frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 25 April 2023
DOI 10.3389/fpls.2023.1139896

OPEN ACCESS

EDITED BY
Peng Wang,
Institute of Botany, Jiangsu Province and (CAS), China

REVIEWED BY
Habtamu Ayalew,
Inari Agriculture, United States
Jiedan Chen,
Tea Research Institute (CAAS), China

*CORRESPONDENCE
Tesfaye Tesso
ttesso@ksu.edu

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 07 January 2023
ACCEPTED 22 March 2023
PUBLISHED 25 April 2023

CITATION
Maulana F, Perumal R, Serba DD and Tesso T (2023) Genomic prediction of hybrid performance in grain sorghum (*Sorghum bicolor* L.). *Front. Plant Sci.* 14:1139896. doi: 10.3389/fpls.2023.1139896

Genomic prediction of hybrid performance in grain sorghum (*Sorghum bicolor* L.)

Frank Maulana[1], Ramasamy Perumal[2], Desalegn D. Serba[3] and Tesfaye Tesso[1*]

[1]Department of Agronomy, Kansas State University, Manhattan, KS, United States, [2]Kansas State University, Agricultural Research Center, Hays, KS, United States, [3]United States Department of Agriculture-Agricultural Research Service (USDA-ARS), U.S. Arid Land Agricultural Research Center, Maricopa, AZ, United States

Genomic selection is expected to improve selection efficiency and genetic gain in breeding programs. The objective of this study was to assess the efficacy of predicting the performance of grain sorghum hybrids using genomic information of parental genotypes. One hundred and two public sorghum inbred parents were genotyped using genotyping-by-sequencing. Ninty-nine of the inbreds were crossed to three tester female parents generating a total of 204 hybrids for evaluation at two environments. The hybrids were sorted in to three sets of 77,59 and 68 and evaluated along with two commercial checks using a randomized complete block design in three replications. The sequence analysis generated 66,265 SNP markers that were used to predict the performance of 204 F1 hybrids resulted from crosses between the parents. Both additive (partial model) and additive and dominance (full model) were constructed and tested using various training population (TP) sizes and cross-validation procedures. Increasing TP size from 41 to 163 increased prediction accuracies for all traits. With the partial model, the five-fold cross validated prediction accuracies ranged from 0.03 for thousand kernel weight (TKW) to 0.58 for grain yield (GY) while it ranged from 0.06 for TKW to 0.67 for GY with the full model. The results suggest that genomic prediction could become an effective tool for predicting the performance of sorghum hybrids based on parental genotypes.

KEYWORDS

genomic-estimated breeding value, ridge regression best linear unbiased prediction, single nucleotide polymorphism, training population, validation population

1 Introduction

Conventional breeding schemes, such as the pedigree method, though time-consuming, remains the most common method used in breeding programs. In sorghum hybrid breeding, populations are initiated from crosses between selected parental sources, and segregating populations are evaluated over multiple seasons, and most promising inbred lines are selected as potential parents often based on their performance in hybrid

combination with other lines. Promising female parents undergo conversion into cytoplasmic male sterility before they can be tested in hybrid combinations. Development of hybrid cultivar is a cumbersome process; it involves synthesis of hundreds of testcross hybrids and evaluation over multiple environments to identify handful of most promising hybrids. It takes significant amount of time and resources to complete the development of hybrid product.

The advent of molecular marker techniques has opened a new horizon for enhancing breeding efficiency through reducing time needed to develop cultivars or improving accuracy during selection (Hasan et al., 2021). Marker-assisted selection (MAS) has shown promise for incorporating quantitative trait loci (QTL) through backcrossing. This approach has been successfully used in different crops, such as yield-related traits in rice (Oryza sativa L.) (Kulkarni et al., 2020), salinity and drought tolerance in maize (Zea mays L.) (Ribaut and Ragot, 2007; Luo et al., 2017), disease resistance in rice (Ni et al., 2015). But MAS has been shown to be more effective for traits under the influence of major effect QTL (Castro et al., 2003; Xu and Crouch, 2008) and thus only a few significant markers with large effects are needed. The small-effect QTL often associated with important agronomic traits are hard to capture using MAS and hence its efficiency for improving complex traits, such as yield, has been limited (Bernardo, 2010). Moreover, many QTL mapped to date are based on simple bi-parental population and their application in MAS is limited to the use of those specific genetic backgrounds as breeding parents. The efficiency of MAS becomes even more limited in hybrid breeding where parental lines that have undergone independent selection are cross combined and tested for expression of the trait in a background different from the one under which they were selected.

Therefore, a less expensive and faster method that allows selection of inbred parents with enhanced hybrid performance is needed. Such method should provide a clue about how the most promising hybrids can be identified without expensive and laborious field testing. Since hybrid performance is the result of putting together of different alleles at several loci associated with the trait of interest (Ben-Israel et al., 2012), new methods should be able to predict how well a given hybrid can do through genetic profiling of its inbred parents (Technow et al., 2012; Cui et al., 2020). Predicting hybrid performance can ultimately reduce the number of hybrids to be evaluated in the field and hence reduce costs associated with synthesizing and phenotyping large number of crosses.

The next generation sequencing (NGS) technologies have provided tools for scanning the entire genome of species instead of few selected genomic regions and capture single nucleotide polymorphisms (SNPs) throughout the genome. Such polymorphisms are often in linkage disequilibrium with alleles responsible for a change in gene functions. Thus, selection approach that takes into account all SNPs across the genome known as genomic selection (GS) may be more powerful than other indirect selection schemes used in the past. Genomic selection is a modified version of MAS that predicts the genetic values of individuals using genome-wide markers without the need for gene and QTL discovery. Unlike MAS, GS permits the use of molecular markers with both major and minor effects on the traits to build the prediction model that is used to predict the phenotypes of untested individuals (Meuwissen et al., 2001). Phenotypes are predicted from the genome information using appropriate prediction models which may provide genomic-estimated breeding values (GEBVs) for each genotype. Prediction of breeding values of the selection candidates is made based on phenotypic data from a set of individuals (training population) randomly drawn from the larger set and marker information of the entire population (Meuwissen et al., 2001).

Genomic selection has been successfully conducted in several crops (Windhausen et al., 2012; Sallam et al., 2015; Spindel et al., 2015). When the accuracy of genomic estimated breeding value (GEBV) is high enough, genomic prediction (GP) can reduce breeding time because the proportion of superior genotypes in a breeding population may increase, and hence accelerate selection gain (Bernardo, 2010; Heffner et al., 2010). To date, several studies have found high prediction accuracies for grain yield and other quantitative traits in maize and wheat (Triticum aestivum L.) using experimental cross-validation (Lorenzana and Bernardo, 2009; Guo et al., 2012). Genomic prediction for single-cross hybrid performance in maize has been shown to outperform marker-assisted recurrent selection (Massman et al., 2013; Zhang et al., 2022). Furthermore, moderate cross-validation prediction accuracies have also been reported for yield and other traits in diverse germplasm and breeding populations of wheat, barley (Hordeum vulgare), and maize (Heffner et al., 2011; Lorenz et al., 2012; Crossa et al., 2014).

In sorghum, GS studies were mainly focused on model training to predict genomic estimated breeding values (GEBVs) of individuals in different sets of populations (Hao et al., 2021). Grain yield and drought adaptation of sorghum hybrids have been assessed using multi-trait model on multi-environment phenotypic performance of 2645 testcross hybrids using their maternal lines genomic and pedigree information (Velazco et al., 2019). They reported that multi-trait genomic evaluation of important agronomic traits enhances genomic prediction of productivity and drought adaptation in grain sorghum. Although full advantage from multi-trait G-BLUP was obtained, only the maternal genomic and pedigree information was considered in this study. Accommodation of genotype-by-environment interaction (GEI) and heterogenous variance of the marker effects through weighted K-BLUP had significant increments in prediction accuracy (Velazco et al., 2020). Comparison of different genomic prediction models incorporating marker-based and pedigree relationships showed higher selection accuracy for marker-based relationship than the pedigree information (Hunt et al., 2018). Moderate to high prediction accuracy for grain composition was obtained for grain sorghum diversity panel and biparental recombinant inbred lines using Bayesian multi-output regressor stacking model than in single-trait single environment models (Sapkota et al., 2020). This approach may be extended to hybrid breeding to replace the extensive hybrid synthesis and evaluation schemes by genome-based prediction. Prediction of hybrid performance based on general (GCA) and specific (SCA) combining abilities applied through genomic-enabled prediction models that incorporated population structure and GEI effects were

used to train classical GCA-SCA-based on genomic (GB) models under a hierarchical Bayesian framework (Fonseca et al., 2021). Using a leave-one-out cross-validation scheme, they effectively predicted hybrid performance and increased prediction accuracy. However, the prediction accuracy of hybrid performance was found to be dependent on repeatability and genetic architecture of the trait, the degree of genetic similarity among parents, the structure of the training set, the method used to perform predictions (genomic or classical GCA-SCA–based models), and the complexity of the models (single or multi-environments). The objective of the present study was to determine whether genomic selection scheme can be effectively used to predict hybrid performance of grain sorghum in the semi-arid mid west with a reasonable accuracy to warrant its application in hybrid breeding program.

2 Materials and methods

2.1 Plant materials

A total of 102 public parental inbred lines, including 99 pollinator lines (fertility-restorer lines) and 3 seed parents (A/B-male sterile lines), bred at Kansas State and Texas A&M Universities, were used in this study. Of these, 59 lines were Acetolactate synthase (ALS) inhibitor herbicide-resistant sorghum pollinator parents (R-lines), 16 were Acetyl co-enzyme-A Carboxylase (ACCase) pollinator parents and 24 were conventional (non-herbicide resistant) pollinators. The lines represented diverse pedigrees in the program and were believed to provide diverse set of hybrids when crossed with three tester females that also represent diversity among the public female inbreds. The female parents were ATx399, ATx3042 and AOK11. A total of 204 F1 hybrids developed from crosses between 99 pollinator lines and the three seed parents were categorized into three subgroups. Group 1 hybrids consisted of crosses between 77 pollinator parents and AOK11 as a female parent, while Group 2 comprised hybrids from crosses between 59 pollinator parents and ATx3042. Group 3 comprised F1 hybrids between 68 pollinator parents and ATx399. Forty-four of the pollinator lines were common across the three populations.

2.2 Field phenotyping

The 204 F1 hybrids were evaluated across four environments at Kansas State University (KSU) Agronomy Research Farm Ashland Bottoms near Manhattan during 2012, 2013 and 2014 seasons and at the Northeast experimental station near Ottawa, KS during 2014. The tests at Ashland bottoms were planted on June 8, 7 and 17 for 2012, 2013 and 2014 seasons, respectively. Field planting at Ottawa was done on June 17, 2014. The experiments were laid in a randomized complete block design with three replications. The gross plot size was 5 m long paired rows spaced 0.75 m apart. On average, the annual precipitation for KSU Agronomy Research Farm Ashland Bottoms was 338, 539 and 576 mm for 2012, 2013 and 2014, respectively.

Data were collected on days to flowering, plant height, grain yield and yield components, including panicle length, panicle weight, panicle yield, number of kernels per panicle, and thousand kernel weight. Days to flowering was determined by recording the number of days from planting to when 50% of plants in each plot reached half-bloom. The plant height was recorded by measuring the distance from soil surface to the tip of the panicle at physiological maturity expressed in centimeters. The grain yield was measured as the weight of the kernels harvested at maturity from each plot recorded in kilograms per hectare.

Prior to harvesting, three panicles from main plants were randomly sampled from each plot for measuring yield components. Mean of the three panicles was used to represent a plot and the moisture content was adjusted to 12.5% for statistical analysis. The panicle length was determined as the mean length of the panicles measured from the base to the tip of the panicle. The panicle weight was recorded as the weight of panicle from individual plant. The panicle yield was measured as the weight of grains threshed from a single panicle. The kernel number was recorded by counting the kernels threshed from each panicle using a laboratory seed counter (Model 850-3, International Marketing and Design Corporation). The thousand kernel weight was determined by measuring the weight of 250 kernels from each panicle and multiplied by four.

2.3 DNA extraction and genotyping

Seeds of the parental lines were planted in the greenhouse at Kansas State University using 96-cell flat trays filled with Metro-mix 360 (Sun Gro, Agawam, MA) growing medium. Two weeks after planting, young leaf tissues were harvested from each line for genomic DNA extraction using the standard cetyltrimethylammonium bromide (CTAB) method (Doyle, 1987). The Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, CA) was used to quantify the concentration of the DNA samples. SNP genotyping and allele calling were carried out using the genotyping-by-sequencing (GBS) platform at the former Institute of Genomic Diversity (currently Cornell Genomic Facility; https://www.biotech.cornell.edu/core-facilities-brc/facilities/genomics-facility) as described in Purcell et al. (2007). The DNA samples were digested with ApeKI restriction enzyme (recognition site: G|CWCG) and 96-plex GBS libraries were constructed as described by Elshire et al. (2011). DNA sequencing was done using either the Illumina Genome Analyzer IIx or Hiseq2000. The Illumina sequencing reads were aligned to the sorghum reference genome v2.1 (http://phytozome.jgi.doe.gov/pz/portal.html; Paterson et al., 2009). SNP calling was conducted using TASSEL 3.0 GBS pipeline (http://www.maizegenetics.net/tassel/; Bradbury et al., 2007; Glaubitz et al., 2014). The GBS data was filtered using minor allele frequency (MAF) of < 5% and missing data of < 20%, which resulted in 66,265 high quality SNPs for downstream analysis. The missing data were imputed using BEAGLE 4.1 (Browning and Browning, 2007). The markers were spread accross the entire genome with the least number of markers 3,950 mapped on to chromosome 7 followed by 4,388 on chromosome 8. The highest number of markers per chromosome of

10,189 was found on chromosome 1 followed by 8,946 on chromosome 2. Chromosomes 3, 4, 5, 6, 9 and 10 had 8,798, 7,162, 5,454, 6,724, 4,965 and 5,689 markers, respectively. The average marker density per chromosome was 6,626.

2.4 Statistical analysis

2.4.1 Variance components and heritability

The variance components were calculated using SAS v.9.3 (SAS Institute, Cary NC, 2011). The following statistical model was used for the analysis of the data across four environments:

$$yijk = \mu + gi + ej + (ge)ij + rk(j) + eijk$$

where *yijk* is the phenotypic observation for ith single cross evaluated in the jth environment, μ is the grand mean for a trait; *gi* represents effect of the ith single cross; *ej* represents the effect of the jth environment; (*ge*)*ij* represents the interaction effect between single cross and environment; *rk*(*j*) represents the effect of replication nested within the jth environment; and *eijk* represents the residual variance. Environment and replication nested within environment effects were modeled as fixed effects while all other effects were treated as random. Error variance was allowed to be heterogeneous among environments.

Broad-sense heritability (H) for each trait was estimated across environments as described by Hallauer et al. (2010):

$$H = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{ge}^2}{e} + \frac{\sigma_e^2}{er}}$$

where σ_g^2, is the genetic variance,σ_{ge}^2 is the genotype-by-environment interaction variance,σ_e^2 is the residual variance, *r* is the number of replications and *e* is the total number of environments.

2.4.2 Population structure and relatedness

To account for population structure that affects prediction accuracy (Riedelsheimer et al., 2013; Lipka et al., 2014), we computed principal component analysis (PCA) on the genotype data of the parental inbred lines using prcomp package in R (Becker et al., 1988). Pairwise genetic distance among the 102 parental inbred lines was estimated by coefficient of co-ancestry directly from 66,265 SNPs among the parents. We also computed kinship matrix as a measure of familial relatedness among the parental inbred lines using the VanRaden method (VanRaden, 2008) in TASSEL 5.2.14 (Bradbury et al., 2007).

2.4.3 Genomic prediction of hybrid performance

Genomic estimated breeding values (GEBVs) were calculated using ridge regression best linear unbiased prediction (RR-BLUP) model implemented in rrBLUP package in R (Endelman, 2011), which assumes that all marker effects are normally distributed and have the same variance (Whittaker et al., 2000). We first generated design matrices for additive and dominance effects from the marker information of the parental lines for the 204 F1 hybrids as described by (Zhao et al., 2013). We predicted the hybrid performance by considering only additive marker effects (partial model) using the following reduced model: y = 1nμ + KAa + ε. We then used both additive and dominance marker effects (full model) in the prediction model to assess if the combined genetic effects would improve the prediction accuracy. The latter was re-run using the full model as follows: y= 1nμ + KAa + KDd +ε; where 1n = a vector of ones, and n and μ represent the number of single cross hybrids and the across environment mean, respectively. KA is the design matrix (n x m) for the additive marker effects, in which m indicates the number of markers, which were coded as -1, 0 and 1, where "-1" and "1" representing homozygous genotypic classes A2A2 and A1A1 and "0" representing heterozygous (A1A2) genotypes. KD is the design matrix for the dominance marker effects coded as 0, 1, 0 with score "0" representing both homozygous genotypes (A2A2 and A1A1) and "1" for the heterozygous (A1A2) genotypes. The additive and dominance effects of the ith marker were represented as a and d, respectively, in the prediction model while ε represents the residual effect for the jth hybrid.

Prediction accuracy, r (ĝ, g), was computed as a measure of the correlation between the observed and predicted phenotypes and divided by the square root of heritability of the trait across environments (Yu et al., 2020). Single-trait prediction accuracy, r (ĝ, g), of hybrid performance was estimated using a five-fold cross-validation (CV) procedure with random sampling method without replacement. The five-fold CV prediction accuracy results were obtained by dividing the 204 F1 hybrids into five random subsets and using 100 iterations. We tested four levels of the TP size (nTP = 41, 82, 122 and 163) to predict the performance of the remaining hybrids as a validation population (VP) using the two models.

3 Results

3.1 Hybrid performance, variance components and heritability

Table 1 summarizes agronomic performance of the 204 F1 hybrids across environments. Flowering time, plant height, and grain yield ranged from 53 to 85 d, from 79.3 to 164 cm, and from 4.0 to 14.5 Mg ha^{-1}, respectively. Overall, each hybrid flowered 65 d after planting, was 111cm tall, and produced 7.9 Mg ha-1 grain yield. Mean panicle length, panicle weight and panicle yield were 25.5 cm, 68.8 g and 47.7 g, respectively. Mean kernel number per panicle and thousand kernel weight were 1,640 and 29.1 g, respectively. Broad-sense heritability varied from 0.23 for grain yield, thousand kernel weight, and panicle weight to 0.81 for flowering time.

3.2 Population structure and relatedness

The first three PCs from the PCA computed across the 102 parental lines accounted for 25.1% of the variance. A plot of PC1 (11.6%), PC2 (7.5%) and PC3 (6.0%) revealed three groups, which generally agrees with pedigree information of the maternal lines (Figure 1). Although most of the lines (97%) were from the KSU

TABLE 1 Summary of eight agronomic traits of sorghum hybrids evaluated across 4 environments at Manhattan in 2012-2014 and Ottawa in 2014 summer seasons.

Trait	Mean*	Range	σ^2_g	σ^2_{ge}	σ^2_e	H
Panicle length (cm)	25.5 ± 2.6	19.1-32.7	1.45	2.3	2.4	0.55
Panicle weight (g)	68.8 ± 14.7	28.4-97.3	17.8	67.9	21.4	0.23
Panicle yield (g)	47.7 ± 9.1	26.3-71.0	25.8	30.5	11.5	0.47
Number of kernels per panicle	1640 ± 307.3	1029-2324	33.6	28.2	12.9	0.52
Thousand kernel weight (g)	29.1 ± 2.4	23.3-38.8	0.43	2.27	3.9	0.23
Days to flowering (days)	65 ± 5.3	53-85	10.2	3.93	6.8	0.81
Plant height (cm)	110.7 ± 14.6	79.3-164	40.9	69.8	52.5	0.47
Grain yield (Mg ha^{-1})	7894.5 ± 2331.3	4014-14475.5	41.1	49.33	31.2	0.23

*Mean with standard errors; σ^2_g, genetic variance; σ^2_{ge}, genotype-by-environment interaction variance; σ^2_e, residual variance; H, broad-sense heritability.

sorghum breeding program, there was clear pattern of genetic differences among the inbred parents. Relative kinship values across pairs of the 102 parental lines ranged from 0 to 1.5 with 98% of the pairs having < 0.5 coefficients and an overall average of 0.1, which suggests that majority of the lines were distantly related (Figure 2).

3.3 Genomic prediction accuracy

Figures 3A, B summarizes the five-fold CV prediction accuracies of hybrids. Both partial model (that incorporated only the additive marker effects) and full model (that used both additive and dominance marker effects) gave moderate to high prediction accuracies of hybrid performance for all traits with the highest accuracy observed for grain yield and the lowest for thousand kernel weight. Prediction accuracy based on additive marker effects alone was slightly lower than when both additive and dominance effects were considered for all traits except for kernel number where the full model had the same level of prediction accuracy with the one based on additive effects alone. For other traits, including panicle length, panicle weight, thousand kernel weight and grain yield, the use of the full model marginally improved prediction accuracy whereas accuracies for plant height and days to flowering were higher with the partial model. For grain yield, which showed an overall higher prediction accuracy, the additive model alone gave r (ĝ, g) of 0.58 versus 0.67 obtained when the full model was used (Figures 3A, B). Although the full model provided better prediction, thousand kernel weight was less predictable for all training population sizes. Other traits, including panicle length and panicle weight also displayed similar trend. On the other hand, the use of the full model decreased the prediction accuracy from 0.24 to 0.17 for panicle length, from 0.18 to 0.14 for days to flowering and from 0.36 to 0.3 for plant height (Figures 3A, B).

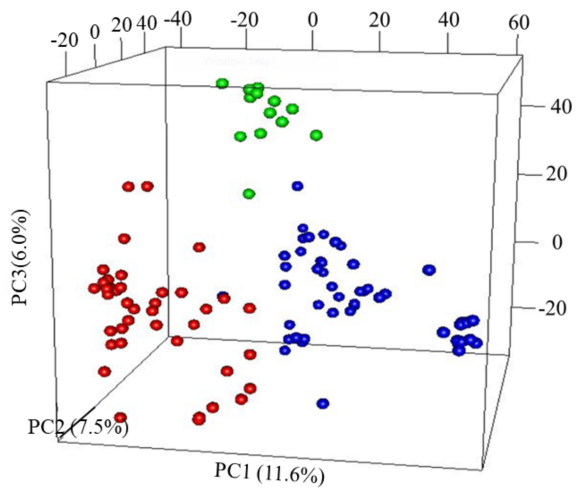

FIGURE 1
Principal component analysis (PCA) results of 102 parental sorghum inbred lines estimated using 66,265 single nucleotide polymorphism (SNP) markers. Subgroup, G1 = Red; G2 = green and G3 = blue.

3.4 Genomic prediction accuracy as influenced by training population size

Prediction of hybrid performance was studied for various TP sizes considering additive marker effects alone as well as for combined additive and dominance effects, the results are summarized in Table 2. The prediction accuracies of hybrid performance for grain yield and yield components based on additive marker effects alone increased as the number of individuals assigned to the TP increased for all traits. Increasing the TP size from 41 (20%) to 163 (80%) increased the prediction accuracy for panicle length, panicle weight, panicle yield and kernel weight by 20, 100, 175 and 89%, respectively. Other traits, including days to flowering, plant height and grain yield had their prediction accuracies increased by 156, 65 and 28%, respectively, when the TP sizes were increased. Prediction accuracy for different traits based on additive effect model was markedly different with grain yield and other yield component traits, namely, panicle weight, kernel number and plant height having higher prediction accuracies while thousand kernel weight, panicle yield and days to flowering showing the lowest prediction accuracy. Similarly, the prediction accuracy of hybrid performance under both additive and

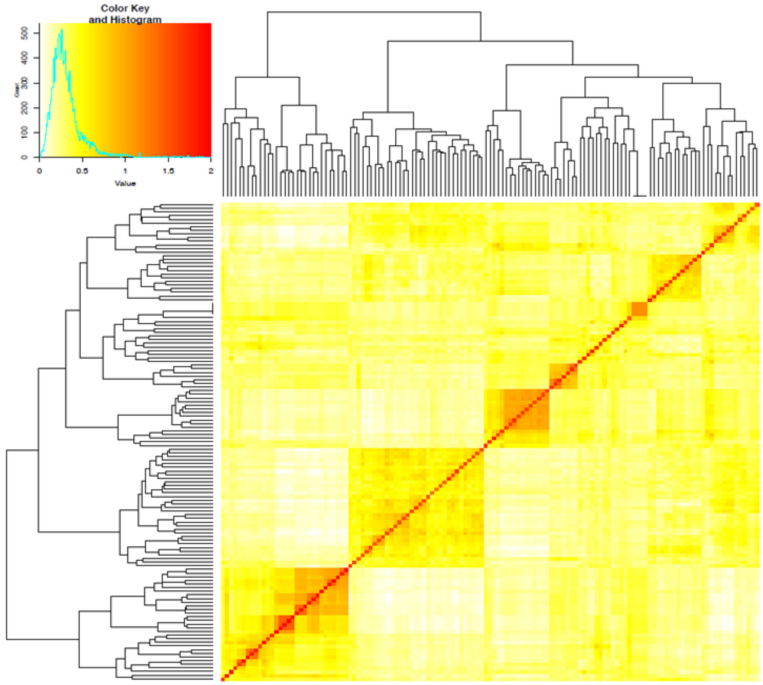

FIGURE 2
Heatmap of pairwise kinship matrix values estimated using VanRaden algorithm for 66,265 single nucleotide polymorphic (SNP) markers among 102 sorghum parental inbred lines. The distribution of coefficients of co-ancestry is shown by the color histogram, and the stronger red color indicates the individuals that are more related to each other.

dominance model was similar to when only the additive effects were considered and for all traits the accuracy increased as the TP size increased.

Prediction accuracy of hybrid performance using five-fold CV where TP and VP are related by common males or females using the partial model are presented in Table 3. When relatedness was only due to common male parental lines in the TP and the VP, the prediction accuracy of hybrid performance for different traits ranged from 0.06 for thousand kernel weight to 0.59 for grain yield. On the other hand, when relatedness was due to common female parents, the average prediction accuracy ranged from 0.17 for panicle weight to 0.56 for grain yield (Table 3).

4 Discussion

The recent breakthrough in genetic marker technology and bioinformatics tools integrating DNA markers with phenotypes has expanded the knowledge of marker effect on phenotype; opening way for MAS to enhance breeding efficiency. While the applicability of MAS was limited to QTL with large effect, a further development based on next-generation sequencing has provided a more powerful tool, genomic selection (GS), to facilitate selection for small effect QTL affecting key traits of agronomic importance (Arruda et al., 2016; Cerrudo et al., 2018). Because GS accounts for all loci with both major and minor effects on the trait, it is expected to address some of the shortcomings of MAS (Cerrudo et al., 2018).

In the present study, GS was used to predict F1 hybrid performance with respect to eight different agronomic traits of sorghum. Prior to building the genomic prediction model, structure analysis was conducted to determine population structure and familial relatedness. The kinship values among the lines were expectedly low and it may be the result of a deliberate attempt by the breeding programs to diversify parental sources in order to maximize hybrid vigor. The grain yield values (7.9 to 14.5 t ha^{-1}) observed in this study may be partly the result of increased heterosis that resulted from the low kinship coefficients among the lines.

Genomic selection utilizes phenotype and genomic data of subset of a population (training population, TP) to predict the performance of the selection candidates based on their genotype only. For GS to be effective, it is very important that high quality genotype data is obtained on the entire population and good quality phenotype data on the TP. This study also looked at the effect of TP size on prediction accuracy of hybrid performance and compared two prediction models, one based on additive marker effects only, and the other considering both additive and dominance effects, to predict F1 hybrid performance in sorghum. The additive and dominance allelic effects were estimated for each marker and used to calculate predicted phenotypes (GEBVs) for untested F1 hybrids using RR-BLUP genomic prediction based on an infinitesimal model where all predictors are maintained in the analysis. This model gave higher prediction accuracies in previous studies (Habier et al., 2007; Zhao et al., 2013).

Previous studies have shown that in cross-validation schemes, prediction accuracy can be overestimated if both TP and validation population (VP) sets contain related lines (Edwards et al., 2019; Lozada et al., 2019; Fraslin et al., 2022). Therefore, in this study, principal component analysis (PCA) was performed on the parental

10.3389/fpls.2023.1139896

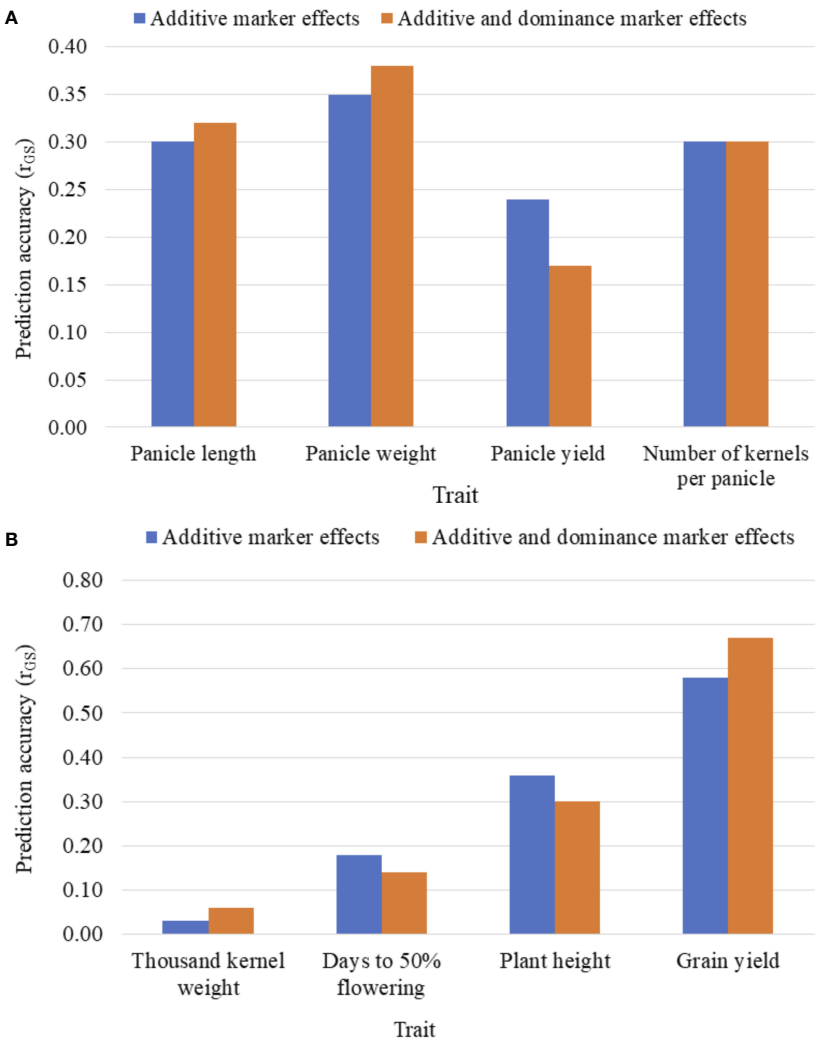

FIGURE 3
Five-fold cross-validated prediction accuracy, r (ĝ, g), of sorghum hybrid performance considering additive marker effects alone versus additive and dominance marker effects: **(A)** panicle characteristics, and **(B)** phenology, plant height, grain yield and seed weight. Prediction accuracy was assessed using 163 and 41 F1 hybrids as the training population (TP) and validation population (VP), respectively.

TABLE 2 Prediction accuracy of hybrid performance for eight agronomic traits as affected by training population size considering additive effects/additive and dominance effects of the markers in the model.

Trait	Prediction accuracy, r (ĝ, g) at different Training population sizes (n_{TP})*			
	n_{TP} = 41	n_{TP} = 82	n_{TP} = 122	n_{TP} = 163
Panicle length	0.25/0.20	0.28/0.24	0.28/0.25	0.30/0.28
Panicle weight	0.19/0.15	0.26/0.18	0.33/0.21	0.38/0.28
Panicle yield	0.08/0.09	0.12/0.15	0.17/0.17	0.22/0.27
Number of kernels per panicle	0.18/0.17	0.26/0.22	0.29/0.24	0.34/0.29
Thousand kernel weight	0.01/0.03	0.02/0.02	0.04/0.02	0.12/0.18
Days to flowering	0.09/0.06	0.12/0.10	0.14/0.13	0.23/0.14
Plant height	0.23/0.26	0.28/0.30	0.33/0.33	0.38/0.34
Grain yield	0.46/0.49	0.53/0.52	0.56/0.56	0.59/0.58

*Additive marker effect and additive & dominance marker effect separated by forward slash, respectively.

TABLE 3 Prediction accuracy of hybrid performance using five-fold cross validation where training sets (n_{TP} = 136, 77) and validation sets (n_{TP} = 68, 127) are related by common males and females, respectively.

Trait	Related by common males, r(ĝ,g)	Related by common females, r(ĝ,g)
Panicle length	0.28	0.33
Panicle weight	0.35	0.17
Panicle yield	0.18	0.19
Number of kernels per panicle	0.26	0.23
Thousand kernel weight	0.06	0.22
Days to flowering	0.16	0.27
Plant height	0.34	0.31
Grain yield	0.59	0.56

lines to determine the genetic structure of the lines before genomic prediction analysis was performed. The results show that the parental lines are structured into three subgroups (G1, G2 and G3 in Figure 3) to some extent based on the maternal lines. Following the PCA results, an alternative cross-validation was considered in which the prediction accuracy of hybrid performance was assessed by assigning F1 hybrids in the TP and VP either with common male or female parents.

In this study, prediction accuracy was markedly different for different traits with grain yield having more than 50% accuracy and thousand kernel weight consistently the lowest. Increase in TP size improved prediction accuracy for all traits but the extent of the increase was different for different traits. Similar results have been reported in previous studies in other crops (Asoro et al., 2011; Heffner et al., 2011; Lorenz et al., 2012; Crossa et al., 2014; Jan et al., 2016). Jan et al. (2016) reported increased prediction accuracies in canola with increase in TP size and no significant increase in accuracy was observed after assigning more than 70% of hybrids in the TP.

Again, grain yield consistently had the highest five-fold CV prediction accuracy among the traits assessed in this study. This result corroborates previous studies that have also reported high prediction accuracy of grain yield in wheat (Crossa et al., 2010; Heffner et al., 2011; Heslot et al., 2012; Zhao et al., 2013) and biomass yield for maize hybrids (De los Campos et al., 2009; Crossa et al., 2010; Albrecht et al., 2011; Crossa et al., 2011; Gonzaílez-Camacho et al., 2012). Furthermore, higher prediction accuracies of hybrid performance were observed for many of the traits with the full model (both additive and dominance effects) than with the reduced model (additive effects only). The result agrees with previous simulation study on maize (Technow et al., 2012) where higher prediction accuracy was reported when dominance effects of the markers were considered in the model. Contrasting results were reported in hybrid wheat by Zhao et al. (2013) where higher prediction accuracies of hybrid performance was observed when dominance effects were not considered in the model. They attributed this to small population size (90 hybrids) used in their study arguing that dominance model is more sensitive to the size of available data for training, suggesting that the dominance effects on prediction accuracy can be better captured when the population size is large. In the present study, 204 F1 sorghum hybrids were used, substantially higher than 90 hybrids studied by Zhao et al. (2013), and perhaps that has contributed to higher prediction accuracies when dominance effects were considered in the model, at least for some of the traits. But perhaps due to the same reseaon for wheat (Zhao et al., 2013), the full model resulted in reduced prediction accuracy for panicle length, days to flowering and plant height in the current study.

Conclusion

This study has shown that it is possible to predict the performance of untested sorghum hybrids for important agronomic traits such as grain yield solely based on the genotype information by using a genomic prediction model. Thus, GS may become a viable tool for predicting the performance of sorghum hybrids prior to committing resources for expensive phenotyping. This intern may help to significantly reduce the number of hybrids to be evaluated and costs associated with phenotyping a large number of hybrids in the field. The fact that genotyping and sequencing costs have been decreasing and knowledge of computational biology expanding, it is becoming possible that public breeding programs can affordably deploy genomic selection platforms to add efficiency and reduce the overall cost of developing a hybrid technology.

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. The data presented in the study are deposited in the github repository, accession number https://github.com/framau2023/SNP_marker_data_GP.

Author contributions

TT conceived the work, acquired funding support for the project and provided supervision, and edited the draft manuscript. FM conducted field experiment, data analysis

preparation of the draft manuscript. RP and DS co-edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by Kansas Grain Sorghum Commission.

Acknowledgments

The authors would like to thank the Kansas Grain Sorghum Commission for financial support of this project and Kansas State University for providing equipment, facilities, and services used to carry out this study. This is Contribution no. xx-xxx-x from Kansas agricultural Experiment Station.

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

References

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

Arruda, M. P., Lipka, A. E., Brown, P. J., Krill, A. M., Turber, 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

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

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The new s language* (Pacific Grove, PA: Wadsworth & Brooks/Cole).

Ben-Israel, I., Kilian, B., Nida, H., and Fridman, E. (2012). Heterotic trait locus (HTL) mapping identifies intra-locus interactions that underlie reproductive hybrid vigor in *Sorghum bicolor*. *PLoS One* 7 (6), e38993. doi: 10.1371/journal.pone.0038993

Bernardo, R. (2010). *Breeding for quantitative traits in plants* (Woodbury, MN: Stemma Press).

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

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi: 10.1086/521987

Castro, A. J., Capettini, F., Corey, A. E., Filichkina, T., Hayes, P. M., Kleinhofs, A., et al. (2003). Mapping and pyramiding of qualitative and quantitative resistance to stripe rust in barley. *Theor. Appl. Genet.* 107, 922–930. doi: 10.1007/s00122-003-1329-6

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. doi: 10.3389/fpls.2018.00366

Crossa, J., De Los Campos, G., Peírez, P., Gianola, D., Burguenão, 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., Peírez, 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* 112 (1), 48–60. doi: 10.1038/hdy.2013.16

Cui, Y., Li, R., Li, G., Zhang, F., Zhu, T., Zhang, Q., et al. (2020). Hybrid breeding of rice *via* genomic selection. *Plant Biotechnol. J.* 18 (1), 57–67. doi: 10.1111/pbi.13170

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 pedigrees. *Genetics* 182, 375–385. doi: 10.1534/genetics.109.101501

Doyle, J. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Edwards, S. M., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., et al. (2019). The effects of training population design on genomic prediction accuracy in wheat. *Theor. Appl. Genet.* 132 (7), 1943–1952. doi: 10.1007/s00122-019-03327-y

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6 (5), e19379. doi: 10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrBLUP. *Plant Genome* 4 (3), 250–255. doi: 10.3835/plantgenome2011.08.0024

Fonseca, J. M. O., Klein, P. E., Crossa, J., Pacheco, A., Perez-Rodriguez, P., Perumal, R., et al. (2021). Assessing combining abilities, genomic data, and genotype × environment interactions to predict hybrid grain sorghum performance. *Plant Genome* 14 (3), e20127. doi: 10.1002/tpg2.20127

Fraslin, C., Yáñez, J. M., Robledo, D., and Houston, R. D. (2022). The impact of genetic relationship between training and validation populations on genomic prediction accuracy in Atlantic salmon. *Aquac. Rep.* 23, 101033. doi: 10.1016/j.aqrep.2022.101033

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 (2), e90346. doi: 10.1371/journal.pone.0090346

Gonzaílez-Camacho, J. M., de los Campos, G., Peírez, P., Gianola, D., Cairns, J., 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, 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 (2), 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 information on genome-assisted breeding values. *Genetics* 177 (4), 2389–2397. doi: 10.1534/genetics.107.081190

Hallauer, A. R., Carena, M. J., and Miranda Filho, J. B. (2010). *Quantitative genetics in maize breeding* (Ames, IA: Iowa State University Press).

Hao, H., Li, Z., Leng, C., Lu, C., Luo, H., Liu, Y., et al. (2021). Sorghum breeding in the genomic era: opportunities and challenges. *Theor. Appl. Genet.* 134 (7), 1899–1924. doi: 10.1007/s00122-021-03789-z

Hasan, N., Choudhary, S., Naaz, N., Sharma, N., and Laskar, R. A. (2021). Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. *J. Genet. Eng. Biotechnol.* 19, 128. doi: 10.1186/s43141-021-00231-1

Heffner, E. L., Jannink, J. L., and Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* 4 (1), 65–75. doi: 10.3835/plantgenome.2010.12.0029

Heffner, E. L., Lorenz, A. J., Jannink, J., 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

Heslot, N., Yang, H. P., Sorrels, 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

Hunt, C. H., van Eeuwijk, F. A., Mace, E. S., Hayes, B. J., and Jordan, D. R. (2018). Development of genomic prediction in sorghum. *Crop Sci.* 58 (2), 690–700. doi: 10.2135/cropsci2017.08.0469

Jan, H. U., Abbadi, A., Lücke, S., Nichols, R. A., and Snowdon, R. J. (2016). Genomic prediction of testcross performance in canola (*Brassica napus*). *PLoS One* 11 (1), e0147769. doi: 10.1371/journal.pone.0147769

Kulkarni, S. R., Balachandran, S. M., Ulaganathan, K., Balakrishnan, D., Praveen, M., Prasad, A. S., et al. (2020). Molecular mapping of QTLs for yield related traits in recombinant inbred line (RIL) population derived from the popular rice hybrid KRH-2 and their validation through SNP genotyping. *Sci. Rep.* 10 (1), 1–21. doi: 10.1038/s41598-020-70637-3

Lipka, A. E., Lu, F., Cherney, J. H., Buckler, E. S., Casler, M. D., and Costich, D. E. (2014). Accelerating the switchgrass (*Panicum virgatum* l.) breeding cycle using genomic selection approaches. *PLoS One* 9 (11), e112227. doi: 10.1371/journal.pone.0112227

Lorenz, A. J., Smith, K. P., and Jannink, J. L. (2012). Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. *Crop Sci.* 52, 1609–1621. doi: 10.2135/cropsci2011.09.0503

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

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), 82. doi: 10.1186/s12863-019-0785-1

Luo, M., Zhao, Y., Zhang, R., Xing, J., Duan, M., Li, J., et al. (2017). Mapping of a major QTL for salt tolerance of mature field-grown maize plants based on SNP markers. *BMC Plant Biol.* 17 (1), 1–10. doi: 10.1186/s12870-017-1090-7

Massman, J. M., Gordillo, A., Lorenzana, R. E., and Bernardo, R. (2013). Genome-wide predictions from maize single-cross data. *Theor. Appl. Genet.* 126, 13–22. doi: 10.1007/s00122-012-1955-y

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic values using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Ni, D., Song, F., Ni, J., Zhang, A., Wang, C., Zhao, K., et al. (2015). Marker-assisted selection of two-line hybrid rice for disease resistance to rice blast and bacterial blight. *Field Crops Res.* 184, 1–8. doi: 10.1016/j.fcr.2015.07.018

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The sorghum bicolor genome and the diversification of grasses. *Nature* 457 (7229), 551–556. doi: 10.1038/nature07723

Purcell, S., Neale, B., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi: 10.1086/519795

Ribaut, J. M., and Ragot, M. (2007). Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. *J. Exp. Bot.* 58 (2), 351–360. doi: 10.1371/journal.pone.0112227

Riedelsheimer, C., Endelman, J. B., Stange, M., Sorrells, M. E., Jannink, J. L., and Melchinger, A. E. (2013). Genomic predictability of interconnected bi-parental maize populations. *Genetics* 194 (2), 493–503. doi: 10.1534/genetics.113.150227

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

Sapkota, S., Boatwright, J. L., Jordan, K., Boyles, R., and Kresovich, S. (2020). Multi-trait regressor stacking increased genomic prediction accuracy of sorghum grain composition. *Agronomy* 10 (9), 1221-1235 doi: 10.3390/agronomy10091221

SAS Institute (2011). *The SAS system for windows. v.9.3* (Cary, NC: SAS Inst.).

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

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

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

Velazco, J. G., Jordan, D. R., Hunt, C. H., Mace, E. S., and van Eeuwijk, F. A. (2020). Genomic prediction for broad and specific adaptation in sorghum accommodating differential variances of SNP effects. *Crop Sci.* 60 (5), 2328–2342. doi: 10.1002/csc2.20221

Velazco, J. G., Jordan, D. R., Mace, E. S., Hunt, C. H., Malosetti, M., and van Eeuwijk, F. A. (2019). Genomic prediction of grain yield and drought-adaptation capacity in sorghum is enhanced by multi-trait analysis. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00997

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

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

Xu, Y., and Crouch, J. H. (2008). Marker-assisted selection in plant breeding: From publications to practice. *Crop Sci.* 48, 391–407. doi: 10.2135/cropsci2007.04.0191

Yu, X., Leiboff, S., Li, X., Guo, T., Ronning, N., Zhang, X., et al. (2020). Genomic prediction of maize microphenotypes provides insights for optimizing selection and mining diversity. *Plant Biotechnol. J.* 18 (12), 2456–2465. doi: 10.1111/pbi.13420

Zhang, A., Pérez-Rodríguez, P., San Vicente, F., Palacios-Rojas, N., Dhliwayo, T., Liu, Y., et al. (2022). Genomic prediction of the performance of hybrids and the combining abilities for line by tester trials in maize. *Crop J.* 10 (1), 109–116. doi: 10.1016/j.cj.2021.04.007

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Sci.* 53 (3), 802–810. doi: 10.2135/cropsci2012.08.0463