can perform internal feature selection, making them robust to the inclusion of irrelevant input variables and to multicollinearity (Hastie et al., 2009; Kuhn et al., 2013). If two variables are strongly correlated, the decision tree will pick either one or the other when deciding upon a split, which should not eventually affect prediction results. Another approach to reduce the number of features and reduce training time is to apply feature extraction, as we did by deriving principal components from the genotype matrix and use these as new predictor variables in the machine learning-based models. This procedure did not seem to affect model performance.

Machine learning models often require an elaborated hyperparameter optimization strategy, implying for example a nested cross-validation approach which can be computationally expensive (Varma and Simon, 2006), since it involves a series of train/validation/test set splits to prevent data leakage. Inadequate model tuning can result in a suboptimal performance of the algorithm. Here, we found that the hyperparameters such as the learning rate or tree depth were relevant regularization parameters to reduce the model complexity, thereby dealing with overfitting. In accordance with these results, other authors had also reported these two hyperparameters as the most important ones for another gradient boosting library similar to LightGBM, Adaboost (Van Rijn and Hutter, 2018). In general, lower values of the learning rate (< 0.01) are recommended to reach the best optimum (Ridgeway, 2007). Nonetheless, as the learning rate is decreased, more iterations are needed to get to the optimum, which implies an increase of the computation time and of additional memory (Ridgeway, 2007; Kuhn et al., 2013). With regard to the tree depth, a relatively low maximal depth generally helped to prevent overfitting, and better results were generally obtained with our data using a tree depth lower than to 8. The deeper a tree is, the more splits it contains, resulting in very complex models which do not generalize well on new data. Knowledge regarding the most important hyperpararmeters to tune is useful if limited computational resources hamper the investigation of numerous hyperparameter combinations during the training phase. Our results demonstrated similar predictive abilities of LightGBM and XGBoost, with a clear speed advantage for LightGBM, which ran often more than twice as fast. This asset relies in particular on a feature implemented in LightGBM, the gradient-based one-side sampling method (GOSS), which implies that not all data actually contribute equally to training. Training instances with large training error (i.e., larger gradients) should be re-trained, while data instances with small gradients are closer to the local minima and indicate that data is well-trained. Hence, this new sampling approach focuses on data points with large gradients and keeps them, while randomly sampling from those with smaller gradient values. A drawback of this method is the risk of biased sampling which might change the distribution of data, but this issue is mitigated in LightGBM by increasing the weight of training instances with small gradients. The main advantage is that it makes LightGBM much faster with comparable accuracy results. Another crucial aspect when applying machine learning models is the adequacy of the dataset for machine learning applications, which should be large enough to allow the algorithm to learn from the data (Géron, 2019). In our case, we benefited from a very large training dataset and a low feature-to-instance ratio (316/18,325).

In our study, on top of prediction applications, tree-based methods were also used to obtain estimates of feature importance, and thereby contributed to a better understanding of key abiotic factors driving the response of the tested genotypes. Feature importance rankings and partial dependence profiles showed that the minimal temperatures and indices related to prolonged heat stress, or to amounts of water received in the field, especially at the flowering stage, ranked among the most important variables for grain yield. When comparing these results with established agronomic knowledge, it was reported that, above a certain threshold, high minimum temperature can lead to an increase of the rate of senescence and reduce the ability of the plant to produce grain across many plant species (Hatfield et al., 2011; Hatfield and Prueger, 2015). Previous research also revealed that increases in average night temperatures were associated with a reduction of grain yield in maize (Millet et al., 2019) and in rice (Welch et al., 2010). In an alternative study on rice cultivars in Colombia, Delerce et al. (2016) identified high minimum temperature (above 22.7°C) as one of the most important environmental factors negatively impacting grain yield by using a machine learning approach based on conditional inference trees. Exposure to temperatures exceeding 35°C during the flowering stage was also a key factor in our study (best predictor variable for grain yield), which can be related to a loss of pollen viability, and consequently to a reduced final kernel set (Hatfield et al., 2011). In our study, water availability at vegetative and grain-filling stages appeared to affect yield, in accordance with the literature outlining that any water deficit during these growth stages can impact grain yield (Denmead and Shaw, 1960; Cakir, 2004), with a more significant impact when water stress occurs during the grain-filling stage (Cakir, 2004). Caution should nonetheless be taken regarding feature importance ranking due to the important correlations between some environmental variables. Furthermore, only 4 years of field trials were used in our analyses, therefore variable importances could be refined with additional data from following years, to mitigate the influence of some environments characterized by adverse climatic conditions and potentially acting as outliers.

4.5. Applications

The usefulness of medium to high prediction accuracies, when predicting the performance in a new environment, must always be related to our predictability of the environmental variation. If the weather fluctuates considerably year to year, then the environmental predictors used to compute these predictions might be very different from the true value in the corresponding year. In addition, even if more precise climate change models were available to improve upon the precision of environmental predictors, predictions of observations falling outside the applicability domain, i.e., the range of predictor space in the training set for which the model can give relativey accurate predictions (Netzeva et al., 2005), might not be trustworthy and should be used cautiously (Kuhn et al., 2013). The degree of similarity of the new test set to the training set should hence always be carefully considered.

While some environmental factors are repeatable from year to year, such as the soil type or agronomic practices, a large part of the GxE variation is attributable to weather patterns. Hence, the success of this type of prediction scenario depends on the relative stability of the climate in the targeted regions across years. Nonetheless, we posit that our approach presents two key advantages to predict performance in future years. First, because they are fundamentally data-directed, the tree-based models can take into account new phenotypic data in the training set in a more flexible manner than classical mixed models, without the need to explicitly specify interactions for example. The development of high-throughput phenotyping technologies announces a future enhancement of rapid and accurate training data (Juliana et al., 2019). The predictive frameworks we presented here can make use of new information to refine the estimated effects of the predictor variables. Secondly, we were able to predict a quantitative phenotype in a new environment by using a novel configuration of genotypic and environmental predictors describing it. A point of interest relates to resource allocation and the possibility to select more efficiently candidates to test in field trials. Based on the exploration of different plausible climatic scenarios—within a range of conditions experienced by the training set—these models can help to evaluate which genotypes might be more adapted to which range of environmental conditions. For regions or target population of environments presenting relatively stable climatic conditions across years, the probability of success of this type of predictive modeling approach is heightened.

5. CONCLUSIONS

Encouraged by the effectiveness of machine learning-based frameworks reported in the recent literature across various research fields, we compared two popular ensemble models with linear random effects models implemented in a Bayesian framework and a regularized linear model. In three CV schemes with the trait grain yield, the use of gradient boosting models resulted in a slight improvement of the average predictive ability but not for plant height. This finding indicates that machine learning-based approaches can be envisaged for genomic prediction but their efficiency may vary according to the trait under study and its degree of responsiveness to environmental variation. For a trait strongly under the influence of environmental factors, machine learning-based models could provide predictive abilities similar or slightly superior to linear random effects, and could additionally be used for interpretation of feature ranking and to build partial dependence plots detailing relationships between predictor variables and outcome. Provided further efficiency gains in machine learning algorithms, as well as the standardization and harmonization of large-scale environmental data, new opportunities in the field of predictive modeling for developing climate resilient varieties appear forthcoming.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. Raw genotypic, phenotypic, weather, and soil data from the Genomes to Fields Initiative can be found at: https://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/GenomesToFields_2014_2017_v1.

AUTHOR CONTRIBUTIONS

CW analyzed the data and wrote the manuscript. TB and HS supervised research. CW, TB, HS, GM, and PT designed the study. TB, HS, GM, SdS, and PT supported with statistical advice. CW, TB, HS, GM, SdS, PT, MS, and J-CR participated in the interpretation of results and contributed to discussion. All authors contributed to the writing of the final draft and approved the manuscript.

FUNDING

Financial support for CW was provided by KWS SAAT SE by means of a Ph.D. fellowship. Additional financial support was provided by the University of Göttingen and by the Center for Integrated Breeding Research. We acknowledge support by the Open Access Publication Funds of the Göttingen University.

ACKNOWLEDGMENTS

The authors would like to thank the G2F Consortium for making data publicly available, sharing them among collaborators, and for their constructive feedback with this study. The authors acknowledge support by the computing center for the university of Göttingen (GWDG) through the use of their High Performance Computing resources. We would like to thank two reviewers for their thoughtful ideas and comments that improved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.699589/full#supplementary-material

REFERENCES

Abdollahi-Arpanahi, R., Gianola, D., and Peagaricano, 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

AlKhalifah, N., Campbell, D. A., Falcon, C. M., Gardiner, J. M., Miller, N. D., Romay, M. C., et al. (2018). Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. *BMC Res. Notes* 11:452. doi: 10.1186/s13104-018-3508-1

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). *Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements-Fao Irrigation and Drainage Paper 56, Vol. 300*. Rome: Fao. D05109.

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* 9, 3691–3702. doi: 10.1534/g3.119.400498

Bandeira e Sousa, M., Cuevas, J., de Oliveira Couto, E. G., Pérez-Rodríguez, P., Jarquín, D., Fritsche-Neto, R., et al. (2017). Genomic-enabled prediction in maize using kernel models with genotype× environment interaction. *G3* 7, 1995–2014. doi: 10.1534/g3.117.042341

Baskerville, G. L., and Emin, P. (1969). Rapid estimation of heat accumulation from maximum and minimum temperatures. *Ecology* 50, 514–517. doi: 10.2307/1933912

Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J. W., et al. (2014). How do various maize crop models vary in their responses to climate change factors? *Glob. Chang Biol.* 20, 2301–2320. doi: 10.1111/gcb.12520

Bates, D., Mchler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw. Articles* 67, 1–48. doi: 10.18637/jss.v067.i01

Behravan, H., Hartikainen, J. M., Tengström, M., Pylkäs, K., Winqvist, R., Kosma, V.-M., et al. (2018). Machine learning identifies interacting genetic variants contributing to breast cancer risk: a case study in Finnish cases and controls. *Sci. Rep.* 8, 1–13. doi: 10.1038/s41598-018-31573-5

Bellot, P., de Los Campos, G., and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics* 210, 809–819. doi: 10.1534/genetics.118.301298

Bernal-Vasquez, A.-M., Gordillo, A., Schmidt, M., and Piepho, H.-P. (2017). Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. *BMC Genet.* 18:51. doi: 10.1186/s12863-017-0512-8

Biecek, P. (2018). Dalex: Explainers for complex predictive models in r. *J. Mach. Learn. Res.* 19, 3245–3249.

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, 2633–2635. doi: 10.1093/bioinformatics/btm308

Brock, J., Lange, M., Tratalos, J. A., More, S. J., Graham, D. A., Guelbenzu-Gonzalo, M., et al. (2021). Combining expert knowledge and machine-learning to classify herd types in livestock systems. *Sci. Rep.* 11, 1–10. doi: 10.1038/s41598-021-82373-3

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

Bustos-Korts, D., Boer, M. P., Malosetti, M., Chapman, S., Chenu, K., Zheng, B., et al. (2019). Combining crop growth modeling and statistical genetic modeling to evaluate phenotyping strategies. *Front. Plant Sci.* 10:1491. doi: 10.3389/fpls.2019.01491

Butler, E. E., and Huybers, P. (2015). Variations in the sensitivity of US maize yield to extreme temperatures by region and growth phase. *Environ. Res. Lett.* 10, 034009. doi: 10.1088/1748-9326/10/3/034009

Cakir, R. (2004). Effect of water stress at different development stages on vegetative and reproductive growth of corn. *Field Crops Res.* 89, 1–16. doi: 10.1016/j.fcr.2004.01.005

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*, 785–794.

Chenu, K. (2015). Characterising the crop environment – Nature, significance and applications. In: *Crop Physiology. Applications for Genetic Improvement and Agronomy*, eds Sadras V. and Calderini D. London: Elsevier, 321–348. doi: 10.1016/B978-0-12-417104-6.00013-3

Cicchino, M., Edreira, J. I. R., Uribelarrea, M., and Otegui, M. E. (2010). Heat stress in field-grown maize: response of physiological determinants of grain yield. *Crop Sci.* 50, 1438–1448. doi: 10.2135/cropsci2009.10.0574

Cooper, M., and DeLacy, I. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theor. Appl. Genet.* 88, 561–572. doi: 10.1007/BF01240919

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2020a). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity* 126, 92–106. doi: 10.1038/s41437-020-00353-1

Costa-Neto, G. M. F., Júnior, O. P. M., Heinemann, A. B., de Castro, A. P., and Duarte, J. B. (2020b). A novel gis-based tool to reveal spatial trends in reaction norm: upland rice case study. *Euphytica* 216, 1–16. doi: 10.1007/s10681-020-2573-4

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ. Res. Lett.* 13, 114003. doi: 10.1088/1748-9326/aae159

Crossa, J., Neto, R.-F., Montesinos-López, O. A., Costa-Neto, G. M. F., Dreisigacker, S., Montesinos-Lopez, A., et al. (2021). The modern plant breeding triangle: optimising the use of genomics, phenomics and enviromics data. *Front. Plant Sci.* 12:332. doi: 10.3389/fpls.2021.651480

Cuevas, J., Granato, I., Fritsche-Neto, R., Montesinos-Lopez, O. A., Burgueño, J., Bandeira e Sousa, M., et al. (2018). Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. *G3* 8, 1347–1365. doi: 10.1534/g3.117.300454

De Los Campos, G., Pérez-Rodríguez, P., Bogard, M., Gouache, D., and Crossa, J. (2020). A data-driven simulation platform to predict cultivars performances under uncertain weather conditions. *Nat. Commun.* 11, 1–10. doi: 10.1038/s41467-020-18480-y

Delerce, S., Dorado, H., Grillon, A., Rebolledo, M. C., Prager, S. D., Pati, V. H., et al. (2016). Assessing weather-yield relationships in rice at local scale using data mining approaches. *PLoS ONE* 11:e0161620. doi: 10.1371/journal.pone.0161620

Denmead, O., and Shaw, R. H. (1960). The effects of soil moisture stress at different stages of growth on the development and yield of corn 1. *Agron. J.* 52, 272–274. doi: 10.2134/agronj1960.00021962005200050010x

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *J. Anim. Ecol.* 77, 802–813. doi: 10.1111/j.1365-2656.2008.01390.x

Ersoz, E. S., Martin, N. F., and Stapleton, A. E. (2020). On to the next chapter for crop breeding: convergence with data science. *Crop Sci*. 60, 639–655. doi: 10.1002/csc2.20054

Estévez, J., Gavilán, P., and Giráldez, J. V. (2011). Guidelines on validation procedures for meteorological data from automatic weather stations. *J. Hydrol*. 402, 144–154. doi: 10.1016/j.jhydrol.2011.02.031

Falcon, C. M., Kaeppler, S. M., Spalding, E. P., Miller, N. D., Haase, N., AlKhalifah, N., et al. (2020). Relative utility of agronomic, phenological, and morphological traits for assessing genotype-by-environment interaction in maize inbreds. *Crop Sci*. 60, 62–81. doi: 10.1002/csc2.20035

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat*. 29, 1189–1232. doi: 10.1214/aos/1013203451

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V., and Müller, J. (2013). Random Forests modelling for the estimation of mango (*Mangifera indica* L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agric. Water Manage*. 116, 142–150. doi: 10.1016/j.agwat.2012.07.003

Gage, J. L., Jarquin, D., Romay, C., Lorenz, A., Buckler, E. S., Kaeppler, S., et al. (2017). The effect of artificial selection on phenotypic plasticity in maize. *Nat. Commun*. 8, 1–11. doi: 10.1038/s41467-017-01450-2

Géron, A. (2019). *Hands-on Machine Learning With Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

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

González-Recio, O., Jiménez-Montero, J., and Alenda, R. (2013). The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. *J. Dairy Sci*. 96, 614–624. doi: 10.3168/jds.2012-5630

Gräler, B., Pebesma, E., and Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *R J*. 8, 204–218. doi: 10.32614/RJ-2016-014

Greaves, J. A. (1996). Improving suboptimal temperature tolerance in maize-the search for variation. *J. Exp. Bot*. 47, 307–323. doi: 10.1093/jxb/47.3.307

Haley, C., and Visscher, P. (1998). Strategies to utilize marker-quantitative trait loci associations. *J. Dairy Sci*. 81, 85–97. doi: 10.3168/jds.S0022-0302(98)70157-2

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, 2nd Edn*. New York, NY: Springer.

Hatfield, J. L., Boote, K. J., Kimball, B., Ziska, L., Izaurralde, R. C., Ort, D., et al. (2011). Climate impacts on agriculture: implications for crop production. *Agron. J*. 103, 351–370. doi: 10.2134/agronj2010.0303

Hatfield, J. L., and Prueger, J. H. (2015). Temperature extremes: effect on plant growth and development. *Weather Climate Extremes* 10, 4–10. doi: 10.1016/j.wace.2015.08.001

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J.-L. (2014a). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet*. 127, 463–480.

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J.-L. (2014b). 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

Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., et al. (2014). Apsim-evolution towards a new generation of agricultural systems simulation. *Environ. Model. Softw*. 62, 327–350. doi: 10.1016/j.envsoft.2014.07.009

Hutter, F., Hoos, H., and Leyton-Brown, K. (2014). "An efficient approach for assessing hyperparameter importance," in *International Conference on Machine Learning* (PMLR), 754–762.

Jarquin, D., De Leon, N., Romay, M. C., Bohn, M. O., Buckler, E. S., Ciampitti, I. A., 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*. 11:1819. 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

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

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., et al. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE* 11:e0156571. doi: 10.1371/journal.pone.0156571

Juliana, P., Montesinos-López, O. A., Crossa, J., Mondal, S., Pérez, L. G., Poland, J., et al. (2019). Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. *Theor. Appl. Genet*. 132, 177–194. doi: 10.1007/s00122-018-3206-3

Kagawa, R., Kawazoe, Y., Ida, Y., Shinohara, E., Tanaka, K., Imai, T., et al. (2017). Development of type 2 diabetes mellitus phenotyping framework using expert knowledge and machine learning approach. *J. Diabetes Sci. Technol*. 11, 791–799. doi: 10.1177/1932296816681584

Kassambara, A., and Mundt, F. (2017). Package factoextra. Extract and visualize the results of multivariate data analyses 76.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "Lightgbm: a highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 30, 3146–3154.

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., et al. (2003). An overview of apsim, a model designed for farming systems simulation. *Eur. J. Agron*. 18, 267–288. doi: 10.1016/S.1161-0301(02)00108-9

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., and Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci. Rep*. 10, 1–12. doi: 10.1038/s41598-020-68771-z

Köppen, W., and Geiger, R. (1930). *Handbuch der Klimatologie, Vol. 1*. Gebrüder Borntraeger Berlin.

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling, Vol. 26*. New York, NY: Springer.

Kuhn, M., and Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. Available online at: https://www.tidymodels.org

Lampa, E., Lind, L., Lind, P. M., and Bornefalk-Hermansson, A. (2014). The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. *Environ. Health* 13:57. doi: 10.1186/1476-069X-13-57

Li, B., Zhang, N., Wang, Y.-G., George, A., 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

Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E., and Peng, B. (2019). Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the united states. *Glob. Chang Biol*. 25, 2325–2337. doi: 10.1111/gcb.14628

Lizaso, J., Ruiz-Ramos, M., Rodríguez, L., Gabaldon-Leal, C., Oliveira, J., Lorite, I., et al. (2018). Impact of high temperatures in maize: phenology and yield components. *Field Crops Res*. 216, 129–140. doi: 10.1016/j.fcr.2017.11.013

Lobell, D. B., Roberts, M. J., Schlenker, W., Braun, N., Little, B. B., Rejesus, R. M., et al. (2014). Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. *Science* 344, 516–519. doi: 10.1126/science.1251423

Malosetti, M., Bustos-Korts, D., Boer, M. P., and van Eeuwijk, F. A. (2016). Predicting responses in multiple environments: issues in relation to genotype environment interactions. *Crop Sci*. 56, 2210–2222. doi: 10.2135/cropsci2015.05.0311

Malosetti, M., Voltas, J., Romagosa, I., Ullrich, S., and Van Eeuwijk, F. (2004). Mixed models including environmental covariables for studying qtl by environment interaction. *Euphytica* 137, 139–145. doi: 10.1023/B:EUPH.0000040511.46388.ef

McFarland, B. A., AlKhalifah, N., Bohn, M., Bubert, J., Buckler, E. S., Ciampitti, I., et al. (2020). Maize genomes to fields (g2f): 2014-2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. *BMC Res. Notes* 13, 1–6. doi: 10.1186/s13104-020-4922-8

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

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

Mimić, G., Brdar, S., Brkić, M., Panić, M., Marko, O., and Crnojević, V. (2020). engineering meteorological features to select stress tolerant hybrids in maize. *Sci. Rep*. 10, 1–10. doi: 10.1038/s41598-020-60366-y

Moisen, G. G., Freeman, E. A., Blackard, J. A., Frescino, T. S., Zimmermann, N. E., and Edwards Jr, T. C. (2006). Predicting tree species presence and basal area in utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecol. Modell*. 199, 176–187. doi: 10.1016/j.ecolmodel.2006.05.021

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* 5, 2383–2390. doi: 10.1534/g3.115.021667

Monteverde, E., Gutierrez, L., Blanco, P., Prez de Vida, F., Rosas, J. E., Bonnecarrre, V., et al. (2019). Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (*Oryza sativa L.*) grown in subtropical areas. *G3* 9, 1519–1531. doi: 10.1534/g3.119.400064

Mushore, T., Manatsa, D., Pedzisai, E., Muzenda-Mudavanhu, C., Mushore, W., and Kudzotsa, I. (2017). Investigating the implications of meteorological indicators of seasonal rainfall performance on maize yield in a rain-fed agricultural system: case study of mt. darwin district in zimbabwe. *Theor. Appl. Climatol*. 129, 1167–1173. doi: 10.1007/s00704-016-1838-2

Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T., Gramatica, P., et al. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: the report and recommendations of ecvam workshop 52. *Alternat. Lab. Anim*. 33, 155–173. doi: 10.1177/026119290503300209

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 Proc*. 5, 1–5. doi: 10.1186/1753-6561-5-S3-S11

Olivoto, T., Nardino, M., Carvalho, I., Follmann, D., Ferrari, M., Szareski, V., et al. (2017). Reml/blup and sequential path analysis in estimating genotypic values and interrelationships among simple maize grain yield-related traits. *Genet. Mol. Res*. 16, 1–19. doi: 10.4238/gmr16019525

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers Geosci.* 30, 683–691. doi: 10.1016/j.cageo.2004.03.012

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-Rodríguez, P., Crossa, J., Bondalapati, K., De Meyer, G., Pita, F., and de los Campos, G. (2015). A pedigree-based reaction norm model for prediction of cotton yield in multienvironment trials. *Crop Sci*. 55, 1143–1151. doi: 10.2135/cropsci2014.08.0577

Pérez-Rodríguez, P., Crossa, J., Rutkoski, J., Poland, J., Singh, R., Legarra, A., et al. (2017). Single-step genomic and pedigree genotype× environment interaction models for predicting wheat lines in international environments. *Plant Genome* 10:plantgenome2016-09. doi: 10.3835/plantgenome2016.09.0089

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

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rahmstorf, S., Foster, G., and Cazenave, A. (2012). Comparing climate projections to observations up to 2011. *Environ. Res. Lett*. 7, 044035. doi: 10.1088/1748-9326/7/4/044035

Ridgeway, G. (2007). Generalized boosted models: a guide to the gbm package. *Update Univ S C Dep Music. 1, 2007*.

Rincent, R., Kuhn, E., Monod, H., Oury, F.-X., Rousset, M., Allard, V., et al. (2017). Optimization of multi-environment trials for genomic selection based on crop models. *Theor. Appl. Genet*. 130, 1735–1752. doi: 10.1007/s00122-017-2922-4

Rincent, R., Malosetti, M., Ababaei, B., Touzy, G., Mini, A., Bogard, M., et al. (2019). Using crop growth model stress covariates and ammi decomposition to better predict genotype-by-environment interactions. *Theor. Appl. Genet*. 132, 3399–3411. doi: 10.1007/s00122-019-03432-y

Ritchie, S. W., Hanway, J. J., Benson, G. O., Herman, J. C., and Lupkes, S. J. (1993). *How a Corn Plant Develops. Iowa State University Cooperative*. Extension Special report 48.

Roe, K. D., Jawa, V., Zhang, X., Chute, C. G., Epstein, J. A., Matelsky, J., et al. (2020). Feature engineering with clinical expert knowledge: a case study assessment of machine learning model complexity and performance. *PLoS ONE* 15:e0231300. doi: 10.1371/journal.pone.0231300

Rogers, A. R., Dunne, J. C., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I. A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3*. 11:jkaa050. doi: 10.1093/g3journal/jkaa050

Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G., and Hugot, J.-P. (2019). Comparative performances of machine learning methods for classifying crohn disease patients using genome-wide genotyping data. *Sci. Rep*. 9, 1–18. doi: 10.1038/s41598-019-46649-z

Schlenker, W., and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proc. Natl. Acad. Scie. U.S.A*. 106, 15594–15598. doi: 10.1073/pnas.0906865106

Shahhosseini, M., Hu, G., and Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. *Front. Plant Sci*. 11:1120. doi: 10.3389/fpls.2020.01120

Snoek, J., Larochelle, H., and Adams, R. P. (2012). "Practical bayesian optimization of machine learning algorithms," in *Proceedings of the 25th International Conference on Neural Information Processing Systems, Vol. 2, NIPS'12* (Red Hook, NY: Curran Associates Inc.), 2951–2959.

Sparks, A. (2018). nasapower: a nasa power global meteorology, surface solar energy and climatology data client for r. *J. Open Source Softw*. 3:1035. doi: 10.21105/joss.01035

Sukumaran, S., Crossa, J., Jarquín, D., and Reynolds, M. (2017). Pedigree-based prediction models with genotype× environment interaction in multienvironment trials of cimmyt wheat. *Crop Sci*. 57, 1865–1880. doi: 10.2135/cropsci2016.06.0558

Sukumaran, S., Jarquin, D., Crossa, J., and Reynolds, M. (2018). Genomic-enabled prediction accuracies increased by modeling genotype× environment interaction in durum wheat. *Plant Genome* 11, 1–11. doi: 10.3835/plantgenome2017.12.0112

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., and Bennett, M. (2017). Plant phenomics, from sensors to knowledge. *Curr. Biol.* 27, R770–R783. doi: 10.1016/j.cub.2017.05.055

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

Tiezzi, F., de Los Campos, G., Gaddis, K. P., and Maltecca, C. (2017). Genotype by environment (climate) interaction improves genomic prediction for production traits in us holstein cattle. *J. Dairy Sci*. 100, 2042–2056. doi: 10.3168/jds.2016-11543

Trnka, M., Rtter, R. P., Ruiz-Ramos, M., Kersebaum, K. C., Olesen, J. E., Ealud, Z., et al. (2014). Adverse weather conditions for european wheat production will become more frequent with climate change. *Nat. Clim. Chang* 4, 637–643. doi: 10.1038/nclimate2242

Troy, T. J., Kipgen, C., and Pal, I. (2015). The impact of climate extremes and irrigation on us crop yields. *Environ. Res. Lett*. 10:054013. doi: 10.1088/1748-9326/10/5/054013

van Eeuwijk, F. A., Denis, J. B., and Kang, M. S. (1996). "Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables." in *Genotype-by-Environment Interaction,* eds M. S. Kang and H. G. Gauch (Boca Raton, FL: CRC Press Inc.), 15–50.

Van Rijn, J. N., and Hutter, F. (2018). "Hyperparameter importance across datasets," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2367–2376.

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7:91. doi: 10.1186/1471-2105-7-91

Welch, J. R., Vincent, J. R., Auffhammer, M., Moya, P. F., Dobermann, A., and Dawe, D. (2010). Rice yields in tropical/subtropical asia exhibit large but opposing sensitivities to minimum and maximum temperatures. *Proc. Natl. Acad. Sci. U.S.A*. 107, 14562–14567. doi: 10.1073/pnas.1001222107

Williams, C. K., and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning,* Vol. 2. Cambridge, MA: MIT Press.

Yu, J., Shi, S., Zhang, F., Chen, G., and Cao, M. (2019). Predgly: predicting lysine glycation sites for homo sapiens based on xgboost feature optimization. *Bioinformatics* 35, 2749–2756. doi: 10.1093/bioinformatics/bty1043

Zahumenský, I. (2004). *Guidelines on Quality Control Procedures for Data From Automatic Weather Stations*. World Meteorological Organization.

Zhu, P., Zhuang, Q., Archontoulis, S. V., Bernacchi, C., and Müller, C. (2019). Dissecting the nonlinear response of maize yield to high temperature stress with model-data integration. *Glob. Chang Biol.* 25, 2470–2484. doi: 10.1111/gcb.14632

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., et al. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci.* 11:25. doi: 10.3389/fpls.2020.00025

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.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.

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 Plant Science

METHODS
published: 19 November 2021
doi: 10.3389/fpls.2021.734512

Comparing Genomic Prediction Models by Means of Cross Validation

Matías F. Schrauf[1,2]*, *Gustavo de los Campos*[3] *and Sebastián Munilla*[1,4]

[1] Facultad de Agronomía, Universidad de Buenos Aires, Buenos Aires, Argentina, [2] Animal Breeding & Genomics, Wageningen Livestock Research, Wageningen University & Research, Wageningen, Netherlands, [3] Departments of Epidemiology, Biostatistics, Statistics, and Probabilty, Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI, United States, [4] Instituto de Investigaciones en Producción Animal (INPA), CONICET-Universidad de Buenos Aires, Buenos Aires, Argentina

In the two decades of continuous development of genomic selection, a great variety of models have been proposed to make predictions from the information available in dense marker panels. Besides deciding which particular model to use, practitioners also need to make many minor choices for those parameters in the model which are not typically estimated by the data (so called "hyper-parameters"). When the focus is placed on predictions, most of these decisions are made in a direction sought to optimize predictive accuracy. Here we discuss and illustrate using publicly available crop datasets the use of cross validation to make many such decisions. In particular, we emphasize the importance of paired comparisons to achieve high power in the comparison between candidate models, as well as the need to define notions of relevance in the difference between their performances. Regarding the latter, we borrow the idea of equivalence margins from clinical research and introduce new statistical tests. We conclude that most hyper-parameters can be learnt from the data by either minimizing REML or by using weakly-informative priors, with good predictive results. In particular, the default options in a popular software are generally competitive with the optimal values. With regard to the performance assessments themselves, we conclude that the paired k-fold cross validation is a generally applicable and statistically powerful methodology to assess differences in model accuracies. Coupled with the definition of equivalence margins based on expected genetic gain, it becomes a useful tool for breeders.

Keywords: genomic selection, cross validation, plant breeding, genomic models, model selection

OPEN ACCESS

Edited by:
Fernando H. Toledo,
International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Michael Benjamin Kantar,
University of Hawaii, United States
Manish Kumar Pandey,
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India

****Correspondence:***
Matías F. Schrauf
matiasfschrauf@agro.uba.ar

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *01 July 2021*
Accepted: *26 October 2021*
Published: *19 November 2021*

Citation:
Schrauf MF, de los Campos G and Munilla S (2021) Comparing Genomic Prediction Models by Means of Cross Validation. Front. Plant Sci. 12:734512. doi: 10.3389/fpls.2021.734512

1. INTRODUCTION

In essence, genomic models relate genotypic variation as present in dense marker panels to phenotypic variation in a given population. These models were first introduced in breeding (Meuwissen et al., 2001) as a change of paradigm with respect to traditional marker assisted selection. They are currently used to accelerate genetic gain in many plant breeding programs with the focus placed on improving predictive ability while remaining agnostic to the causative nature of the genotype-phenotype relation. When fitting genomic models to data, practitioners need to make multiple decisions, sometimes without a clear guide or approach on how to take them. Besides the decision of choosing which model to use among the increasing number available (Whittaker et al., 2000; Meuwissen et al., 2001; VanRaden, 2008; de Los Campos et al., 2010; Habier et al., 2011; Ober et al., 2015), the practitioners also need to make many minor choices for those parameters which are not directly estimated by the data (so called "hyper-parameters"). When the focus is placed on predictions, as it is usual with genomic models, most of these decisions are made in a direction sought to optimize predictive accuracy. This accuracy is usually estimated in practice by means of cross validations.

Because of the impact of the prediction accuracy on genetic gain, many benchmarks have been done seeking to compare such accuracies among competing models. Most conclude that there is no better model in general (Heslot et al., 2012), with the recommendation that practitioners evaluate the entertained models with their own data and for the specific prediction tasks at hand (Azodi et al., 2019). The present work illustrates how the different performance assessments and comparisons can be made with cross validations, with a focus placed on both identifying differences of practical relevance and the decision making required for model selection and hyper-parameter tuning. We emphasize the importance of conducting paired cross validations to achieve higher statistical power, and propose the use of equivalence margins to identify the differences in accuracy which are relevant in practice.

With these goals in mind, the present work is organized as follows: we first assess the predictive ability of G-BLUP (VanRaden, 2008), probably the most known genomic model, in a well studied dataset, where we discuss the general aspects of cross validation. We then move on to the comparison of predictive abilities, which we first use to select the model complexity of BayesA by tuning the prior average variance of marker effects. We then consider general hyper-parameter tuning and evaluate the impact each hyper-parameter has on the accuracy for a variety of models. We explore general model comparisons, and describe tools to identify relevant differences in accuracy. To show an assessment of accuracy differences across multiple datasets we explore whether a pattern observed in the previous section can be generally extrapolated. We close with some final remarks.

2. MATERIALS AND METHODS

2.1. The Datasets

In the present work we used public datasets from three main crops: wheat, rice and maize. The first dataset consists of 599 CIMMYT wheat lines, genotyped with 1,279 DArT markers. The wheat lines were grown in four different environments and grain yield was recorded for each line and each environment (Crossa et al., 2010). This dataset is easily accessed from the R package BGLR (Perez and de los Campos, 2014) and its relatively small size allowed us to assess a greater number of models and parameter combinations.

The remaining two datasets include both more lines and genotyping by sequencing. They were included in the last section 'Comparison across datasets'. The rice dataset consists of 1,946 lines, which were genotyped by the 3,000 Rice Genomes Project (Wang et al., 2018). We used four quantitative traits available on a high number of lines: grain weight, width and length and the date on which 80% of the plants are heading. Finally, the maize dataset consists of lines from the "282" Association Panel and the NAM population. These lines were genotyped by the project "Biology of Rare Alleles in Maize and its Wild Relatives" (Glaubitz et al., 2014). For these lines we used four contrasting traits: the germination count, the number of leaves, the days to tassel, and plant height.

2.2. The Genomic Models

In the current work we assessed the performance of a variety of statistical models coming from two families of common use in genomic selection. The first family of models we considered is the so-called "Bayesian alphabet" (Gianola et al., 2009) and consists of regressions of phenotypes on markers. The second family comprises models that use the markers to build genomic relationship matrices (GRM), used in turn to model the covariance among genetic effects. These latter models stem from the linear mixed models tradition in breeding, which can be traced back to Henderson (cf. Henderson, 1984).

Models of the first family, the Bayesian alphabet, are usually formulated in a hierarchical structure of the form:

$$
\begin{aligned}
y &= \mu + X\beta + \varepsilon \\
\beta &\sim F(\Theta)
\end{aligned}
$$

where y is an n-length vector of trait phenotypes, μ is the vector of means (possibly dependent on fixed-effect predictors), X is an incidence matrix of the marker effects in the p-length vector β, and ε is an n-length vector of normally distributed errors (with environmental and unmodelled effects confounded). As the number of markers (p) typically exceeds the number of different genotypes (n), the regression equation is over-parameterized. Bayesian alphabet models deal with this "$n \ll p$" situation by assuming a prior distribution F(Θ) for the marker effects. Each model is distinguished by the distribution of such priors, which we briefly describe in **Box 1**.

Note that, after Gianola et al. (2009), it is usual to marginalize the marker effect distribution over all other marker-specific

BOX 1 | Priors of marker effects in models of the "Bayesian alphabet" used in this work.

rrBLUP: $\beta_j \sim$ normal distribution
$\beta_j \sim \mathcal{N}(0, \sigma^2_\beta)$
see Whittaker et al. (2000),

BayesA: $\beta_j \sim$ scaled t-student distribution
$\beta_j|\sigma^2_{\beta_j} \sim \mathcal{N}(0, \sigma^2_{\beta_j})$
$\sigma^2_{\beta_j} \sim$ Scaled-inv-$\chi^2(\nu, S)$
see Meuwissen et al. (2001),

BayesB: $\beta_j \sim$ spike-slab with scaled t-student distribution
$\beta_j|\sigma^2_{\beta_j} \sim \mathcal{N}(0, \sigma^2_{\beta_j})$
$\sigma^2_{\beta_j} = 0$, with probability π
$\sigma^2_{\beta_j}$ ~Scaled-inv-$\chi^2(\nu, S)$, with probability $(1 - \pi)$
see Meuwissen et al. (2001),

BayesC: $\beta_j \sim$ spike-slab with normal distribution
$\beta_j = 0$, with probability π
$\beta_j \sim \mathcal{N}(0, \sigma^2_\beta)$, with probability $(1 - \pi)$
see Habier et al. (2011).

parameters in the prior. As an example, by marginalizing over the marker-specific variance ($\sigma^2_{\beta_j}$), BayesA is usually characterized as having a scaled t-student distribution for the markers effects priors. Also in the literature, priors with a mass probability at zero are called spike-slab (like those used in BayesB and BayesC). In this work we do not interpret these Bayesian priors as statements of belief, but rather as regularization devices (Gelman and Shalizi, 2013). They stabilize estimates and predictions by making fitted models less sensitive to certain details of the data, and thus alleviate the over-parameterization problem in genomic models.

The second family of models considered consists of mixed linear models, where marker information is used to build-up relationship matrices. All these models may be specified as follows:

$$
\begin{aligned}
&y = X\beta + \sum_i Z^{(i)}u^{(i)} + \varepsilon \\
&u^{(i)} \sim \mathcal{N}(0, G^{(i)}\sigma^2_{u^{(i)}}) \\
&\varepsilon \sim \mathcal{N}(0, I\sigma^2_e)
\end{aligned}
$$

where y is an n-length vector of trait phenotypes, X is an incidence matrix of the fixed effects in β, each $Z^{(i)}$ is an incidence matrix of the individual genetic values in the n-length vector $u^{(i)}$, and ε is an n-length vector of errors (with environmental and unmodelled effects confounded). Each model is distinguished by different (often one, possibly many) genomic relationship matrices $[G^{(i)}]$ described in **Box 2**. These genomic relationship matrices (GRMs) specify the covariance structure of the genetic values.

BOX 2 | Genomic relationship matrices and the mixed models which use them.

G-BLUP:
$G \propto (M - 2 \cdot 1P)(M - 2 \cdot 1P)'$
see VanRaden (2008),

EG-BLUP:
$H \propto G \odot G$, (where $\odot$ is the hadamard product)
see Ober et al. (2015), Martini et al. (2016),

Categorical Epistasis:
$Cm_{ij} \propto \frac{1}{p} \cdot \sum_k [M_{ik} = M_{jk}]$, (where $[proposition] := 1$ if true, else 0)
$Ce \propto \frac{1}{2}(Cm \odot Cm + Cm)$
see Martini et al. (2017),

Gaussian Kernel:
$D_{ij} = \frac{1}{p} \cdot \sum_k |M_{ik} - M + jk|^2$ (alternatively, $D_{ij} = G_{ii} + G_{jj} - 2G_{ij}$)
$K_{ij} \propto exp(D_{ij}/h)$, (elementwise exponentiation)
see de Los Campos et al. (2010) and Alves et al. (2019),

Symbols:
M_{ik}: allele incidence matrix
P: allele frequencies
GRMs are defined up to a multiplicative constant, which can be absorbed into the corresponding variance parameter (σ^2_g) in the mixed model.

2.3. Cross Validations for Model Assessment

In this work we used k-fold cross validation in order to assess each model's predictive performance (cf. Friedman et al., 2001). This procedure consists of dividing a dataset with n cases (including both phenotypes and genotypic information) into a number of folds (k) of approximately equal size. Data in k-1 folds are used for training the model to predict phenotypes in the remaining fold (the testing fold), given the realized genotypes. The prediction task is repeated using one fold at a time for testing, and overall results are then combined. When the partitioning into folds is repeated, say r times, the procedure is called an r-replicated k-fold cross validation.

An important aspect in the design of a cross validation test is to define an appropriate error measure to be minimized. In this regard, a reasonable choice would be the mean square error (MSE), which penalizes every departure in predictions from the observed values. However, in the context of breeding this measure can be too strict, as any constant or scaling factor afflicting all predictions will inflate the MSE but will not change the ranking. Instead, breeders have focused on estimating the predictive accuracy (accuracy, for short), measured as the correlation between predictions and observations.

In practice, genetic values are usually the ultimate prediction targets rather than phenotypes. To account for this, the accuracy can be re-scaled dividing by the square root of a heritability estimate (notice it is important to use the same heritability estimate for all accuracies compared to each other). It is possible, though, to go one step further and directly focus on estimating the expected genetic gain, which is easily obtained if we assume truncation selection. We used this new re-scaling into expected genetic gain in the section "Comparison across datasets" (in results and discussion). The scaling factor can be easily derived from the standard genetic gain formula (cf. Falconer and Mackay, 1996, in "Response to selection"):

$$
\begin{aligned}
\Delta G &= i_q \cdot r_g \cdot \sigma_G \\
\Delta G &= i_q \cdot (r_{ph}/h) \cdot \sigma_G \\
\Delta G &= i_q \cdot r_{ph} \cdot (\sigma_P/\sigma_G) \cdot \sigma_G \\
\Delta G &= i_q \cdot r_{ph} \cdot \sigma_P \\
\Delta G/\sigma_P &= i_q \cdot r_{ph}
\end{aligned}
$$

where r_g is the predictive accuracy with respect to the (unobserved) true genetic values, r_{ph} is the predictive accuracy with respect to phenotypes, i_q is the selection intensity (i.e., the mean of a standardized Normal distribution truncated at the q selection quantile), and $\Delta G/\sigma_P$ is the estimate of genetic gain (in phenotypic standard deviations). This genetic gain measure is quite simplistic (as it assumes selection by truncation and random mating), but on the other hand is easily interpretable and of practical relevance.

There are two further important issues with regard to cross validations. One concerns the partitioning between training and testing sets. While here we always used random partitions, in specific cases it can be more appropriate to use other schemes such as splitting the dataset in generations, half-sib families

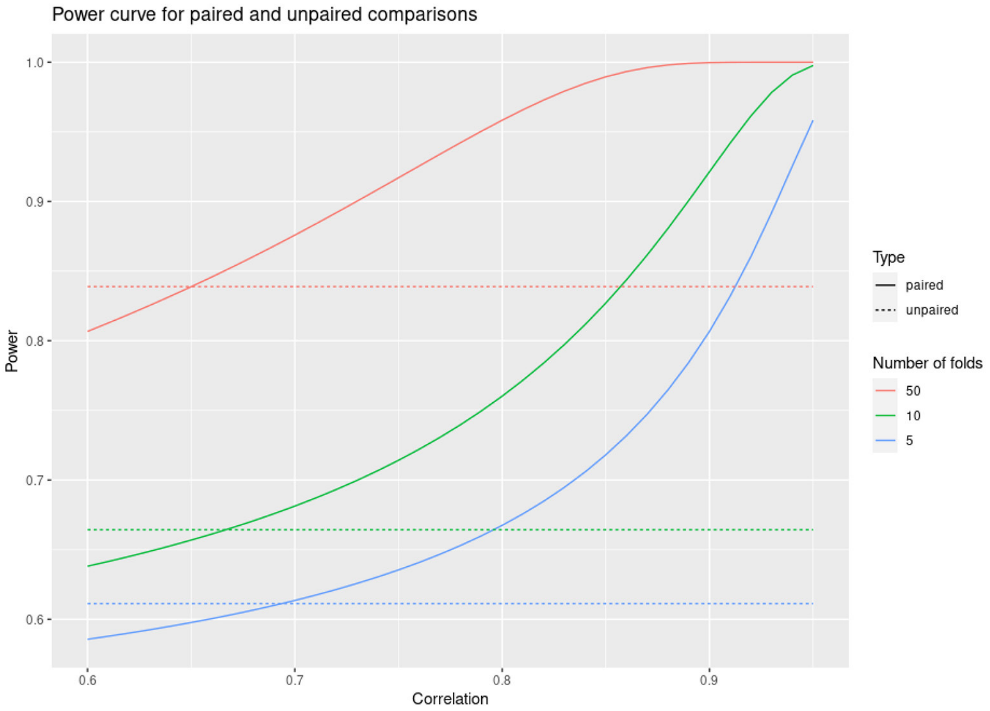


FIGURE 1 | Power curve for paired and unpaired cross validations as a function of fold correlation between compared models.

or sub-populations. Also, in the context of multi-trait models, available information about the different traits can vary between selection candidates at the time of prediction; thus blurring the distinction between training and testing sets (Runcie and Cheng, 2019).

The second issue concerns whether we are interested in estimating the accuracy conditional on the available training set or as a marginal expectation; i.e., averaged over different possible training sets. In the context of a breeding plan, where the genomic model gets updated with new data, the marginal predictive accuracy might be the more appropriate. Fortunately, this is the version of the accuracy which is thought to be better estimated by a k-fold cross validation, while a leave-one-out cross validation might be better tailored to estimate the conditional predictive accuracy (cf. Friedman et al., 2001, section 7.12). In summary, the cross validation should be designed to accurately simulate the real-world usage of the genomic model.

2.4. Paired Cross Validations for Model Comparison

The problem of identifying a superior model is different from the performance assessment task such as discussed in the previous section. While one could conduct model selection simply by choosing the model with the highest estimated performance, it is important to take the variability of those estimates into account, as well as to provide some control for error probabilities according to statistical established practice. When applying an r-replicated k-fold cross validation procedure, variability in the performance estimates arises from the r replicates and the k folds. However, using the variability estimate of each assessment independently (surprisingly an extended practice) ignores that most variability is shared among models.

A much more reasonable approach when comparing predictive accuracies between models is to perform paired comparisons within the same partitioning of folds (Hothorn et al., 2005). That is, for each fold one summarizes the difference in accuracies between the compared models rather than the individual accuracies. This often results in a huge reduction in the variance of the performance estimates, because most of the variability is usually shared across the different models. For example, if the correlation across folds of the accuracy scores for two models is over 0.8, then the variance of the estimate of the accuracy difference can be reduced five times by taking this approach, with a corresponding increase in statistical power (see **Figure 1**). We employed this approach in all our model comparisons.

2.5. Equivalence, Non-inferiority and Superiority Tests

The comparison of model accuracies using paired differences of cross validation can have high statistical power. This allows detecting with high confidence very small differences in performance. Such statistically significant differences of small magnitude can be uninteresting because they are superseded by considerations other than accuracy, or they might not be robust to any changes in the application of the models. As the saying goes, *"With great power there must also come great responsibility"*. Here it is the responsibility of practitioners to evaluate the differences, not only by the statistical ability to detect them, but also by their assessed practical relevance.

To help with this task, we propose defining an “equivalence margin” [-Δ, Δ] within which model performances are deemed equivalent in practice. These kinds of equivalence margins are standard used in clinical studies (e.g., Da Silva et al., 2009) but, to the best of our knowledge, their use is not widespread in plant breeding or the agricultural and environmental sciences in general. Then, in addition to the conventional test for statistical differences (sd)

- $H_0 : d = 0$,

we use the machinery of statistical tests to provide assertions on the practical relevance of these differences with some degree of error control. Specifically, by conducting tests of

- Equivalence (eq), $H_0 : |d| > \Delta$
- Non-inferiority (noi), $H_0 : d < -\Delta$
- Superiority (sup), $H_0 : d < \Delta$

The hypothesis for these tests are illustrated in **Box 3**.

BOX 3 | Representation of the null (H_0) and alternative (H_1) hypothesis for specific tests:

```
sd:   <-------)[](--------->
eq:   <----](-----)[------->
noi:  <----](-------------->
sup:  <------------](------>
      <----|---+---|------->
          -Δ   0  +Δ
```

Null hypotheses represented in gray, alternative hypotheses in black.

We can use these tests to assess the practical relevance of differences in predictive accuracy. With the result of these tests we can produce labels similar to the “significance letters”, which we argue have some advantages with regard to their interpretation:

- Equivalence letters: models sharing the same letter have an accuracy difference confidently within the equivalence margin (and thus are deemed equivalent for practical purposes).
- Non-inferiority ranking: models with the same or higher ordinal are confidently non-inferior (the accuracy difference is within or above the equivalence margin).
- Superiority ranking: models with higher ordinal are confidently superior (the accuracy difference is above the equivalence margin).

To build these labels we use directed graphs where the nodes are the models compared and they are connected by an edge if the null hypothesis for the comparison is rejected.

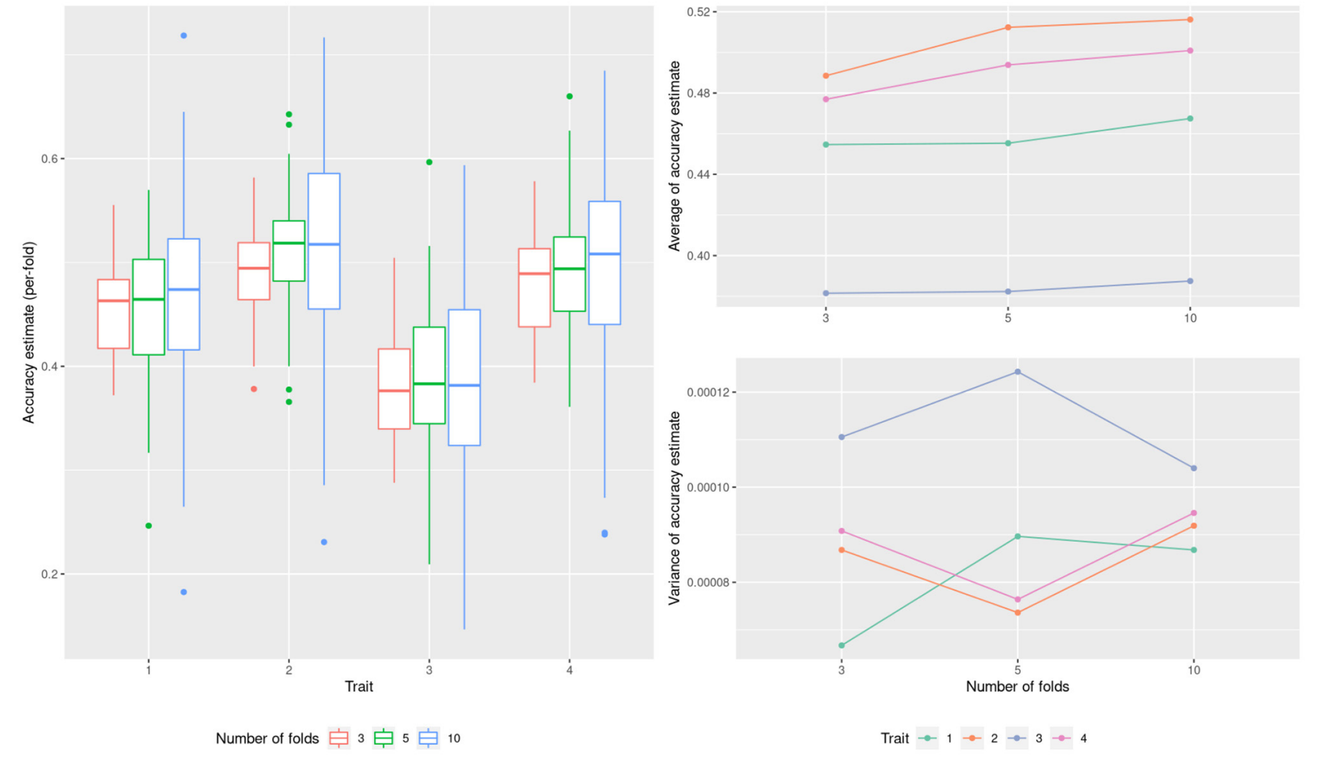


FIGURE 2 | Model performance estimation for the wheat dataset with varying number of cross-validation folds.

- Equivalence letters: One letter is assigned to each clique of the graph (which is effectively an undirected graph due to the reflexivity of the equivalence test).
- Non-inferiority and Superiority rankings: The rankings are built from the consensus ordering of all possible topological orders for their respective directed graphs.

These algorithms are similar in nature to those used by statistical software to compute the traditional significance letters. We note, though, that traditional significance letters should not be interpreted as meaning that elements with the same letter are equivalent which, instead, is the correct interpretation for the equivalence letters built with the construction above. Finally, we would like to mention that the hypothesis tests covered in this section have a general scope of application and are not restricted to the comparisons of model performance.

2.6. Software

The GRMs were built with custom code in the Julia programming language (Bezanson et al., 2017), available upon request from the corresponding author. The remaining analyses were done in the R programming language (R Core Team, 2021). In particular, the "Bayesian alphabet" models were fitted with the BGLR package (Perez and de los Campos, 2014) and the mixed models were fitted with the EMMREML package (Akdemir and Godfrey, 2015). We used the bootstrap utilities from the package "boot" (Davison and Hinkley, 1997; Canty and Ripley, 2021). Finally, the functions for the analysis of cross validation results and equivalence margin testing were organized into the R package "AccuracyComparer" (available at https://github.com/schrauf/AccuracyComparer).

3. RESULTS AND DISCUSSION

3.1. Model Predictive Ability Assessment

As a starting point and to illustrate the use of the cross validation technique we estimate the ability of a G-BLUP model to predict CIMMYT wheat yield across four environments. **Figure 2** shows the accuracies estimated by the K-fold cross validation when using different numbers of folds (K = 3, 5 and 10). The means bias downward for a smaller number of folds (panel b) but the effect is small. For the variance of the estimate there is no clear tendency (panel c). This is because of two competing effects that balanced out. For one, as the size of the testing set increases (less folds), this reduces the variance of the estimate at each fold (panel a). In the opposite direction, as the number of folds increases, the variance of the whole cross validation estimate reduces. To estimate the marginal predictive error, both 5-fold and 10-fold seem reasonable choices, with smaller bias than 3-fold cross validation and similar variances. As briefly mentioned in materials and methods, a greater number of folds should not be used unless the goal is to estimate the conditional

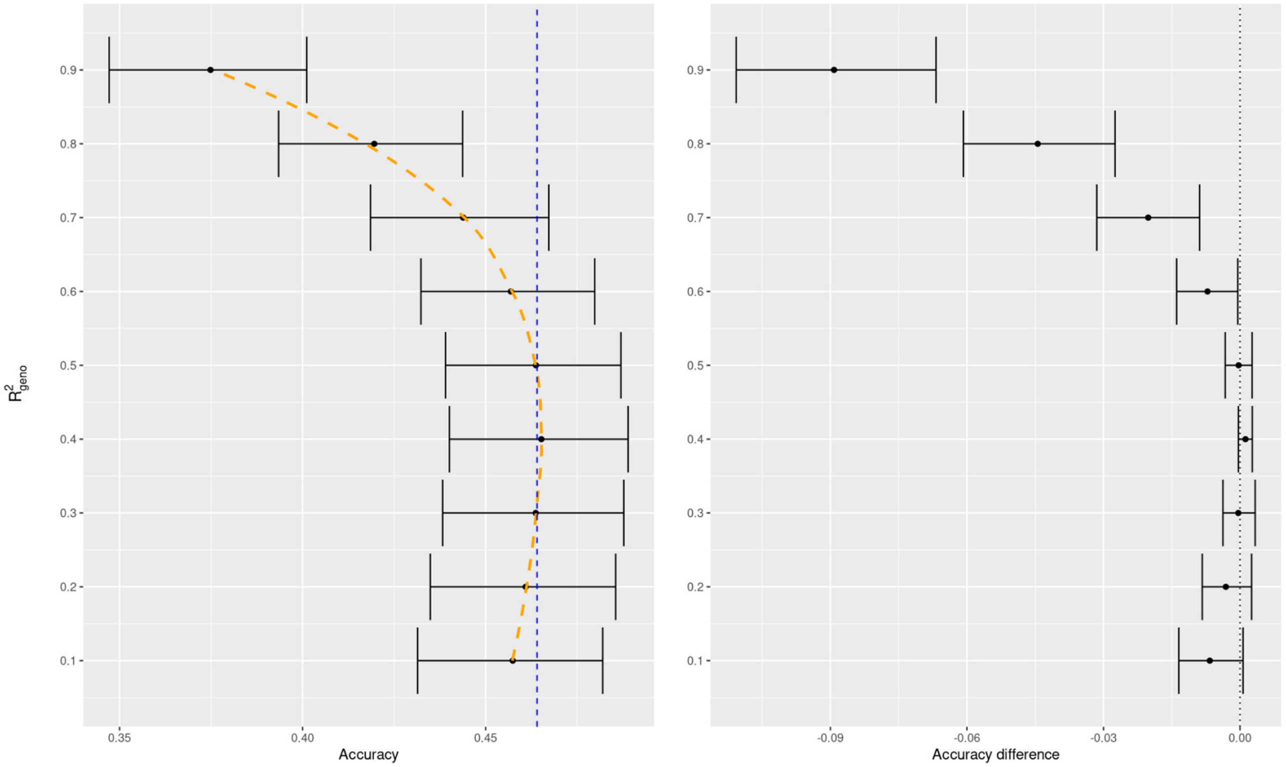


FIGURE 3 | BayesA predictive accuracy as function of prior mean of R^2_{geno} for trait 1 of the wheat dataset. Average predictive accuracy of the default model in BGLR (a weakly-informative prior for R^2_{geno}) in dashed blue for the left panel. All accuracy differences in the right panel are taken with respect to the default model.

predictive accuracy. In all the following sections we used 10-fold cross validations.

3.2. Model Selection

3.2.1. Model Complexity and Penalization Parameter

Most genomic models have some penalization parameters which regulate how flexibly the model adjusts to sample observations. Finding an optimal value for these parameters is a typical task for cross validation. Alternatively, these penalization parameters can be learnt from the data by either minimizing the REML criterion in mixed models (where the penalization parameters are variance components, see Bates et al., 2014) or by using non or weakly-informative priors in Bayesian alphabet models.

As an example, take the case of BayesA, where model penalization is mainly controlled by the scale parameter of the chi-squared distribution (S, in **Box 1**), which in turn determines the a priori average variance of the marker effects ($\mathbb{E}[\sigma^2_{\beta_j}]$). With BGLR we can choose the value of this parameter by specifying the proportion of phenotypic variance a-priori expected to be explained by the marker effects (in the following "R^2_{geno}" see Perez and de los Campos, 2014), which allows for an easier interpretation.

To illustrate the usefulness of cross validation to elicit these parameters, we conducted a 10-fold cross validation for BayesA with a grid of values for R^2_{geno} when fitting wheat yield data. From this we can observe a textbook accuracy curve which results from the classical bias-variance tradeoff (cf. Friedman et al., 2001). Starting with low values of R^2_{geno} we have rigid models, whose accuracies improve with increasing R^2_{geno}, until the models begin to overfit and the accuracy rapidly deteriorates (**Figure 3**, left panel). This resulted in an intermediate optimal value.

In addition, we compared the difference in accuracies of the specific variance proportions in a model with a weakly-informative prior which is the default in BGLR (**Figure 2**, right panel). We can see that the model which learns the variance proportion from the data performs competitively with the best pre-specified values of R^2_{geno}. We know that REML is a sound criterion for learning variance components (Thompson, 2019) and known theoretical results match REML estimates to the mode of the posterior distribution of the parameter when a non-informative prior is set in a Bayesian model (cf. Sorensen and Gianola, 2002, chapter 9). It is possible then, that the soundness of REML applies not only to Bayesian mixed models but also, at least approximately, to other Bayesian regressions when using weakly informative priors.

3.2.2. Hyper-Parameter Tuning

Beside penalization parameters, there are many hyper-parameters without a clear impact on accuracy in the priors of Bayesian regressions. On the other hand, mixed linear models have fewer ones, a notable exception being the bandwidth of the Gaussian kernel. Here we have summarized the impact of

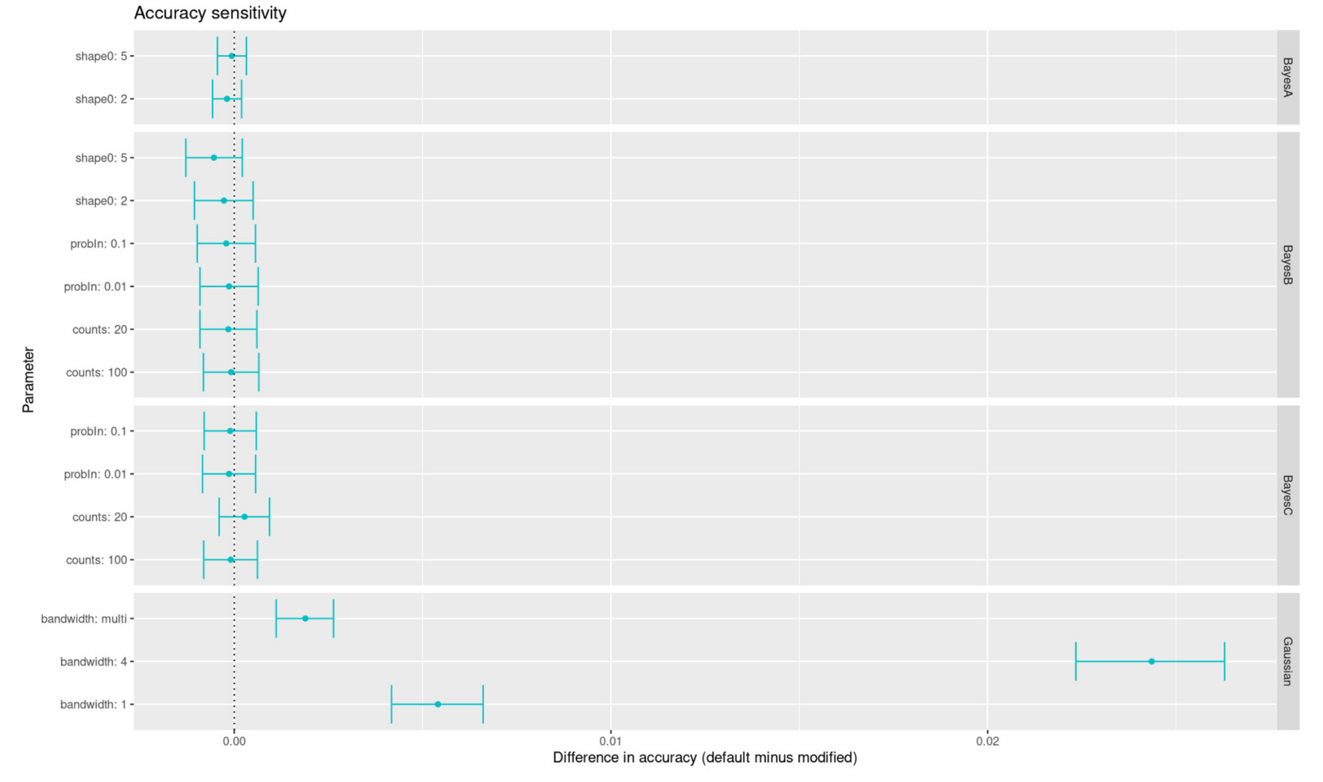


FIGURE 4 | Accuracy sensitivity of multiple models to changes in the values of their hyper-parameters (alternative values compared with respect to defaults in BGLR software).

many of those hyper-parameters affecting the models when fitting the wheat yield data. For each parameter we show the change in accuracy with respect to the default value in BGLR (**Figure 4**). We can see that, of these parameters, only changes in the kernel's bandwidth impact the accuracy with a statistically significant change. An alternative to arbitrary choices is to use multiple kernels in the same model, each kernel with a different bandwidth (Alves et al., 2019). These multi-kernel models have the ability to weight the contribution of each kernel, with results close to optimal.

3.2.3. General Model Comparison

Beyond setting hyper-parameters, which in our exploration resulted in minor changes in accuracy, the practitioners may want to compare between distinctly different models. Here we see how one may proceed to compare between more than two models, when they are not necessarily organized by a specific parameter. We used clustering to help interpretation, and we chose the G-BLUP as a reference model to compare accuracy differences (**Figure 5**). Still referring to wheat yield, BayesB performed the worst and the Gaussian kernel methods

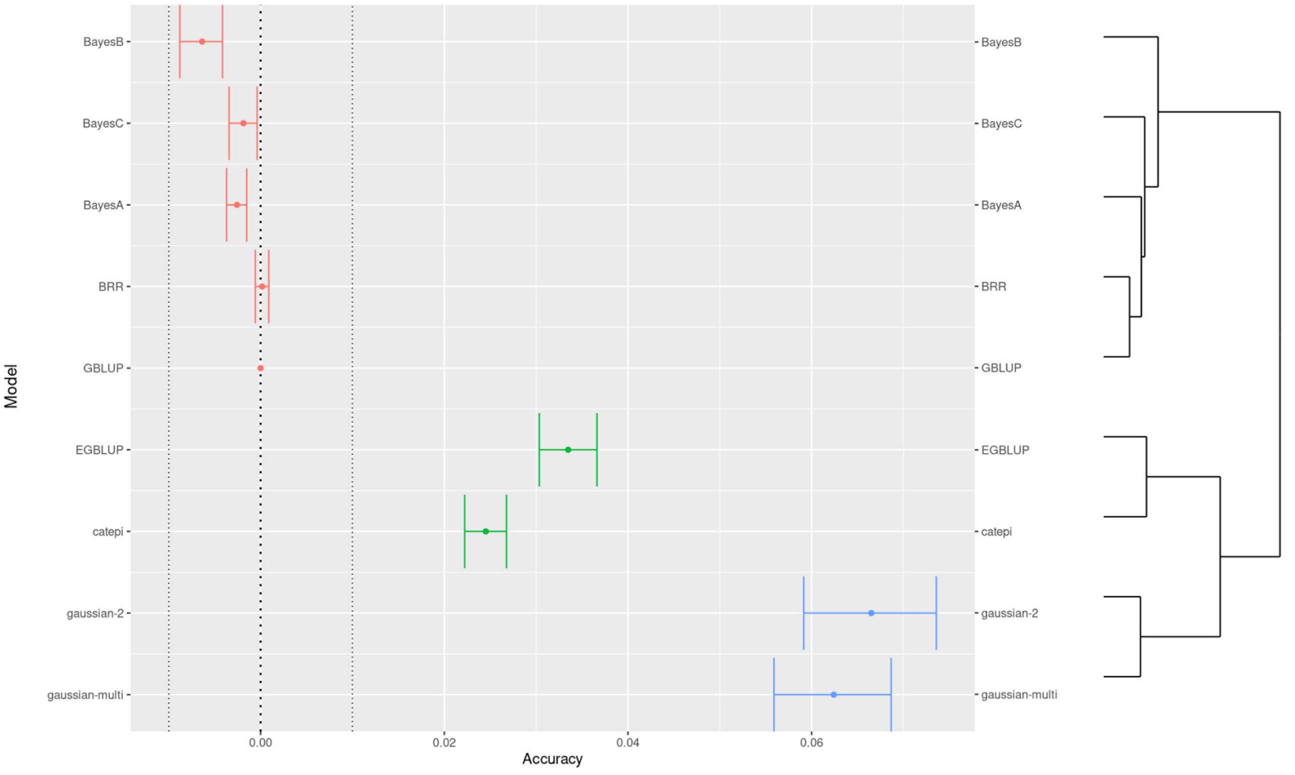


FIGURE 5 | Comparison of predictive accuracies across models for the trait 1 of the wheat dataset. Colors show a 3-group clustering of the models based on the accuracy with a hierarchical clustering on the right side.

TABLE 1 | General model comparison for the wheat dataset.

	Accuracy difference			Hypothesis tests		
Model	**Mean**	**Lower**	**Upper**	**sd**	**eq**	**Sup**
BayesB	–0.006	–0.009	–0.004	a	A	1
BayesC	–0.002	-0.003	0.000	b	A	1
BayesA	–0.003	-0.004	–0.002	b	A	1
GBLUP	0.000	-	-	c	A	1
BRR	0.000	–0.001	0.001	c	A	1
Catepi	0.025	0.022	0.027	d	B	2
EGBLUP	0.033	0.030	0.037	e	B	2
Gaussian-multi	0.062	0.056	0.069	f	C	3
Gaussian-2	0.067	0.059	0.074	f	C	3

Accuracy differences measured with respect to the G-BLUP model. Hypothesis tests (sd, statistical differences; eq, equivalence; sup, superiority). Meaning of letters and rankings in section 2.5.

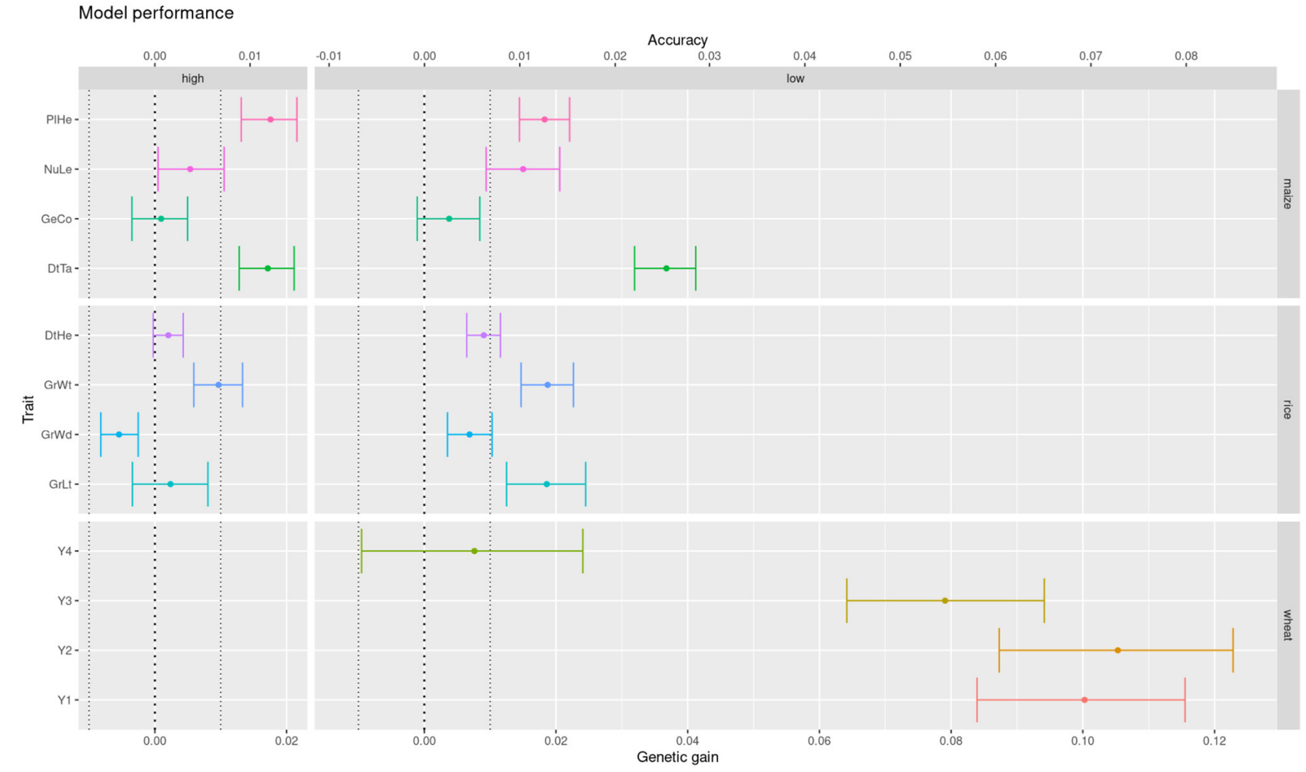


FIGURE 6 | Difference in expected genetic gain from using a Gaussian kernel with respect to a GBLUP prediction model, for low and high panel densities, across multiple crops and traits (color coded for legibility).

the best. Because of high power, we defined an "equivalence margin" to identify the relevant differences. This allowed us to identify easily interpreted groups of statistically equivalent models (**Table 1**). Concretely, the 3 equivalent groups were the additive models (A), models with only pairwise interactions between markers (B), and finally the models with higher order interactions (C). We explore further this relation between marker interactions and predictive performance in the following section.

3.2.4. Comparison Across Datasets

Sometimes we need to compare models across different datasets or prediction tasks. For instance, we would like to see here if the difference between additive and epistatic models observed in the previous section is particular to the wheat dataset. Schrauf et al. (2020) showed that marker density could be a relevant factor for the advantage of models with marker interactions. Recall that wheat lines were genotyped at low density with DaRTs, whereas rice and maize lines were sequenced. So we compared the models performance across datasets with both low and high density marker panels from these latter species. Also, to assess the relevance of the differences in accuracy, we converted them to differences in expected genetic gain (assuming truncating selection of the highest 10% genetic values). This scale could help practitioners in deciding on relevant equivalence margins for the equivalence, non-inferiority and superiority hypothesis tests.

We can see that the advantage for the Gaussian kernel over the GBLUP model observed for wheat in the previous section is much less clear for the maize and rice datasets (**Figure 6**). Further, the improvements that can be observed are reduced when going from a low density marker panel to a high density one. In particular, the traits where the models were statistically equivalent rose from under 10% with low density panels to half at high density panels (**Table 2**). These results are in accordance with the phenomena of phantom epistasis (Schrauf et al., 2020).

3.3. Final Remarks

In the present work we explored a variety of aspects related to the performance assessment of genomic models via cross validations. We identified several strategies which can help practitioners avoid arbitrary decisions when implementing a particular genomic prediction model. For instance, many hyper-parameters can be effectively learnt from the data by either minimizing REML or by using weakly-informative priors. In particular, the default values of those hyper-parameters in the software used (BGLR) are generally competitive with the optimal values. An exception is the choice of bandwidth in a gaussian kernel, for which different values can result in qualitatively different predictive performances of the model. For this particular case we recommended the use of multi-kernel models.

TABLE 2 | Difference in expected genetic gain from using a Gaussian kernel with respect to a GBLUP prediction model, for low and high panel densities.

			Genetic gain difference			Hypothesis tests			
Markers	**Species**	**Trait**	**Mean**	**Lower**	**Upper**	**sd**	**eq**	**noi**	**Sup**
High	Maize	Days to tassel	0.017	0.013	0.021	*		*	*
		Germination count	0.001	–0.003	0.005		*	*	
		Number of leaves	0.005	0.000	0.011	*		*	
		Plant height	0.018	0.013	0.022	*		*	*
	Rice	Grain length	0.002	–0.003	0.008		*	*	
		Grain width	–0.005	–0.008	–0.003	*	*	*	
		Grain weight	0.010	0.006	0.013	*		*	
		Days to heading	0.002	0.000	0.004		*	*	
						0.63	0.5	1	0.25
Low	Maize	Days to tassel	0.037	0.032	0.041	*		*	*
		Germination count	0.004	–0.001	0.008		*	*	
		Number of Leaves	0.015	0.009	0.021	*		*	
		Plant height	0.018	0.014	0.022	*		*	*
	Rice	Grain length	0.019	0.012	0.025	*		*	*
		Grain width	0.007	0.004	0.010	*		*	
		Grain weight	0.019	0.015	0.023	*		*	*
		Days to heading	0.009	0.006	0.012	*		*	
	Wheat	Yield 1	0.100	0.084	0.116	*		*	*
		Yield 2	0.105	0.087	0.123	*		*	*
		Yield 3	0.079	0.064	0.094	*		*	*
		Yield 4	0.008	-0.010	0.024			*	
						0.83	0.08	1	0.58

Hypothesis tests (sd, statistical differences; eq, equivalence; noi, non-inferiority; sup, superiority), and proportion of models with rejected nulls for high and low panel densities. Asterisks indicate rejected nulls for the corresponding test (see ***Box 3****).*

Throughout the work we used paired cross validations to compare methods. This was motivated by the fact that cross validation estimates are greatly correlated between models. While the cross validation estimate of the performance of a model can have a high variability, the estimate of the difference in performance between two models is usually much more precise and allows for their comparison with higher statistical power. We concluded that paired k-fold cross validations result in a generally applicable and statistically powerful methodology to assess differences in model accuracies.

Finally, we introduced the idea of equivalence margins as a means to identify when those significant differences have practical relevance for decision making and model selection. This is important because with high statistical power small differences become detectable, which might not be of interest, or might not be robust to even small changes between the validation and the application of the models. We suggest to couple the tool of equivalence margins, and the associated hypothesis tests, with informative performance scales for the tasks at hand. In a breeding context, such scale could be the potential genetic gain from truncation selection.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.panzea.org/; https://doi.org/10.7910/DVN/HGRSJG; https://cran.r-project.org/web/packages/BGLR/index.html.

AUTHOR CONTRIBUTIONS

MS carried out the analyses and wrote the manuscript with support from SM. GC provided key ideas and revised intermediate stages. All authors discussed the results and contributed to the final manuscript.

FUNDING

This publication resulted (in part) from research supported by multiple institutions: MS received funding from the National Scientific and Technical Research Council (CONICET) in the form of a doctoral scholarship. GC received funding from the National Institute for Food and Agriculture (NIFA) of the USDA (award #2021-67015-33413). SM received funding from UBACyT 2020 (proyect: 20020190200324BA).

ACKNOWLEDGMENTS

We would like to thank Pablo Inchausti (UdelaR, Uruguay) and Federico Gómez (FAUBA, UBA, Argentina) for valuable discussions about statistical inference of effect sizes and relevant differences.

REFERENCES

Akdemir, D., and Godfrey, O. U. (2015). EMMREML: fitting mixed models with known covariance structures. *R package version* 3.1.

Alves, F. C., Granato, Í. S. C., Galli, G., Lyra, D. H., Fritsche-Neto, R., and de Los Campos, G. (2019). Bayesian analysis and prediction of hybrid performance. *Plant Methods* 15, 1–18. doi: 10.1186/s13007-019-0388-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* 9, 3691–3702. doi: 10.1534/g3.119.400498

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*. doi: 10.18637/jss.v067.i01

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev*. 59, 65–98. doi: 10.1137/141000671

Canty, A., and Ripley, B. D. (2021). boot: Bootstrap R (S-Plus) Functions. *R package version* 1.3–28.

Crossa, J., Campos, G., d. l., Pérez, P., Gianola, D., Burgueno J, Luis Araus J, 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

Da Silva, G. T., Logan, B. R., and Klein, J. P. (2009). Methods for equivalence and noninferiority testing. *Biol. Blood Marrow Transplant*. 15, 120–127. doi: 10.1016/j.bbmt.2008.10.004

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

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

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

Friedman, J., Hastie, T., Tibshirani, R. (2001). *The Elements of Statistical Learning, Vol. 1*. New York, NY: Springer Series in Statistics New York.

Gelman, A., and Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *Br. J. Math. Stat. Psychol*. 66, 8–38. doi: 10.1111/j.2044-8317.2011.02037.x

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

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

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12, 1–12. doi: 10.1186/1471-2105-12-186

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph, ON: University of Guelph.

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

Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. (2005). The design and analysis of benchmark experiments. *J. Comput. Graph. Stat*. 14, 675–699. doi: 10.1198/106186005X59630

Martini, J. W., Gao, N., Cardoso, D. F., Wimmer, V., Erbe, M., Cantet, R. J., et al. (2017). Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended gblup and properties of the categorical epistasis model (ce). *BMC Bioinformatics* 18, 1–16. doi: 10.1186/s12859-016-1439-1

Martini, J. W., Wimmer, V., Erbe, M., and Simianer, H. (2016). Epistasis and covariance: how gene interaction translates into genomic relationship. *Theor. Appl. Genet*. 129, 963–976. doi: 10.1007/s00122-016-2675-5

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

Ober, U., Huang, W., Magwire, M., Schlather, M., Simianer, H., and Mackay, T. F. (2015). Accounting for genetic architecture improves sequence based genomic prediction for a drosophila fitness trait. *PLoS ONE* 10:e0126880. doi: 10.1371/journal.pone.0126880

Perez, 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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

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

Schrauf, M. F., Martini, J. W., Simianer, H., de Los Campos, G., Cantet, R., Freudenthal, J., et al. (2020). Phantom epistasis in genomic selection: on the predictive ability of epistatic models. *G3* 10, 3137–3145. doi: 10.1534/g3.120.401300

Sorensen, D., and Gianola, D. (2002). *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. New York, NY: Springer.

Thompson, R. (2019). Desert island papers—a life in variance parameter and quantitative genetic parameter estimation reviewed using 16 papers. *J. Anim. Breed. Genet*. 136, 230–242. doi: 10.1111/jbg.12400

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

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of asian cultivated rice. *Nature* 557, 43–49. doi: 10.1038/s41586-018-0063-9

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

Author Disclaimer: The contents are solely the responsibility of the authors and do not necessarily represent the official views of the funding institutions.

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 Plant Science

OPINION
published: 19 November 2021
doi: 10.3389/fpls.2021.728567

A Comparison of the Adoption of Genomic Selection Across Different Breeding Institutions

Mahmood Gholami[1]*, Valentin Wimmer[1], Carolina Sansaloni[2], Cesar Petroli[2], Sarah J. Hearne[2,3], Giovanny Covarrubias-Pazaran[3], Stefan Rensing[4], Johannes Heise[4], Paulino Pérez-Rodríguez[5], Susanne Dreisigacker[6], José Crossa[2,5] and Johannes W. R. Martini[2]*

[1] KWS SAAT SE & Co. KGaA, Einbeck, Germany, [2] Genetic Resources Program, International Maize and Wheat Improvement Center, Texcoco, Mexico, [3] Excellence in Breeding Platform, Consultative Group for International Agricultural Research, Texcoco, Mexico, [4] IT Solutions for Animal Production (vit - Vereinigte Informationssysteme Tierhaltung w.V.), Verden, Germany, [5] Department of Statistics, Colegio de Postgraduados, Montecillos, Mexico, [6] Global Wheat Program, International Maize and Wheat Improvement Center, Texcoco, Mexico

Keywords: genomic selection, breeding, plant breeding, animal breeding, dairy breeding, selection gain

OPEN ACCESS

Edited by:
Lee Hickey,
The University of Queensland, Australia

Reviewed by:
Owen Powell,
University of Queensland, Australia
Gregor Gorjanc,
University of Edinburgh, United Kingdom

***Correspondence:**
Mahmood Gholami
gholami.mahmoud@gmail.com
Johannes W. R. Martini
jwrmartini@gmail.com

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

***Received:** 21 June 2021*
***Accepted:** 30 September 2021*
***Published:** 19 November 2021*

Citation:
Gholami M, Wimmer V, Sansaloni C, Petroli C, Hearne SJ, Covarrubias-Pazaran G, Rensing S, Heise J, Pérez-Rodríguez P, Dreisigacker S, Crossa J and Martini JWR (2021) A Comparison of the Adoption of Genomic Selection Across Different Breeding Institutions. Front. Plant Sci. 12:728567. doi: 10.3389/fpls.2021.728567

INTRODUCTION

Within the last 20 years, after the landmark paper by Meuwissen et al. (2001), genomic selection (GS) has been widely incorporated in plant and animal breeding (Crossa et al., 2017; Hickey et al., 2017). However, adoption happened at different speeds and with distinct focus.

Here, we give a short description of the history and the current state of GS implementation in German dairy cattle breeding (as an example in animal breeding), at the private plant breeding company KWS SAAT SE & Co. KGaA, and at the public breeding programs of the International Maize and Wheat Improvement Center (CIMMYT) and the Consultative Group for International Agricultural Research (CGIAR) in general. We close by highlighting some differences in organizational structure and objectives of the considered breeding institutions, and comment on how these differences may have influenced the adoption of GS.

GENOMIC SELECTION IN DAIRY CATTLE BREEDING

Dairy cattle breeding provided good conditions for the introduction of GS. Selection decisions had been based for decades purely on additive genetic effects reflected in a sire's breeding value, and the use of pedigree-based estimated breeding values (PEBVs) had already been common practice. However, reliabilities of early estimated breeding values from information on parents only were low. Therefore, a testing scheme was used, in which bulls were mated to a more or less representative sample of cows in a first step. The resulting daughters were then raised until their performance could be measured, thus improving the reliabilities of their sires' breeding values. Only then, the best test bulls were selected and used broadly. This costly waiting period led to a generation interval of more than five years. Using genomically estimated breeding values (GEBVs) of young bulls, which are more reliable than PEBVs, permitted to reduce this waiting period, and thus to increase selection gain per time. Although the accuracy of the breeding value of a bull which has been extensively progeny tested over years is of higher accuracy than a young bull's GEBV, the costs in terms of waiting time do not pay off for the breeding program, when comparing a more accurate late selection to a less accurate early selection based on the GEBV instead of the PEBV.

With this setup, genomic breeding values for Holsteins and Jerseys were first published in the USA in 2009 (Wiggans et al., 2017), about a decade after the release of the first commercial SNP chip (Wang et al., 1998). In Europe, four breeding organizations (UNCEIA: France; VikingGenetics: Denmark, Finland, and Sweden; DHV-vit: Germany; CRV: The Netherlands, Flanders) joined forces and put a reference population together with 4,000 bulls each (Lund et al., 2011). After 1.5 years of development, from August 2010 onwards, genomic breeding values, based on the joint reference population, were published in these European countries. This rapid evolution was only possible due to a long-established international data infrastructure with Multiple Across Country Evaluations (MACE) being in place since the 1990s at the international evaluation center Interbull. MACE allows the expression and use of estimated breeding values on the scale of each participating country (Schaeffer, 1994). Since 2010, breeding progress has more than doubled for all traits in German Holsteins as seen from **Figure 1**, mostly due to the sharply decreased generation interval for bulls.

The initial 50k Illumina SNP set is still the reference SNP set for genomic evaluations at vit in Germany, although dozens of different SNP chips have been integrated since then, especially many low density chips. With dropping genotyping costs and low density 10k SNP chips, female animals came also into the focus. In 2019, cows were integrated in the German reference population. As of the routine genetic evaluation in April 2021, there were 43,699 bulls and 249,363 cows in the reference population for milk traits. Current efforts aim at implementing Single Step methodology (Aguilar et al., 2010) in the genetic evaluation systems of most countries, which is a computationally demanding task with big populations, requiring specialized algorithms (e.g., Liu et al., 2014).

GENOMIC SELECTION AT KWS

Around 2008, KWS started own research activities in the field of GS and participated in several large collaborations (e.g., Albrecht et al., 2011; Hofheinz et al., 2012). Only a few years later, GS became an established part of the breeder's toolbox for all KWS field crops.

The reason for this rapid adoption of GS is its attractiveness for addressing several components of the breeder's equation simultaneously: Shorten the breeding cycle by replacing phenotypic evaluation steps through a genomic evaluation, increasing accuracy by integrating information from relatives and multiple environments, and increasing selection intensity in case that genotyping is cheaper than phenotyping.

Advances in genome analysis of major crops over the past 15 years led to the availability of a vast number of molecular markers, a pre-requisite for GS application. New genotyping technologies reduced costs of genotyping to a fraction of the costs of phenotyping an individual in field trials.

As a consequence of these developments, GS influenced the design of breeding schemes. With this tool at hand, predictive breeding is used to plan crosses, to reduce breeding cycle length, and to select for more stable performance using multi-year training sets. Genomic prediction is now practiced on many complex traits including yield, quality, biotic, and abiotic stress.

For instance in sugar beet breeding, GS has become an essential component to address the trait "sugar yield," which is a composite of "sugar content" and "yield." These two traits are addressed by both (i) within cycle prediction, which allows higher selection intensity and (ii) across cycle prediction, which allows early selection. Predictive ability in each breeding program is constantly monitored. Besides routine application, KWS does very active research to further enhance the efficiency of this tool. Two factors have been the focus of genomic prediction

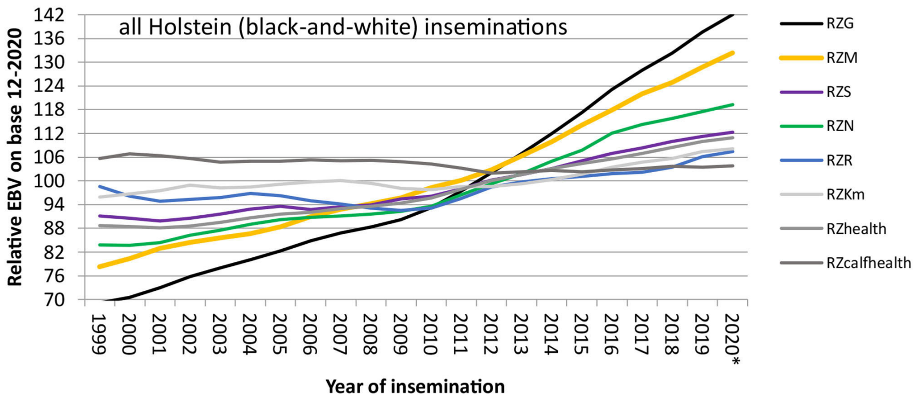


FIGURE 1 | Breeding progress in important traits in German Holsteins, measured as yearly mean EBVs of bulls, weighted by the number of inseminations with their semen. The label "RZ" denotes that all breeding values are standardized to a genetic standard deviation of 12, and a mean of 100 in the female base population (year of birth 2014–2016), all breeding values are expressed such that more positive values are more desirable from the breeder's perspective. RZG, total merit index; RZM, milk production index; RZS, somatic cell score; RZN, longevity; RZR, fertility index; RZKm, index of maternal calving traits; RZhealth, health trait index; RZcalfhealth, calf survival. *Year 2020: incomplete data. Slightly modified from IT Solutions for Animal Production (vit - IT Solutions for Animal Production, 2021).

research: chip design and size and composition of training sets. For instance, for sugar beet, we saw that approximately 2,000 markers are sufficient for genomic prediction, potentially due to high linkage disequilibrium in the breeding material. The required training set size is highly dependent of the relationship between training set and prediction set as well as the heritability of the trait. We observe a diminishing return on prediction accuracy for the phenotype of sugar yield when having more than 300 individuals in the training set (which may also be a consequence of the high linkage disequilibrium in breeding populations).

Today, GS has become a routine application in breeding programs at KWS. Thousands of GS analyses are performed every year. Therefore, KWS has optimized genotyping processes and analysis pipelines. With GS being implemented widely in all breeding programs, KWS is extending prediction methods using artificial intelligence and genotype by environment (GxE) interactions.

GENOMIC SELECTION AT THE INTERNATIONAL MAIZE AND WHEAT IMPROVEMENT CENTER (CIMMYT)

CIMMYT has started to explore GS more aggressively as a new breeding tool since 2010 (de los Campos et al., 2009; Crossa et al., 2010, 2019; Dreisigacker et al., 2021). The estimation of GEBVs for the germplasm is routinely implemented for the maize and the wheat program, but it is a decision of the respective breeder which weight is given to this information in the selection process. The initial focus of GS application has been on greater selection intensity in stage I yield trials by predicting the GEBVs of germplasm which had not been included in the trials. Recent projects aim to use GS for early selection and to shorten cycle time. Standardized workflows for data storage, processing, and subsequent analyses are currently advanced by the Excellence in Breeding (EiB) platform and various projects at CIMMYT and other CGIAR centers. CIMMYT has also worked on genomic prediction of traits of germplasm bank accessions (Crossa et al., 2016) to explore its potential for harnessing genetic resources (Martini et al., 2021). The center has built the basis for more informed screening of novel allelic diversity in the germplasm collection by genotyping a substantial part of the available accessions (Sansaloni et al., 2020).

The question which impact GS had on the annual genetic gain for yield across breeding pipelines is more difficult to answer than for the dairy cattle example presented above. Estimates of genetic gain vary and GS has been used to different extend across breeding pipelines. Since programs introduced GS gradually, it is difficult to separate a potential increase in genetic gain due to the use of GS, from other aspects which may have improved the breeding pipelines. A recent publication by Gerard et al. (2020) reports estimated yearly selection gains of 0.93% for low-rainfall environments and 3.8% for high-rainfall environments for the period of 2007–2016 for grain yield in wheat. However, we cannot clearly attribute the credit of this selection gain to GS, since this period is too short after GS has been implemented. However, several dedicated experiments in maize outlined the potential of GS. For instance, Beyene et al. (2015) used GS to select from bi-parental maize populations for yield under drought stress and reported a higher selection gain than for conventional breeding methods. Comparing to previous studies, the authors concluded that "the average gain observed under drought in our study using GS was two- to fourfolds higher than what has been reported from conventional phenotypic selection under drought stress." Moreover, CIMMYT's Global Maize Program designed a rapid cycle genomic selection (RCGS) of multi-parental crosses (Zhang et al., 2017). Two cycles per year were performed, and the authors found that "the genetic gains from the RCGS [...] are at the same or higher level than those observed in other studies under phenotypic selection [...]." Also, Beyene et al. (2019) compared selection gain of phenotypic selection (PS) and GS for two different environments (well-watered and water stressed) and observed a higher selection gain for PS for well-watered conditions, and a higher selection gain for GS under water stress. The authors highlighted that GS provides "the potential to bypass stage I trial evaluation and move material directly into stage II" which "would reduce both the costs and cycle time but will require accurate predictions from training sets composed of historical data" (Beyene et al., 2019). This potential to reduce cycle time has not yet been included in the study.

IMPLEMENTATION OF GENOMIC SELECTION CGIAR-WIDE

The CGIAR has entered a phase of pushing the application of GS for all crops, from maize to bananas (Nyine et al., 2017; Wolfe et al., 2017; Ahmadi et al., 2020; Gemenet et al., 2020; Atanda et al., 2021). The EiB platform provides technical assistance and practical guidelines for the implementation of GS and the modernization of breeding programs (see for instance Covarrubias-Pazaran et al., 2021). Before EiB, several initiatives advanced the use of GS in specific crops. For example, the NextGen Cassava project took important steps toward the successful implementation of GS for root, tuber, and banana (RTB) crops (Wolfe et al., 2017; Maxmen, 2019). Those steps included the development of a robust database system, matching the genotyping logistics with the growing season, and automating analytical pipelines. Similar steps have been taken by initiatives at IRRI and CIMMYT (Crossa et al., 2017; Gao et al., 2020).

Crops currently using GS to reduce cycle time are cassava and maize (Atanda et al., 2021; Esuma et al., 2021). Genomic selection is being used to increase selection intensity in cassava, maize, rice, and wheat (Ahmadi et al., 2020; Dreisigacker et al., 2021). Finally, GS is used for increasing the selection accuracy of yield trials by all the aforementioned and yams (Agre et al., 2018). Other crops, including beans, pulses, forages, bananas, and potato are developing and validating the necessary logistics and tools to manage the data, genotyping, analytical pipelines, and costs. This picture is rapidly changing since the ambition of all breeding programs in the CGIAR is to use genome-assisted prediction methodologies to reduce the length of the breeding cycle to 2–3 years.

CONCLUSION

Dairy Cattle Breeding Compared to Plant Breeding

Genomic selection was adopted in dairy cattle breeding almost instantly after genotyping costs dropped below the anticipated break-even point, presumably because the routine use of pedigree-based predictions, and a culture of centrally processing data of fragmented production units, had already been established (Schaeffer, 1994; Wiggans et al., 2017).

In contrast, plant breeding programs are traditionally dedicated to more specific geographical regions aiming to adapt the germplasm to certain environmental conditions, and the data used for selection decisions have almost exclusively focused on the most recent trials of the respective program. An overarching approach for handling data across programs or selection cycles had not been necessary. Moreover, pedigree information had hardly been used for pedigree-based predictions, since the pedigree information has often been incomplete and "relatively wide" crosses of unrelated material have been used (Dreisigacker et al., 2021). Moreover, a PEBV may not provide additional information, since it cannot capture the segregation within a family generated by a certain cross.

Also, plant breeders traditionally tend to focus on product development that is on identifying varieties, rather than on population improvement, that is identifying parents for new crosses. In other words, breeders are more interested in the genotypic value comprising the complete genetic contribution to the phenotype than in the additive genetic value (the breeding value). A focus on the latter is natural in dairy breeding, where the sire's breeding value is defined indirectly by the performance of its offspring, not by its own phenotype (Mrode, 2014).

Only in recent years some concepts from animal breeding, such as the focus on the breeding value, have been transferred in more formal and more rigorous ways to plant breeding. An example is the separation of population improvement from product development (Gaynor et al., 2017) which allows to focus on the breeding value for the population improvement step. The impact of this paradigm shift on genetic gain is to be observed in coming decade(s).

Public Compared to Private Plant Breeding

In general, the timelines for the exploration of the potential of GS were relatively similar between the considered public and private plant breeding organizations. CIMMYT and the CGIAR are public research organizations that also pursue the publication of novel, creative approaches, and follow in parts a (research) project-based organization. In contrast, private institutions naturally tend to focus more on the standardization and optimization of routine processes for GS, which may have had a lower priority in the public sector. The EiB platform and associated projects are currently addressing a stronger standardization of data storage and related analysis pipelines. Moreover, the project-based organization in public institutions comes with a variance in funding which leads to challenges for mid to long-term planning on the use of GS.

Finally, CGIAR centers are plant improvement-breeding centers that focus on delivering germplasm to National Agricultural Research institutions (NARs), in particular in Africa and Asia. This implies other priorities for traits, different frameworks for the evaluation of material, and different cost structures compared to, for instance, a commercial program in North America. The economics of implementing GS may therefore differ from those at private companies.

Overall, we think that the advent of GS has provided a tipping point to catalyze the ongoing reform of plant breeding institutions to data processing focused organizations. This transformation will leverage both the historic data resources amassed and the data generated annually to more effectively drive breeding decisions. However, with the increasing number of phenotypic records, and genotypic and environmental information, we now face the challenge of how to use "big data" most efficiently.

AUTHOR CONTRIBUTIONS

MG and VW wrote the section about GS at KWS SAAT SE & Co. KGaA. CS, CP, SJH, SD, PP-R, JC, and JM wrote the section on GS at CIMMYT. GC-P wrote the section about CGIAR-wide implementation of GS. Authors from vit (SR and JH) wrote the section about GS in dairy cattle breeding. All authors contributed to the conclusion. MG and JM organized the joint effort.

ACKNOWLEDGMENTS

The authors affiliated to CIMMYT and EiB thank the Bill and Melinda Gates Foundation for financial support for increasing genetic gains [INV-003439 BMGF/FCDO Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods (AGG)] 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 (F.F.L.) and the Agricultural Agreement Research Fund (J.A.) in Norway through NFR grant 267806, the CIMMYT CRP-WHEAT and the USDA National Institute of Food and Agriculture Grant Nos. 2020-67013-30904 and 2018-67015-27957 to DER and Hatch project 1010469.

REFERENCES

Agre, A. P., Beauchet, G., Dekoyer, D., Yusuf, M., Gisel, A., Abberton, M., et al. (2018). "Designing SNP-array for guinea yams (Dioscorea spp.) for routine use in breeding program," in *Plant and Animal Genome XXVI Conference* (San Diego, CA).

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93, 743–752. doi: 10.3168/jds.2009-2730

Ahmadi, N., Bartholomé, J., Tuong-Vi, C., and Grenier, C. (2020). "Genomic selection in rice: empirical results and implications for breeding," *Quantitative Genetics, Genomics and Plant Breeding, 2nd Edn*., ed M. Kang (Wallingford; Oxon: CABI Publishing), 243–258. doi: 10.1079/9781789240214.0243

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

Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., et al. (2021). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theor. Appl. Genet.* 134, 279–294. doi: 10.1007/s00122-020-03696-9

Beyene, Y., Gowda, M., Olsen, M., Robbins, K. R., Pérez-Rodríguez, P., Alvarado, G., et al. (2019). Empirical comparison of tropical maize hybrids selected through genomic and phenotypic selections. *Front. Plant Sci.* 10:1502. doi: 10.3389/fpls.2019.01502

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

Covarrubias-Pazaran, G., Martini, J. W. R., Quinn, M., and Atlin, G. (2021). Strengthening public breeding pipelines by emphasizing quantitative genetics principles and open source data management. *Front. Plant Sci.* 12:681624. doi: 10.3389/fpls.2021.681624

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., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., et al. (2016). Genomic prediction of gene bank wheat landraces. *G3 (Bethesda)*. 6, 1819–1834. doi: 10.1534/g3.116.029637

Crossa, J., Martini, J. W. R., Gianola, D., Pérez-Rodríguez, P., Jarquín, D., Juliana, P., et al. (2019). Deep kernel and deep learning for genome-based prediction of single traits in multienvironment breeding trials. *Front. Genet.* 10:1168. doi: 10.3389/fgene.2019.01168

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

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

Dreisigacker, S., Crossa, J., Pérez-Rodríguez, P., Montesinos-López, O. A., Rosyara, U., Juliana, P., et al. (2021). Implementation of genomic selection in the CIMMYT global wheat program, findings from the past 10 years. *Crop Breed. Genet. Genom.* 3:e210005. doi: 10.20900/cbgg20210005

Esuma, W., Ozimati, A., Kulakow, P., Gore, M. A., Wolfe, M. D., Nuwamanya, E., et al. (2021). Effectiveness of genomic selection for improving provitamin A carotenoid content and associated traits in cassava. *G3 (Bethesda)*. 11:jkab160. doi: 10.1093/g3journal/jkab160

Gao, S. Y., Hagen, T. J., Robbins, K., Jones, E., Karkkainen, M., Dreher, K. A., et al. (2020). "Transforming breeding through enterprise breeding system and analytics," in *Plant and Animal Genome XXVIII Conference* (San Diego, CA).

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

Gemenet, D. C., da Silva Pereira, G., De Boeck, B., Wood, J. C., Mollinari, M., Olukolu, B. A., et al. (2020). Quantitative trait loci and differential gene expression analyses reveal the genetic basis for negatively associated β-carotene and starch content in hexaploid sweetpotato [*Ipomoea batatas* (L.) Lam.]. *Theor. Appl. Genet.* 133, 23–36. doi: 10.1007/s00122-019-03437-7

Gerard, G. S., Crespo-Herrera, L. A., Crossa, J., Mondal, S., Velu, G., Juliana, P., et al. (2020). Grain yield genetic gains and changes in physiological related traits for CIMMYT's high rainfall wheat screening nursery tested across international environments. *Field Crops Res.* 249:107742. doi: 10.1016/j.fcr.2020.107742

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

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

Liu, Z., Goddard, M. E., Reinhardt, F., and Reents, R. (2014). A single-step genomic model with direct estimation of marker effects. *J. Dairy Sci.* 97, 5833–5850. doi: 10.3168/jds.2014-7924

Lund, M. S., Roos, A. P., Vries, A. G., Druet, T., Ducrocq, V., Fritz, S., et al. (2011). A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genet. Sel. Evol.* 43:43. doi: 10.1186/1297-9686-43-43

Martini, J. W. R., Molnar, T. L., Hearne, S., Crossa, J., and Pixley, K. V. (2021). Opportunities and challenges of predictive approaches for harnessing the potential of genetic resources. *Front. Plant Sci.* 12:674036. doi: 10.3389/fpls.2021.674036

Maxmen, A. (2019). How African scientists are improving cassava to help feed the world. *Nature* 565, 144–147. doi: 10.1038/d41586-019-00014-2

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

Mrode, R. A. (2014). *Linear Models for the Prediction of Animal Breeding Values*. Oxfordshire: CABI. doi: 10.1079/9781780643915.0000

Nyine, M., Uwimana, B., Swennen, R., Batte, M., Brown, A., Christelová, P., et al. (2017). Trait variation and genetic diversity in a banana genomic selection training population. *PLoS ONE* 12:e0178734. doi: 10.1371/journal.pone.0178734

Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., et al. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-020-18404-w

Schaeffer, L. (1994). Multiple-country comparison of dairy sires. *J. Dairy Sci.* 77, 2671–2678. doi: 10.3168/jds.S0022-0302(94)77209-X

vit - IT Solutions for Animal Production (2021). *Geschäftsbericht 2020/21*. Available online at: https://www.vit.de/fileadmin/Wir-sind-vit/Geschaeftsberichte/vit_GB_2020_2021.pdf (accessed October 13, 2021).

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., et al. (1998). Large-scale identification. mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082. doi: 10.1126/science.280.5366.1077

Wiggans, G. R., Cole, J. B., Hubbard, S. M., and Sonstegard, T. S. (2017). Genomic selection in dairy cattle: the USDA experience. *Annu. Rev. Anim. Biosci.* 5, 309–327. doi: 10.1146/annurev-animal-021815-111422

Wolfe, M. D., Del Carpio, D. P., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., Kayondo, I. S., et al. (2017). Prospects for genomic selection in cassava breeding. *Plant Genome* 10. doi: 10.3835/plantgenome2017.03.0015

Zhang, X., Pérez-Rodríguez, P., Burgueño, J., Olsen, M., Buckler, E., Atlin, G., et al. (2017). Rapid cycling genomic selection in a multiparental tropical maize population. *G3 (Bethesda)*. 7, 2315–2326. doi: 10.1534/g3.117.043141

Conflict of Interest: MG and VW are employed by the company KWS SAAT SE & Co. KGaA. Moreover, JH and SR are employed by IT Solutions for Animal Production (vit).

The remaining authors declare that the research was conducted in the absence of any commercial or financial interest that could be construed as potentially conflicting with the research.

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 Plant Science

ORIGINAL RESEARCH
published: 11 January 2022
doi: 10.3389/fpls.2021.718611

Response to Early Generation Genomic Selection for Yield in Wheat

David Bonnett[1,2]*, Yongle Li[3], Jose Crossa[1,4], Susanne Dreisigacker[1]*, Bhoja Basnet[1], Paulino Pérez-Rodríguez[4], G. Alvarado[1]†, J. L. Jannink[5,6]*, Jesse Poland[7] and Mark Sorrells[6]

[1] International Maize and Wheat Improvement Center, Texcoco, Mexico, [2] BASF Wheat Breeding, Sabin, MN, United States, [3] School of Agriculture, Food and Wine, Faculty of Sciences, The University of Adelaide, Adelaide, SA, Australia, [4] Colegio de Postgraduados, Texcoco, Mexico, [5] USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, United States, [6] Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, United States, [7] Department of Plant Pathology, Kansas State University, Manhattan, KS, United States

OPEN ACCESS

Edited by:
Diego Rubiales,
Institute for Sustainable Agriculture,
Spanish National Research Council
(CSIC), Spain

Reviewed by:
Arron Carter,
Washington State University,
United States
Valerio Hoyos-Villegas,
McGill University, Canada
Lance Merrick,
Washington State University Pullman,
United States, in collaboration with
reviewer AC

***Correspondence:**
David Bonnett
david.bonnett@basf.com
Susanne Dreisigacker
s.dreisigacker@cgiar.org
J. L. Jannink
jj332@cornell.edu

† To the memory of "Goyito"

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *01 June 2021*
Accepted: *22 October 2021*
Published: *11 January 2022*

Citation:
Bonnett D, Li Y, Crossa J, Dreisigacker S, Basnet B, Pérez-Rodríguez P, Alvarado G, Jannink JL, Poland J and Sorrells M (2022) Response to Early Generation Genomic Selection for Yield in Wheat. Front. Plant Sci. 12:718611. doi: 10.3389/fpls.2021.718611

We investigated increasing genetic gain for grain yield using early generation genomic selection (GS). A training set of 1,334 elite wheat breeding lines tested over three field seasons was used to generate Genomic Estimated Breeding Values (GEBVs) for grain yield under irrigated conditions applying markers and three different prediction methods: (1) Genomic Best Linear Unbiased Predictor (GBLUP), (2) GBLUP with the imputation of missing genotypic data by Ridge Regression BLUP (rrGBLUP_imp), and (3) Reproducing Kernel Hilbert Space (RKHS) a.k.a. Gaussian Kernel (GK). F2 GEBVs were generated for 1,924 individuals from 38 biparental cross populations between 21 parents selected from the training set. Results showed that F2 GEBVs from the different methods were not correlated. Experiment 1 consisted of selecting F2s with the highest average GEBVs and advancing them to form genomically selected bulks and make intercross populations aiming to combine favorable alleles for yield. F4:6 lines were derived from genomically selected bulks, intercrosses, and conventional breeding methods with similar numbers from each. Results of field-testing for Experiment 1 did not find any difference in yield with genomic compared to conventional selection. Experiment 2 compared the predictive ability of the different GEBV calculation methods in F2 using a set of single plant-derived F2:4 lines from randomly selected F2 plants. Grain yield results from Experiment 2 showed a significant positive correlation between observed yields of F2:4 lines and predicted yield GEBVs of F2 single plants from GK (the predictive ability of 0.248, $P < 0.001$) and GBLUP (0.195, $P < 0.01$) but no correlation with rrGBLUP_imp. Results demonstrate the potential for the application of GS in early generations of wheat breeding and the importance of using the appropriate statistical model for GEBV calculation, which may not be the same as the best model for inbreds.

Keywords: early generation genomic selection, linear and non-linear kernels genomic matrices, wheat breeding, breeding methodology, response to selection

INTRODUCTION

Genomic selection (GS) (Meuwissen et al., 2001; Bernardo and Yu, 2007) has become possible through the rapid development of next-generation sequencing technologies that allow the use of abundant and low-cost molecular markers. Evidence in plant breeding literature has shown that GS provides an important increase in prediction accuracy compared to pedigree and marker-assisted selection for low heritability traits (de los Campos et al., 2009, 2010, 2013;

Crossa et al., 2010, 2011, 2013, 2014; González-Camacho et al., 2012, 2016; Heslot et al., 2012, 2014; Hickey and Gorjanc, 2012; Pérez-Rodríguez et al., 2012; Riedelsheimer et al., 2012; Windhausen et al., 2012; Zhao et al., 2012). An initial review of the main activities of GS in the International Maize and Wheat Improvement Center (CIMMYT) maize and wheat breeding programs was published by Crossa et al. (2014). Simultaneously, breeding programs around the world have been studying GS, initially performing extensive research, and the development of new statistical models for incorporating pedigree, genomic, and environmental covariables (climatic and soil data). Models that incorporated genomic × environment and marker × environment and genomic × environmental covariables were earlier developed to improve the accuracy for predicting unobserved cultivars in new environments (Burgueño et al., 2012; Heslot et al., 2014; Jarquín et al., 2014; Lopez-Cruz et al., 2015; Crossa et al., 2016).

After these initial studies, an increasing number of research articles have been published effectively testing the integration of GS into conventional plant breeding pipelines for different traits measured in different environments (Crossa et al., 2017; Dreisigacker et al., 2021). The application of GS has focused on two approaches. One approach predicts the complete genetic values of individuals and focuses on both additive and non-additive effects, thereby estimating the genetic performance of candidate cultivars (Crossa et al., 2017). Additive or genetic values are predicted in breeding generations using as much phenotypic information as possible obtained from different environments in a complete or incomplete (sparse) multi-environment testing scheme (Jarquin et al., 2020). A second approach is predicting additive effects in early generations (bi-parental F2, or multi-parental populations) to achieve a rapid selection cycle with a short interval (Vivek et al., 2017; Zhang et al., 2017; Beyene et al., 2021). In these instances, the main focus is on the prediction of breeding values of the genotypes. The application of GS offers attractive benefits but comes with challenges when implemented into current conventional breeding systems.

Genomic selection is affected by a range of factors occurring at different levels. For example, one complexity arises while incorporating genotype × environment (G × E) interaction into statistical models. Also important are the genome interactions related to G × E interactions for multi-traits and the complexity of the traits (complex vs. simple) evaluated in multiple environments. Some of these complexities can be addressed using parametric models where the effect of phenotypic lines can be replaced by g_j expressed as a linear regression of the line phenotype on marker covariates (this approximates the genetic value of the line). The matrix $\boldsymbol{G}$ is a genomic relationship matrix with markers centered and standardized (VanRaden, 2007), which leads to what is known as Genomic Best Linear Unbiased Predictor (GBLUP). The genomic relationship matrix $\boldsymbol{G}$ is the most common parametric linear kernel that accounts for the additive relationship between lines. Also, the effect of the line can be replaced by $\boldsymbol{A}$, the additive relationship matrix of the linear kernel is derived from pedigree and proportional to the identical by descent (IBD) probabilities.

Semi-parametric genomic regression methods are efficient for capturing non-additive variation. The Reproducing Kernel Hilbert Space (RKHS) method was initially used in animal breeding (Gianola et al., 2006; Gianola and Van Kaam, 2008; Gonzalez-Recio et al., 2008) and in wheat genomic-assisted plant breeding with very promising practical results (de los Campos et al., 2009, 2010; Crossa et al., 2010; González-Camacho et al., 2012; Pérez-Rodríguez et al., 2012). Semi-parametric models use kernel methods capturing non-linear relationships between the phenotype and genotype for complex traits, such as grain yield. The Gaussian Kernel (GK) or RKHS method is a non-linear kernel (González-Camacho et al., 2012) that captures major and complex marker effects in addition to their interaction effects. Note that the non-linear kernels and the linear kernels can be employed for a single environment model and on a genomic multi-environment model, such as G × E. According to de los Campos et al. (2009); Crossa et al. (2010); Pérez-Rodríguez et al. (2012), and Cuevas et al. (2017), it is well known that the GK is efficient for capturing additive × additive epistasis interactions in multi-environment trials.

While GS is routinely deployed in the stage 1 yield trials of the CIMMYT Global Wheat Program, genomic prediction has not yet been applied in early generations due to a number of factors including, but not limited to, genetic complexity of the crop, logistics, and expense of establishing a faster cycle integrated into the existing shuttle breeding method, which involves moving seed within and/or outside Mexico each breeding generation. However, from the 2009–2010 to 2014–2015 seasons, a large GS proof-of-concept experiment was carried out with the objective of incorporating genomic prediction for increased yield in the early stages of population improvement in the context of the standard methodology applied in the CIMMYT Wheat Breeding Program in Mexico. Here, we present the results of this initial experiment, which is the first reported in wheat applying GS as early as the F2 generation. Note that the genome-based models incorporating G × E were not yet available during the time this experiment was conducted, so were not applied in this study.

MATERIALS AND METHODS

Training and Prediction Sets

Composition of the Base Training Set

The training set was comprised of 1,334 entries from the 17th and 18th Semi-Arid Wheat Yield Trials (17th and 18th SAWYT), and International Bread Wheat and Semi-Arid Wheat Screening Nurseries (29th and 30th SAWSN, 45th IBWSN; **Figure 1** and **Supplementary Table 1**).

Development of Populations to Validate Early Generation Genomic Prediction

This study sought to incorporate genomic prediction for increased yield in the early stages of population development in the context of the standard breeding methodology applied at CIMMYT in Mexico. This method used selected bulks and two field generations per year alternating between the CIMMYT Experimental Station in Toluca (Lat 19° N, Long

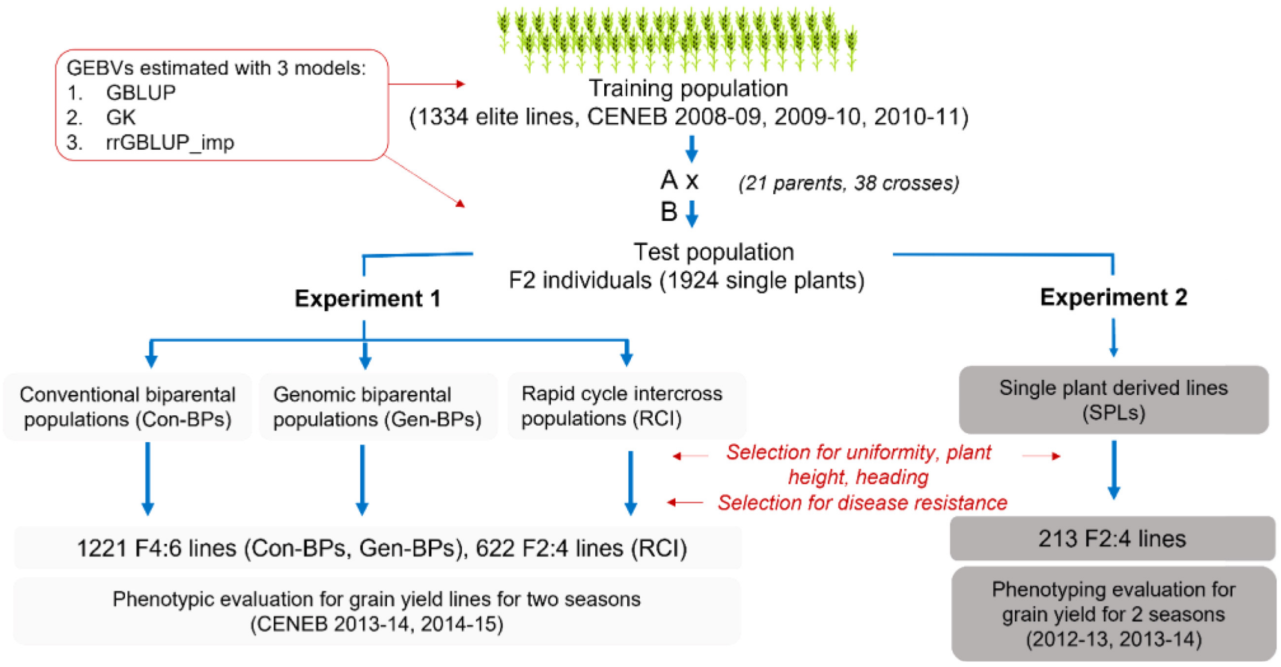

GEBVs estimated with 3 models:
1. GBLUP
2. GK
3. rrGBLUP_imp
Training population
(1334 elite lines, CENEB 2008-09, 2009-10, 2010-11)
A x
B
(21 parents, 38 crosses)
Test population
F2 individuals (1924 single plants)
Experiment 1
Experiment 2
Conventional biparental populations (Con-BPs)
Genomic biparental populations (Gen-BPs)
Rapid cycle intercross populations (RCI)
Single plant derived lines (SPLs)
Selection for uniformity, plant height, heading
Selection for disease resistance
1221 F4:6 lines (Con-BPs, Gen-BPs), 622 F2:4 lines (RCI)
Phenotypic evaluation for grain yield lines for two seasons (CENEB 2013-14, 2014-15)
213 F2:4 lines
Phenotyping evaluation for grain yield for 2 seasons (2012-13, 2013-14)

FIGURE 1 | Overview of the development of populations to validate early generation genomic prediction. Con-BPs, conventional biparental populations; Gen-BPs, genomic biparental populations; GBLUP, genomic best linear unbiased prediction; rrGBLUP, Ridge Regression BLUP; SPLs, single plant-derived lines; GK, Gaussian Kernel.

99° W Elevation 2,640 masl) and the Campo Experimental Norman E. Borlaug (CENEB) station at Cd. Obregon (Lat 27° N, Long 110° W, Elevation 39 masl). The phenotypic selection in segregating generations was for semidwarf plant height, phenology equivalent to parents and checks and disease resistance; notably stripe rust (*Puccinia striiformis* f. sp. *tritici*), leaf rust (*Puccinia triticina*), and septoria tritici blotch (*Zymoseptoria tritici*).

Thirty-eight biparental breeding populations were generated from crosses between 21 parent lines selected from the training set and advanced to F_2 (**Supplementary Table 2**). Parents were selected to limit segregation for height and phenology. From these crosses, four sets of sub-populations were derived as follows (**Figure 1** and **Supplementary Table 2**).

Conventional Biparental Populations

Conventional biparental populations (Con-BPs) comprised lines derived from a random sample of approximately 1,000 F2 seeds per cross. These Con-BP F2s were each sown in a 10 m × 1.6 m plot at the CENEB station in the 2011–2012 season, and approximately 50 F2 plants with desirable height, phenology, and disease reaction were selected to form an F3 bulk. Approximately 1,000 seeds from each F3 bulk were planted in 10 m × 1.6 m plots in Toluca in May 2012, and 50 plants with desirable plant type and disease reaction were selected and harvested to form an F4 bulk. Again, 1,000 seeds of each F4 bulk were planted in the same plot configuration, 50 plants per plot were selected for plant type and disease reaction and each harvested individually to form F4:5 single plant selections. These were increased in single 2 m double-row beds over summer 2013 at the Toluca Station. We aimed to select 20 lines from each cross based on uniformity, plant type, and disease reaction. Selected rows were individually harvested and threshed to generate the F4:6 lines that were planted in field trials at the CENEB station in the 2013–2014 and 2014–2015 seasons.

Genomic Biparental Populations

Genomic biparental populations (Gen-BPs) were formed from 50 F2 plants per cross that were space planted at the same time and in the same field location with the Con-BP F2 subpopulations. DNA was extracted from leaf tissue of F2 individuals for genotyping-by-sequencing (GBS) and calculation of Genomic Estimated Breeding Values (GEBVs). Individuals from each cross were selected on the basis of GEBV, plant type, and disease reaction. As GEBVs from the different prediction methods were not highly correlated (see section "Results"), with no way to know which was most predictive, F2s with the highest average GEBV across the three prediction methods were selected. Selfed seed from selected F2 plants within each cross was combined to form F3 bulks. Gen-BPs were advanced from F3 bulk to F4:6 line concurrently, with the same methods and in the same field nurseries as the Con-BPs. In other words, selection methodologies and intensities were identical for Gen-BPs and Con-BPs from the F3 bulk stage. Similar numbers of lines were derived from the Gen-BP and Con-BP subpopulations of most crosses. Six crosses did not produce progeny with acceptable

TABLE 1 | Training set experiments summary data.

Trial	17SAWYT		18SAWYT		29SAWSN			30SAWSN			45IBWSN		
Season	**2009**	**2010**	**2009**	**2010**	**2009**	**2010**	**2010**	**2010**	**2011**	**2011**	**2010**	**2011**	**2011**
Experiment													
Type	Bed	Flat	Bed	Bed	Bed	Bed	Flat	Bed	Bed	Flat	Bed	Flat	Bed ZT
Entries	50	50	43	43	264	264	264	264	264	264	780	780	780
Reps	2	2	1	2	1	3	3	1	3	3	3	3	3
Mean DTH	84.7	83.6	80.5	81.5	80.1	85.1	–	84.6	91.0	87.4	88.6	91.9	95.1
Mean HT	112.0	93.3	102.3	100.0	99.1	109.8	–				104.7	113.0	100.0
Mean YLD	7.15	6.87	6.85	5.65	7.20	7.16	6.95	7.12	7.17	5.82	7.70	7.66	6.63
H^2 YLD	0.67	0.45	0.58	0.70	0.63	0.83	0.81	0.64	0.72	0.78	0.88	0.89	0.68
CV YLD	7.76	11.98	7.72	7.40	6.90	9.84	12.05	8.91	7.61	12.43	8.07	10.00	10.46

Data for the training sets included wheat lines tested in several international trails (17–18 SAWYT, 20SAWSN, 30SAWSN, and 45IDWSN) during seasons 2009, 2010, and 2011 using two planting systems bed and flat. Heritability (H^2) and coefficient of variation (CV) of grain yield (YLD, ton/ha) and an average of days to heading (DTH, days), and height (HT, cm).

combinations of plant type and disease reaction for Gen-BP and Con-BP subpopulations.

Rapid Cycle Intercross Populations

Rapid cycle intercross populations (RCIs) were generated by crossing selected F2 individuals, those with the highest average GEBVs across prediction methods, within and between Gen-BP subpopulations. For intercrosses within a population, average GEBV and genetic distance based on the kinship matrix (VanRaden, 2008) among individuals were used to increase the probability of combining distinct, favorable alleles. All plants selected for crossing also produced enough selfed seed to contribute approximately the same number to the Gen-BP F3 bulks as plants that were not selected for intercrossing. A total of 37 RCI populations were generated. RCI F1s were space planted by cross at the Toluca Research Station in the summer of 2012 in the same field and under the same conditions as the Gen-BP and Con-BP F3 bulks. Plants were selected based on plant type and disease reaction. Selected plants were bulked by cross to form RCIF2 bulks and were then advanced concurrently with the same selection methods as for the Con-BP and Gen-BP subpopulations to produce RCI F2:4 lines for field trials at the CENEB station in the 2013–2014 and 2014–2015 seasons. The 37 RCI populations were represented by a variable number of selections although a total of 26 populations produced 16 or more selections and only 5 populations produced fewer than 10 selections. Overall, 622 lines were derived from RCI populations.

Single Plant-Derived Lines

Single plant-derived lines (SPLs) were developed from a subset of 240 F2 plants from across the Gen-BP subpopulations. The selection of F2 plants was based on a visual assessment of acceptable plant height, phenology, and agronomic type, without consideration of disease reaction or GEBV. F3 seed from each selected plant was sown in a single row at CIMMYTs El Batan Research Station in May of 2012. Rows were sprayed with a fungicide to control diseases and were assessed for uniformity, height, and phenology. Rows were discarded only if they expressed excessive height, slow phenological development, or high levels of within-row variability. From the 240 rows, 213 F2:4 SPLs were selected for field testing to assess response to selection for F2 GEBV using each of the three different GEBV calculation methods. SPLs were obtained from 36 of the 38 Gen-BP subpopulations and tested in field trials at CENEB in the 2012–2013 and 2013–2014 seasons.

Field Trials and Phenotyping

Training Set

Phenotypic data for the training set of 1,334 lines were generated in field trials at CENEB, Cd. Obregon over the 2008–2009, 2009–2010, and 2010–2011 growing cycles under irrigated conditions with management to achieve high yield according to local best practice. Summary data for these trials are outlined in **Table 1**.

Testing Set

Field trials of the developed populations were conducted at CENEB across three growing cycles (2012–2013, 2013–2014, and 2014–2015) with equivalent management to that applied to the training set. Plots were of 4.8 m^2 (3 m × 1.6 m). Each trial was conducted in two consecutive seasons. Trials in each growing season were planted in late November or early December and harvested in early May. Data were collected for grain yield, plant height, and heading date. Details specific to the trials related to each of the following components of our research are provided in the following sections and summarized in **Table 2**.

Validation of Genomic Predictions for Wheat Grain Yield

Experiment 1 – Conventional Biparental, Genomic Biparental, and Rapid Cycle Intercross Populations

Phenotypic data for the Con-BP, Gen-BP, and RCI-derived lines (591, 630, and 622 lines, respectively) were generated in field trials at CENEB, Cd. Obregon over the 2013–2014 and 2014–2015 crop cycles (**Table 3**). Entries were randomly assigned to 1 of 10 different sub-experiment blocks with each sub-experiment being a two-rep row-column design.

TABLE 2 | Summary field trial data for Experiments 1 and 2.

Trial	Experiment 1 Con-BP vs. Gen-BP and RCI		Experiment 2 SPL F2 GEBV validation	
Season	**2013–2014**	**2014–2015**	**2012–2013**	**2013–2014**
Experiment				
Type	Bed	Bed	Bed	Bed
Design	Row-column	Row-column	Row-column	Row-column
Entries	2,000	2,000	240	240
Reps	2	2	2	2
Mean DTH[a]	76	72.6	80.5	78.4
H^2 DTH[a]	0.88	0.74	0.95	0.94
CV% DTH[a]	1.82	2.17	3.5	3.7
Mean HT[b]	99.6	105.1	104.2	101.3
H^2 HT[b]	0.48	0.72	0.66	0.65
CV% HT[b]	5.08	3.21	4.1	4.2
Mean YLD[c]	5.98	4.72	7.12	5.98
H^2 YLD[c]	0.60	0.41	0.73	0.82
CV% YLD[c]	9.63	13.41	9.27	6.95

Experiment 1 compares prediction accuracy of genomic bi-parental (Gen-BP), with conventional bi-parental (Con-BP), and rapid cycling intercross population (RCI) populations in cycles 2013–2014 and 2014–2015. Experiment 2 has single plant-derived lines (SPL) F2:4 validation.

[a] *Days to heading from sowing.*

[b] *Height in cm to tip of ear.*

[c] *Grain yield in tons per hectare.*

TABLE 3 | Experiment 1: Least significant difference (LSD), mean yield comparison of different breeding populations and checks evaluated at Cd. Obregon during 2013–2014 (Year-1) and 2014–2015 (Year-2) growing seasons.

Class	*N*	Mean (ton/ha)	Tukey grouping	
YEAR-1	[LSD (0.05) = 0.1341]			
Checks	80	6.212	A	
Con-BP	1,182	6.048	B	
Gen-BP	1,260	5.990	B	C
Parents	190	5.988	B	C
RCI	1,288	5.902		C
YEAR-2	[LSD (0.05) = 0.1288]			
Checks	80	4.916	A	
Parents	190	4.800	A	B
Gen-BP	1,260	4.741	C	B
Con-BP	1,182	4.741	C	B
RCI	1,288	4.664	C	
Combined	[LSD (0.05) = 0.093]			
Checks	160	5.564	A	
Con-BP	2,364	5.394	B	
Parents	380	5.394	B	
Gen-BP	2,520	5.366	B	C
RCI	2,576	5.283		C

Breeding populations included the genomic bi-parental (Gen-BP), conventional bi-parental (Con-BP), and rapid cycling intercross population (RCI) and parents.

All sub-experiments included common checks. Parents of the populations were included in the experiments and assigned randomly across sub experiments. Grain yield data of the different population types were compared, and differences were determined using the least significant difference (LSD at 5% significance). The expected response to selection was derived by multiplying the narrow sense heritability by the selection differential ($H^2 \times S$). The latter was calculated by dividing the mean of the selected lines by the mean of the full population.

Experiment 2 – Validation of F2 Grain Yield Genomic Estimated Breeding Values in Single Plant-Derived F2:4 Lines

Single plant-derived lines were tested in field trials at CENEB, Cd. Obregon in the 2012–2013 and 2013–2014 crop seasons. Experiments were two replicate row-column designs. Grain yield data for the F2:4 lines were examined for correlation to GEBVs of their respective, individual F2 progenitor plant.

Genotyping

The wheat genotypes included in the training set and F_2 plants, indexed by their genotypic identification number (GID), were characterized using GBS following the same procedure as described in Poland et al. (2012). Briefly, genomic DNA was extracted from seedling leaf tissue using the procedure described in Dreisigacker et al. (2016). Two enzymes PstI (CTGCAG) and MspI (CCGG) were used to digest genomic DNA. Individual samples were ligated with barcoded adapters and pooled by plate into a single library. Each library was sequenced on a single lane of Illumina HiSeq2000. A total of 45,818 single nucleotide polymorphisms (SNPs) markers were initially obtained. The filtering consisted of removing markers whose minor allele frequency (MAF) was less than 5% or had more than 80% missing values. After initial filtering, 29,999 markers were available for further analysis.

Statistical Models and Methods

The Base-Line Phenotype Model for the Training Populations

This part of the analysis was performed on the six field trials that included the 1,334 entries in the training set which are outlined in **Table 1**. Best Linear Unbiased Estimates (BLUEs) for grain yield across trials were generated using the following linear mixed model:

$$Y_{ijkl} = \mu + \mathrm{g_i} + \mathrm{Year}_j + \mathrm{R}_{k(j)} + \mathrm{B}_{l(kj)}(\mathrm{g} \times \mathrm{L})_{ij} + e_{ijkl}$$

where Y_{ijkl} is the phenotype of wheat line i-th at location j-th in replicate k-th within the block l-th, μ is the overall mean, Year_j is the fixed effect of the year j-th, $\mathrm{R}_{k(j)}$ is the fixed effect of the k-th replicate within year j-th, $\mathrm{B}_{l(kj)}$ is the random effect of the incomplete block l-th within replicate k-th and year j-th assumed to be independently and identically normal distributed (iid) with mean zero and variance σ_b^2, $\mathrm{g_i}$ is the fixed effect of genotype i-th, $(\mathrm{g} \times \mathrm{L})_{ij}$ is the fixed effect of the genotype × year interaction, and e_{ijkl} is the random error assumed to be iid normal with mean zero and variance σ_e^2. Broad sense heritability (H^2) was computed on an entry-mean basis according to Bernardo (2010) as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{gy}^2}{y} + \frac{\sigma_e^2}{y \times r}}$$

where σ_g^2 is the genotypic variance, σ_{gy}^2 is the genotype × year interaction variance, σ_e^2 is the estimated of the error variance, y is the number of years, and r is the number of replicates. Note that different trials had different numbers of testing years, 17–18 SAWYT data had trials in years 2009 and 2010, whereas the other three trials had 3 years of testing (**Table 1**).

Genomic-Enabled Prediction Models

Meuwissen et al. (2001) were the first to propose whole-genome regression methods (GS) by jointly fitting hundreds of thousands of markers with major and small effects. In the whole-genome regression methods, the number of markers (p) greatly exceeds the number of data-points (n) available; thus, implementing regression methods poses important statistical and computational challenges. However, new developments in the area of shrinkage estimation procedures allows the implementation of whole-genome regression methods.

We considered three different models: GBLUP using additive genomic relationships (VanRaden, 2008), the GK or RKHS regression (Gianola et al., 2006) which is equivalent to a GBLUP but with a non-linear kernel, and the rrGBLUP_imp where missing markers were imputed (Endelman, 2011). The GBLUP and RKHS models were fitted using routines kindly provided by de los Campos (personal communication). Nowadays, GBLUP, RKHS, and many other models can be fitted in the BGLR package (Pérez-Rodríguez and de los Campos, 2014), which is available on the CRAN website. This software was not available at the time our study was conducted.

The Genomic Best Linear Unbiased Prediction Model

The regression model for wheat lines (i = 1, 2,. . ., n) is given by:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{u} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ is the response vector of n phenotypic observations, μ is the overall mean, and the random vectors of the genetic values $\mathbf{u}$ and the errors $\boldsymbol{\varepsilon}$ are independent variables with $\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \sigma_u^2\mathbf{K})$ and $\varepsilon \sim \mathbf{N}(\mathbf{0},\sigma_\varepsilon^2\mathbf{I})$, respectively, where σ_u^2 is the variance of $\mathbf{u}$, $\boldsymbol{I}$ is the identity matrix, and $\mathbf{K}$ is a symmetric semi-positive definite matrix representing the covariance of the genetic values, and $\boldsymbol{\varepsilon}$ is the vector of random errors with normal distribution and common variance, σ_ε^2. The p bi-allelic centered and standardized molecular markers are represented in incidence matrix $\mathbf{X}$ of order $n \times p$ such that $\mathbf{K} = \mathbf{G} = \frac{\mathbf{XX}'}{p}$ is a linear kernel. Model (1) is known as GBLUP (VanRaden, 2007, 2008).

Under the conditions given above, model (1) estimates the genomic relationship by means of its linear kernel $\mathbf{XX}'/p$, where p is the number of markers. However, a nonlinear kernel, such as the GK, can also be used (Cuevas et al., 2016). The model represented by Eq. 1 is computationally very efficient and convenient when $n >> p$ (de los Campos et al., 2012).

Gaussian Kernel or Reproducing Kernel Hilbert Space Regressions

In general, the parametric genomic linear regression function has a rigid structure comprising a set of assumptions, which may not be met in GS problems. Thus, departures from linearity can be addressed by semi-parametric approaches, such as the GK or RKHS regressions (Kimeldorf and Wahba, 1971; Gianola and Van Kaam, 2008; Gianola, 2013). The GK regression for semi-parametric, genomic-enabled prediction, such as kernel regression, is necessary to reduce the dimension of the parametric space and maybe able to capture complex cryptic interaction among markers (Gianola et al., 2006, 2014). Morota and Gianola (2014) pointed out that most studies carried out so far suggest that whole-genome prediction coupled with combinations of kernels may capture non-additive variation (Gianola et al., 2014).

The basic idea underlying the GK approach to GS (Kimeldorf and Wahba, 1971; Gianola, 2013) is to use the matrix of markers $\mathbf{X}$ to build a covariance structure among genetic values $\mathbf{u}$. Therefore, $\mathbf{u} \sim N(\mathbf{0},\sigma_g^2\mathbf{K}_h)$ is independent of $\boldsymbol{\varepsilon}$ (Crossa et al., 2010; de los Campos et al., 2010), $\mathbf{K}_h$ is a symmetric positive semi-definite matrix of order $n \times n$, known as the reproducing kernel (RK) matrix, which depends on the markers and the bandwidth parameter $h > 0$, $\sigma_g^2 > 0$, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of homoscedastic and independent normal errors.

This general approach requires choosing an RK, for example, a GK function

$$K_h\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-hd_{ij}^2/q_{0.05}\right), \tag{2}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the marker vectors for the i-th and j-th individuals, and $q_{0.05}$ is the fifth percentile of the squared Euclidean distance d_{ij}^2 (González-Camacho et al., 2012).

Ridge Regression Best Linear Unbiased Prediction With Imputed Marker Data

The marker-based, additive relationship matrix was calculated with the function A.mat in R package rrGBLUP, version 4.1 (Endelman, 2011), which centers (but does not standardize) each marker by the population mean (VanRaden, 2008). The relationship matrix was additionally calculated with the imputed markers. Missing data were imputed with the "EM" option in A.mat, which implements a multivariate normal expectation-maximization (EM) algorithm (Poland and Rife, 2012).

A Fivefold Cross-Validation

A fivefold cross-validation was performed to evaluate the prediction performance of the models on the training set. The full dataset was randomly divided into five mutually exclusive subsets, four of which formed the training set for fitting the model, and the fifth was used as a test set. Predictive abilities were calculated as the Pearson's correlation coefficient between the predicted values and the observed phenotypic values of the test set.

RESULTS

Validation of Genomic Prediction Models

Predictions with GBLUP, rrGBLUP_imp, and GK in the training population had similar levels of predictive ability for a yield of 0.42–0.43 as determined by a fivefold cross-validation (**Table 4**). The GEBVs produced by the three methods showed high correlations of between 0.93 and 0.97

TABLE 4 | Predictive power of GEBVs in a training population of 1,334 inbreds and correlation between GEBV predictions for grain yield by three different calculation methods (GBLUP, rrGBLUP_imp, and GK) among inbreds in the training set and in a target population of 1,924 F2s.

Prediction model	Yield (fivefold cross validation)	GBLUP	rrGBLUP_imp
		Training/F2	Training/F2
GBLUP	0.43	–	–
rrGBLUP_imp	0.42	0.96/0.44	–
GK	0.42	0.97/0.37	0.93/0.13

GBLUP, genomic best linear unbiased prediction; rrGBLUP, Ridge Regression BLUP; GK, Gaussian Kernel.

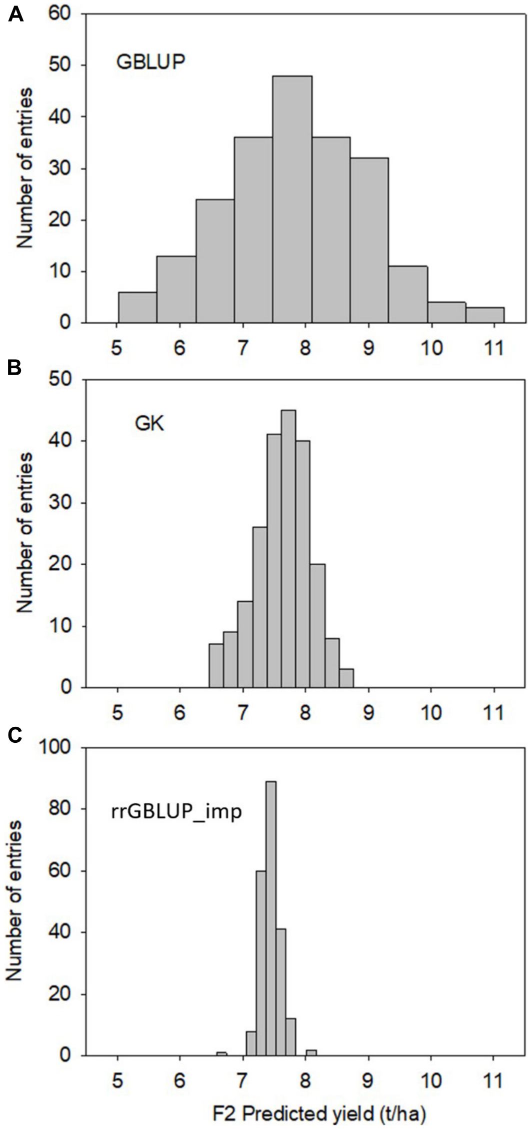


FIGURE 2 | Experiment 2. Distribution of F2-predicted breeding values (grain yield) of 213 randomly selected individuals estimated using GK (top, **A**), GBLUP (Middle, **B**), and rrGBLUP (low, **C**) with imputed markers.

in the training set. In contrast, the models produced divergent predictions in F2 populations. The shapes of the distribution of GEBVs from each prediction method also differed with GBLUP having a wide distribution from 5 to 11 ton/ha while rrGBLUP_imp values were more narrowly grouped between 6.5 and 8.1 ton/ha (**Figures 2A–C**). The lack of correlation and different distributions of values caused uncertainty about which was the most appropriate method to use in the selection of F2 plants to generate genomically selected bulks and to intercross in a rapid cycle intercross strategy. As the GEBVs were uncorrelated, not negatively correlated, individuals with the highest GEBVs averaged across the prediction methods were selected for selfing to form F3 bulks and for intercrossing to form the RCI populations. Phenotypic selection was also applied for plant height, phenology, and disease reaction in the same way as for the Con-BP populations. For the RCI populations, the additional criteria of maximizing genetic differences between F2 individuals, if selected from the same biparental cross, were applied in planning intercrosses.

Because populations were advanced through a selected bulk method to develop the material tested in Experiment 1, this experiment could not address the question of whether one method was superior to another in F2 GEBV calculation. Therefore, a random subsample of F2 plants was chosen to develop single F2 plant-derived lines so a correlation between the yield of a derived line and GEBV of an F2 could be measured. This set of lines was the basis for Experiment 2.

Experiment 1 – Conventional Biparental, Genomic Biparental, and Rapid Cycle Intercross Populations

A total of 1,857 lines were derived from conventional, phenotypically selected biparental (Con-BP), Gen-BP, and RCI breeding methods with roughly similar numbers from each (**Supplementary Table 2**). All methods used phenotypic selection

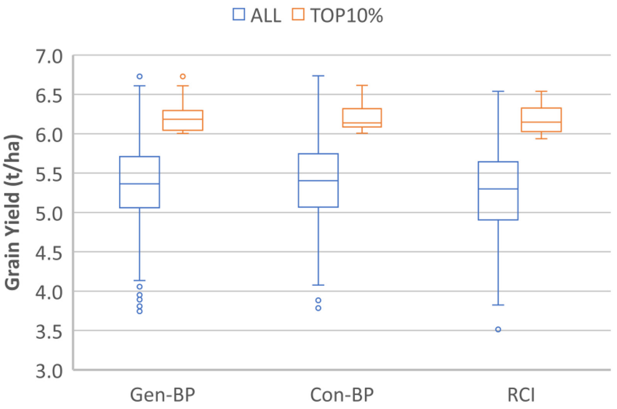


FIGURE 3 | Experiment 1. Box-plots comparing grain yield distribution of all the lines for Genomic Bi-parental (Gen-BP), Conventional Bi-parental (Con-BP), and Rapid Cycle Intercross population (RCI) populations (blue box-plots) and their corresponding top 10% entries (red box-plots).

TABLE 5 | Experiment 1: Comparing the expected response to selection under 5 and 10% selection intensity for different selection schemes for average grain yield (AV_YLD ton/ha) derived from trials performed at the CENEB station in seasons 2013–2014 and 2014–2015.

			TOP5%			TOP10%		
Class	***N***	**AV_YLD**	**AV_YLD**	***S***	***R***	**AV_YLD**	***S***	***R***
Gen-BP	630	5.366	6.326	0.960	0.576	6.189	0.823	0.494
Con-BP	591	5.394	6.347	0.953	0.572	6.216	0.822	0.493
RCI	635	5.286	6.312	1.026	0.615	6.177	0.891	0.534

Con-BPs, conventional biparental populations; Gen-BPs, genomic biparental populations; RCI, rapid cycling intercross.
S = Selection differential (selected mean − population mean).
R = Expected response to selection = $H^2 \times S$.
Expected selection response ($H^2 = 0.6$).

to progress material through selected bulk stages, while the Gen-BP method used GEBVs of F2 plants to add a single cycle of GS and RCI used the F2 GEBVs to select plants for intercrossing to produce new populations that were subsequently passed through the same phenotypic selection methodologies.

Field testing showed that Con-BP lines yielded an average of 2% more than RCI lines ($P < 0.001$), Gen-BP lines yielded an average of 1.5% more than RCI lines ($P < 0.01$), and there was no significant difference between Gen-BP and Con-BP (**Table 3**) populations. Similar comparisons of yield focusing on the top 10% highest yielding lines in each population type showed similar patterns with Con-BP having the highest yield, significantly greater than GS-BP and RCI (**Figure 3** and **Table 5**). Although differences were statistically significant, they were only approximately 1%. Gen-BP subpopulations in the top 10% were not significantly different in yield to the top 10% of RCI lines. Response to selection in the RCI populations was marginally greater than for Con-BP and Gen-BP, but the difference was small and likely reflects the lower mean yield and distribution of grain yield in the RCI populations compared to the other population types (**Table 3** and **Figure 4**).

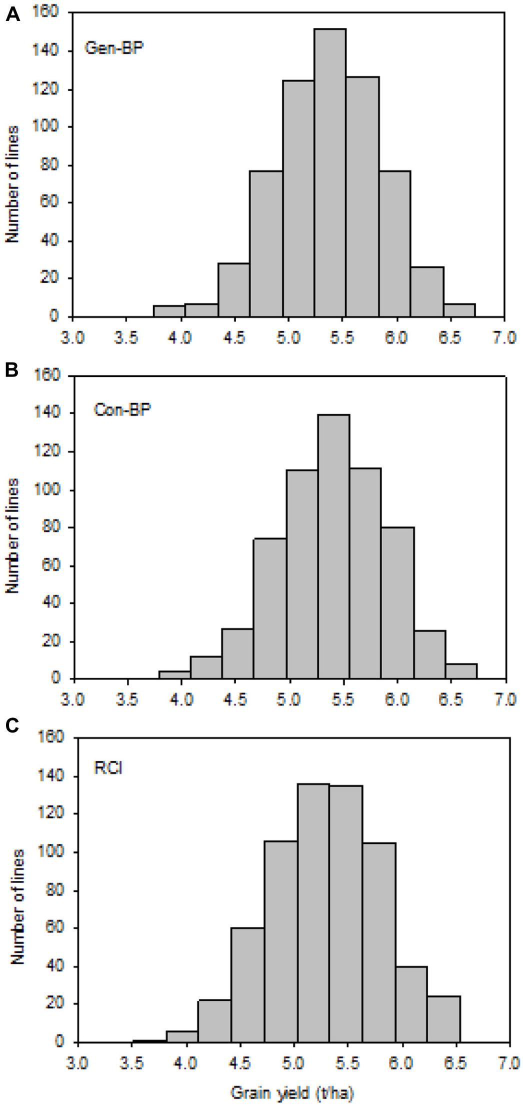


FIGURE 4 | Experiment 1: Distribution of observed grain yield across 2 years under different selection strategies: **(A)** Gen-BP, **(B)** Con-BP, and **(C)** RCI. Con-BPs, conventional biparental populations; Gen-BPs, genomic biparental populations; RCI, rapid cycling intercross.

TABLE 6 | Experiment 2: Correlations between F2:4 GEBVs from three prediction methods (GBLUP, GK, and rrGBLUP_imp) from grain yield of 213 derived F2:4 lines across two seasons.

	F2:4 observed yield	
F2:4 predicted YLD	**Correlation**	***P*-Value**
GBLUP	0.2870	0.000024
GK	0.3020	0.000009
rrGBLUP_imp	−0.0733	0.290000

GBLUP, genomic best linear unbiased prediction; rrGBLUP, Ridge Regression BLUP.

Experiment 2 – Validation of F2 Grain Yield Genomic Estimated Breeding Values in Single Plant-Derived F2:4 Lines

In Experiment 2, we compared the predictive ability of the different GEBV calculation methods in F2 in a set of 213 single plant-derived F2:4 lines from randomly selected F2 plants. Trials of the F2:4 SPLs showed a significant positive correlation with F2 GEBVs from GK and GBLUP (**Table 6** and **Figure 5**). Individuals with the highest 10 and 20% GEBVs predicted by GK, produced F2:4 progeny lines with realized grain yield gains of 4.7 and 4.2%, respectively; significantly higher than the mean of 50 random samples from across the full set of F2s (**Table 7**). The top 10 and 20% of F2s predicted by the GBLUP method showed realized gains of 3.68 and 2.60%, respectively, in their F2:4 progenies; significantly higher than the mean of 50 random samples of the same proportions (**Table 7**). Contrarily, selecting the top 10 and 20% of F2 GEBVs estimated with rrGBLUP_imp did not produce F2:4 progenies with a higher mean performance compared to random samples.

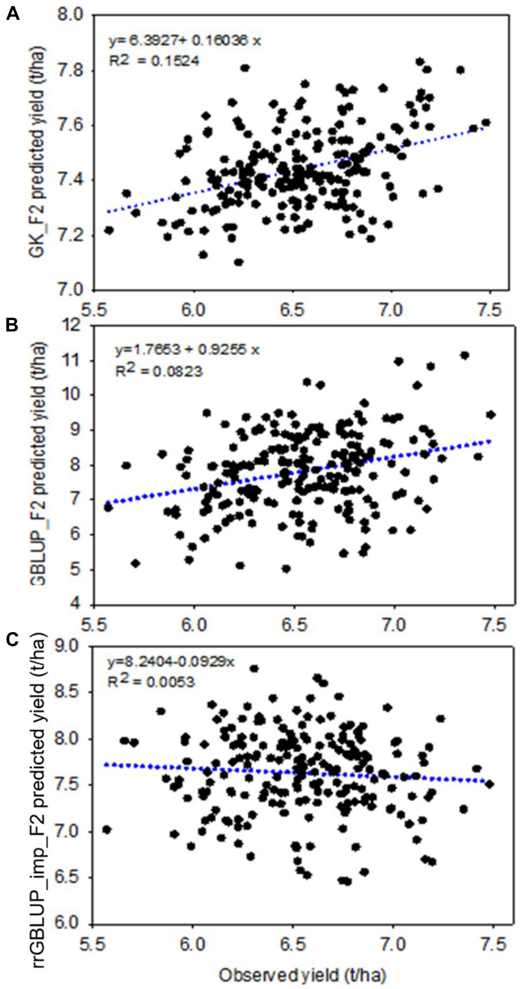


FIGURE 5 | Experiment 2: Scatter plot showing the correlation between F2 predicted yield, estimated with different prediction models GK (top, **A**), GBLUP (middle, **B**), and rrGBLUP_imp (low, **C**), with an observed yield of 213 derived F2:4 lines. GBLUP, genomic best linear unbiased prediction; rrGBLUP, Ridge Regression BLUP; GK, Gaussian Kernel.

Within the subset of 213 F2s which were randomly sampled to produce F2:4 bulks for yield testing, the correlations between prediction methods were also low (**Table 4** and **Figure 6**). **Figures 5A–C** shows scatterplots of F2 single plant GEBVs vs. realized yields of derived F2:4 lines (GEBVs on *Y*-axis, yields on *X*-axis). From these, it is clear that correlations between F2 GEBVs and F2:4 yield for GK and GBLUP are not strong but are not driven by outliers with high leverage. In both cases, the selection of F2s with the highest GEBVs would avoid the selection of the lowest yielding F2:4 lines.

DISCUSSION

The proof-of-concept Experiment 2 reported here demonstrates the potential of early generation genomic prediction to increase genetic gain over conventional selection methods by allowing the ability to increase the number of crossing cycles per year. In Experiment 2 of our study, F2 GEBVs generated by GK and GBLUP methods showed significant positive correlations with the yield of derived lines. The highest 10 and 20% of GEBVs from the GK method showed 4.7 and 4.2% increases, respectively, and the top 10 and 20% of F2s GEBVs predicted by GBLUP showed realized gains of 3.68 and 2.60% over a 50× random sample of the same proportion of lines from the same populations. In contrast, a similar analysis of F2 GEBVs from the rrGBLUP_imp method showed no difference from the mean of the 50× random sampling.

Each of the three prediction methods used in this study produced highly correlated GEBVs in inbreds and the same levels of predictability of inbred performance based on cross validation in a training set of elite CIMMYT inbreds. In contrast, predictions in F2s derived from crosses between inbreds that were part of the training set for the model showed little to no correlation and differing levels of predictive ability compared with a realized yield of F2 SPLs. These differences are likely due to the different abilities of the prediction methods to handle heterozygosity, which is generally not accurately characterized with a GBS genotyping platform and the importance of non-additive variation in wheat. This may be reflected in the much narrower distribution of GEBVs from rrGBLUP_imp compared to GBLUP and particularly GK. The difference in the distribution of the GEBVs between the GBLUP and GK methods is likely due to the different shrinkage applied in each method. On the other hand, differences between GBLUP and rrGBLUP_imp are likely due to the imputation method used.

In Experiment 1, we attempted to incorporate F2 genomic prediction into a selected bulk breeding methodology closely mirroring the typical breeding methodology in the CIMMYT spring bread wheat program. The three different prediction methods generated F2 GEBVs that showed little correlation with one another. It should be noted that the low correlations between the rrGBLUP_imp with GBLUP and GK were considered as a rare result, especially knowing the equivalence between the GBLUP and the rrGBLUP_imp. The reasons for the failure of the rrGBLUP_imp in generating similar predictions to GBLUP are unknown but may be attributable to different factors. For example the nature of the imputation algorithm or convergence issues with the Expectation-Maximization algorithm in rrGBLUP_imp. Since the three methods had similar ability to predict yield of inbreds and predictions were correlated, it was difficult to discard one of the models based on observed phenotypes and we decided to use an average of the methods. If we had conducted additional research to confirm that GK was the most predictive method or that GBLUP also showed a useful level of predictability, we would likely have made better selections of F2 individuals to form selected bulks and to make early generation intercrosses. Given that our selections were probably no better than random and the number of F2s selected was less than in parallel conventionally selected populations, it is hardly surprising that a lower level of genetic variance (presumed by planting only 5% of the number of F2s in Gen-BP vs. Con-BP) did not result in a yield advantage in the genomically selected

TABLE 7 | Yield of F2:4 lines based on selection of the top 10 and 20% of GEBVs from different methods (GBLUP, GK, and rrGBLUP_imp) compared to a random sample of 10 and 20% of all F2:4 lines, with 50× sampling (the top 10 and 20% *t*-test: Two-Sample Assuming Unequal Variances.

	Sample	Best 20% F2:4-predicted lines			Sample	Best 10% F2:4-predicted lines		
	20%	**GK**	**GBLUP**	**rrGBLUP_imp**	**10%**	**GK**	**GBLUP**	**rrGBLUP_imp**
Mean yield (ton/ha)	6.52	6.76	6.69	6.49	6.52	6.83	6.80	6.47
Variance	0.0036	0.1536	0.1081	0.0808	0.0077	0.1107	0.1334	0.0989
Observations (*n*)	50	42	42	43	50	21	20	21
Mean difference (*D*)		0.24	0.17	−0.03		0.31	0.28	−0.05
% Mean difference		3.68	2.61	−0.46		4.75	4.29	−0.77
Degree of freedom		43	43	45		21	20	21
t-Value		3.9823	3.3382	−0.6783		4.2767	3.3561	−0.7761
P (*t* < = *t*) one-tail		0.0001	0.0009	0.2505		0.0002	0.0016	0.2232
t critical one-tail		1.6811	1.6811	1.6794		1.7207	1.7247	1.7207
P (*t* < = *t*) two-tail		0.0003	0.0017	0.5011		0.0003	0.0031	0.4463
t critical two-tail		2.0167	2.0167	2.0141		2.0796	2.086	2.0796

GBLUP, genomic best linear unbiased prediction; rrGBLUP, Ridge Regression BLUP; GK, Gaussian Kernel.

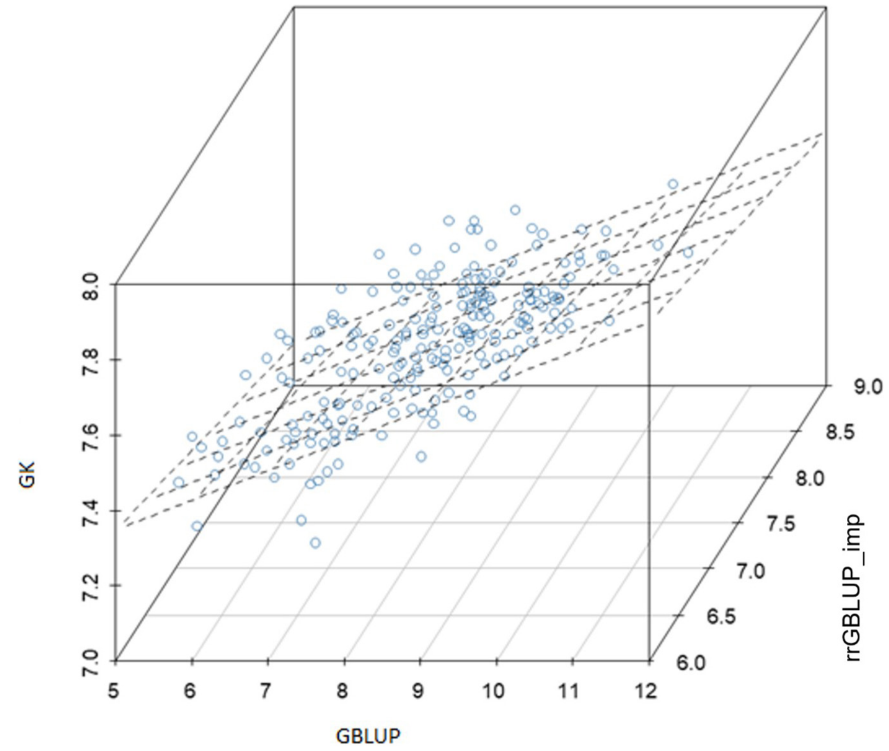


FIGURE 6 | Experiment 2: Relationship between genomic-enabled predictive values of 213 F2 which were later advanced to F4 (F2:4) using models GK, GBLUP, and rrGBLUP_imp. GBLUP, genomic best linear unbiased prediction; rrGBLUP, Ridge Regression BLUP; GK, Gaussian Kernel.

biparental-derived inbreds (Gen-BP) and the early generation intercross derived (RCI) inbred populations; both on average and in the highest yielding 20% of lines from each of the population types.

When comparing genome-based predictions, we should also emphasize that in this study the accuracy of the three methods (GBLUP, GK, and rrGBLUP_imp) for predicting F2 plants was measured at the F2:4. Therefore, any attempt to make a precise estimate of errors among the three methods and benchmarking results from genome-based methods with those under conventional breeding methods in terms of biases and errors are complex and out of the scope of this research. It would be worthwhile to investigate further methods to optimize prediction power in early generation wheat populations. If a robust method can be determined, there are useful increases in genetic gain from early generation genomic prediction in wheat, particularly, in populations that are not varying for some of the obvious drivers of yield that are easily selected phenotypically, such as height or flowering time. Considering there are roughly one million F2 plants generated per year in the CIMMYT spring bread wheat program, early generation genomic prediction will likely be best targeted to certain types of populations that provide the greatest probability of higher response to selection or where there is little obvious variation amenable to phenotypic selection.

These evaluations give the first indications of genetic gains from early generation GS for a highly complex trait in a practical wheat breeding program.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: https://data.cimmyt.org/dataset.xhtml?persistentId=hdl:11529/10548576.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

FUNDING

This project was supported through the Bill & Melinda Gates Foundation grant no. OPP101638 Genomic Selection: the next frontier for rapid gains in maize and wheat improvement project.

ACKNOWLEDGMENTS

The authors are grateful to their CIMMYT colleagues for intensive discussions on the subject, CIMMYT field and lab technicians who helped to record phenotypic and genotypic data, collaborators worldwide, as well as scientists in national programs. The authors acknowledge the contributions of Yann Manes, formerly at CIMMYT, for being an early champion of the project and planning the crosses in which the early generation selection experiments were conducted. The authors would also like to acknowledge the contributions of Jeff Endelman formerly at Cornell University, to some of the early analysis of genomic data and model building.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.718611/full#supplementary-material

REFERENCES

Bernardo, R. (2010). “Genotype × environment interaction,” in *Breeding for Quantitative Traits in Plants*, ed. R. Bernardo (Woodbury: Stemma Press), 177–203.

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. (2021). Genetic gains in grain yield through genomic selection in eight bi-parental Maize populations under drought stress. *Crop J.* 55, 154–163. doi: 10.2135/cropsci2014.07.0460

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

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. *Genom. Sel.* 3, 1903–1926. doi: 10.1534/g3.113.008227

Crossa, J., De Los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., and Pérez-Rodríguez, P. (2016). Extending the marker × Environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. *Crop Sci.* 56, 2193–2209. doi: 10.2135/cropsci2015.04.0260

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, P., de los Campos, G., Mahuku, G., Dreisigacker, S., and Magorokosho, C. (2011). Genomic selection and prediction in plant breeding. *J. Crop Improv.* 25, 239–261. doi: 10.1080/15427528.2011.558767

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 (Edinb)* 112, 48–60. doi: 10.1038/hdy.2013.16

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

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P., and de los Campos, G. (2017). Bayesian genomic prediction with genotype × environment interaction kernel models. *G3 Genes Genom. Genet.* 7, 41–53. doi: 10.1534/g3.116.035584

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., Campos, G., et al. (2016). Genomic prediction of genotype × environment interaction Kernel regression models. *Plant Genome* 9:lantgenome2016.03.0024. doi: 10.3835/plantgenome2016.03.0024

de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res. (Camb).* 92, 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.

de los Campos, G., Klimentidis, Y. C., Vazquez, A. I, and Allison, D. B. (2012). Prediction of expected years of life using whole- genome markers. *PLoS One* 7:e40964. doi: 10.1371/journal.pone.0040964

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

Dreisigacker, S., Crossa, J., Pérez-rodríguez, P., Montesinos-L°ópez, O., Rosyara, U., Juliana, P., et al. (2021). Implementation of genomic selection in the CIMMYT global wheat program, findings from the past 10 years. *Crop Breed. Genet. Genom.* 3:e210005. doi: 10.20900/cbgg20210005

Dreisigacker, S., Sehgal, D., Reyes Jaimez, A., Luna Garrido, B., Muñoz Savala, S., and Núñez Ríos, C. (2016). *CIMMYT Wheat Molecular Genetics: Laboratory Protocols and Applications to Wheat Breeding*, eds J. Mollins and S. Mall (Mexico: CIMMYT).

Endelman, J. B. (2011). Ridge regression and other Kernels for genomic selection with R Package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

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

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

Gianola, D., and Van Kaam, J. B. C. H. M. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303. doi: 10.1534/genetics.107.084285

Gianola, D., Weigel, K. A., Krämer, N., Stella, A., and Schön, C. C. (2014). Enhancing genome-enabled prediction by bagging genomic BLUP. *PLoS One* 9:e91693. doi: 10.1371/journal.pone.0091693

González-Camacho, J. M., Crossa, J., Pérez-Rodríguez, P., Ornella, L., and Gianola, D. (2016). Genome-enabled prediction using probabilistic neural network classifiers. *BMC Genom.* 17:208. doi: 10.1186/s12864-016-2553-1

González-Camacho, J. M., de los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125, 759–771. doi: 10.1007/s00122-012-1868-9

Gonzalez-Recio, O., Gianola, D., Long, N., Weigel, K. A., Gonza, O., Rosa, G. J. M., et al. (2008). Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* 178, 2305–2313. doi: 10.1534/genetics.107.084293

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

Heslot, N., Yang, H., Sorrells, M. E., and Jannink, J. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52, 146–160. doi: 10.2135/cropsci2011.06.0297

Hickey, J. M., and Gorjanc, G. (2012). Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. *G3 Genes Genomes Genet.* 2, 425–427. doi: 10.1534/g3.111.001297

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., Howard, R., Liang, Z., Gupta, S. K., Schnable, J. C., and Crossa, J. (2020). Enhancing hybrid prediction in pearl millet using genomic and / or multi- environment phenotypic information of inbreds. *Front. Genet.* 10:1294. doi: 10.3389/fgene.2019.01294

Kimeldorf, G. S., and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* 33, 82–95.

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 Genet.* 5, 569–582. doi: 10.1534/g3.114.016097

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

Morota, G., and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Front. Genet.* 5:363. doi: 10.3389/fgene.2014.00363

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

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., and Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 Genes Genomes Genet.* 2, 1595–1605. doi: 10.1534/g3.112.003665

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome J.* 5:103. doi: 10.3835/plantgenome2012.06.0006

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* 5, 92–102. doi: 10.3835/plantgenome2012.05.0005

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., et al. (2012). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8872–8877. doi: 10.1073/pnas.1120813109

VanRaden, P. M. (2007). Genomic measures of relationship and inbreeding. *Interbull. Bull.* 25, 33–33.

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

Vivek, B. S., Krishna, G. K., Vengadessan, V., Babu, R., Zaidi, P. H., Kha, L. Q., et al. (2017). Use of genomic estimated breeding values results in rapid genetic gains for drought tolerance in Maize. *Plant Genome* 10, 1–8. doi: 10.3835/plantgenome2016.07.0070

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

Zhang, X., Pérez-rodríguez, P., Burgueño, J., Olsen, M., Buckler, E., Atlin, G., et al. (2017). Rapid cycling genomic selection in a multiparental tropical Maize population. *G3 Genes Genomes Genet.* 7, 2315–2326.

Zhao, Y., Gowda, M., Liu, W., Ranc, N., and Reif, J. C. (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theor. Appl. Genet.* 124, 769–776. doi: 10.1007/s00122-011-1745-y

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