Core ideas

- In this empirical sweet sorghum breeding population, phenotyping replication was the dominant factor explaining variation in genomic predictive ability across traits and environments.
- The benefit of larger training populations and greater training-validation genomic relatedness increased when phenotype estimates were based on more replicates.
- Grain yield, the most environmentally sensitive trait evaluated, showed the largest response to improved replication and training-population design.
- Bayesian models, rrBLUP, and GBLUP showed similar predictive abilities across traits and environments, suggesting that phenotyping and experimental design may be more important than model complexity.

Phenotyping replication is a major determinant of genomic predictive ability in sweet sorghum (Sorghum bicolor [L.] Moench)

Jean Rigaud Charles[1,2], Brian Rice[4], Geoffrey Morris[3], Thierry Tovignan[5] and Gael Pressoir[1*]

1 Chibas, Centre Haïtien d'Innovation en Biotechnologies et pour une Agriculture Soutenable

2 Université Quisqueya, Port-au-Prince, Haïti

3 Soil and Crop Sciences, College of Agricultural Sciences, Colorado State University

4 University of Nebraska, Department of Agronomy and Horticulture

5 AgroUniQ Lab, Université Quisqueya, Port-au-Prince, Haïti

*Corresponding author (gael.pressoir@chibas-haiti.org; jeanrigaud.charles@chibas-haiti.org).

ARTICLE SUMMARY

Genomic selection can accelerate breeding only when the phenotypes used to train prediction models have high reliability. Using a sweet sorghum breeding population evaluated in three Haitian field environments, we quantified how replication number, training population size, training-validation genomic relatedness, and prediction model affected genomic predictive ability for grain yield, plant height, stem weight, and total soluble solids. Replication increased genomic heritability and predictive ability for all traits, with the strongest effects for grain yield. Larger and more connected training populations improved prediction, mainly when replication was adequate. These results provide practical guidance for resource-limited breeding programs.

ABSTRACT

Genomic selection can increase the rate of genetic gain in crop breeding programs, but its effectiveness depends on the reliability of phenotypic data, the size and composition of the training population (TP), and the statistical model used to estimate genomic breeding values. These design choices are especially important in resource-limited breeding programs, where additional replication, larger TPs, and more extensive genotyping compete for the same resources. Using empirical data from a sweet sorghum [Sorghum bicolor (L.) Moench] breeding population, developed by CHIBAS, we evaluated the effects of phenotyping replication, TP size, training-validation genomic relatedness, and genomic prediction (GP) model on predictive ability (PA). Grain yield, plant height, stem weight, and total soluble solids were evaluated across three field environments. Few studies in sorghum have examined these factors together with comparable empirical rigor.

Increasing replication improved genomic heritability and PA for all traits and environments, with the largest gains observed for grain yield. Larger TPs and increased training-validation genomic relatedness also improved PA, but their effects were most significant when phenotype estimates were based on multiple replicates. GP models showed largely comparable PAs across all evaluated traits. Different models produced similar PA, with a few exceptions. These findings provide practical guidance for optimizing genomic selection in resource-limited sorghum breeding programs.

ABBREVIATIONS

BGLR, Bayesian Generalized Linear Regression; BLUE, best linear unbiased estimate; BRR, Bayesian ridge regression; CHIBAS, Centre Haitien d'Innovation en Biotechnologies et pour une Agriculture Soutenable; GBLUP, genomic best linear unbiased prediction; GBS, genotyping-by-sequencing; GP, genomic prediction; GRM, genomic relationship matrix; GS, genomic selection; LOGO-CV, leave-one-group-out cross-validation; PA, predictive ability; rrBLUP, ridge-regression best linear unbiased prediction; SNP, single nucleotide polymorphism; TP, training population; VP, validation population.

INTRODUCTION

Genomic selection (GS) has become an important approach for accelerating genetic gain in plant breeding because it enables the prediction of genetic merit from genome-wide marker data rather than from a small number of marker-trait associations. The approach was first formalized for dense marker maps by Meuwissen et al. (2001) and has since been extended to many crop species as genotyping costs have decreased through next-generation sequencing and

genotyping-by-sequencing platforms (Elshire et al., 2011; Wetterstrand, 2020; Satam et al., 2023). GS enables breeders to screen larger populations and select a smaller, more elite subset of individuals, thereby increasing selection intensity and potential genetic gain.

The realized value of GS depends on prediction ability (PA), and is influenced by marker density, genetic architecture, trait heritability, population structure, training population (TP) size, TP composition, and the quality of phenotypic data used to train prediction models (Heffner et al., 2011; Lorenz, 2013; Tayeh et al., 2015; Norman et al., 2018; Zhang et al., 2019; Alemu et al., 2024). Among these factors, three are directly controlled by breeders at the design stage: the size of the TP, the degree of genetic relatedness between training and selection candidates, and the intensity of field phenotyping. These factors are not independent. A fixed budget may allow breeders to evaluate many genotypes with limited replication or fewer genotypes with more precise phenotypic estimates. Therefore, the practical question is not whether more data improve prediction, but which type of additional information most improves prediction under field conditions.

TP size usually improves PA because larger populations sample more recombination events, capture more allelic diversity, and support more stable estimation of marker effects or genomic relationships (VanRaden et al., 2009; Lorenzana and Bernardo, 2009; Heffner et al., 2011; Technow et al., 2013; Lorenz, 2013; Endelman et al., 2014; Lorenz and Smith, 2015; Cericola et al., 2017; Tan et al., 2017; Lozada et al., 2019). However, increases in TP size may show diminishing returns, especially when additional individuals are weakly related to the prediction candidates or when phenotypic error remains high. Similarly, TP composition strongly affects prediction because PA tends to increase when the TP and prediction candidates are genetically

related (Lorenz et al., 2011; Edwards et al., 2019; Werner et al., 2020; Fraslin et al., 2022). Designing TPs that balance diversity, relatedness, and cost is therefore an ongoing concern in applied genomic-assisted breeding.

Finally, PA is affected by the level of error in phenotype data. For highly polygenic traits strongly influenced by the environment, such as grain yield, low-precision phenotypes can limit prediction even when marker density and TP size are adequate (Falconer and Mackay, 2009). Additional replications reduce experimental error and improve the reliability of genotype means, thereby increasing the heritability of the target phenotype (Bos, 1983; Wricke and Weber, 1986; Falconer and Mackay, 2009; Lorenz, 2013; Lourenço et al., 2020; Yan, 2021). Yet replication is costly. For small breeding programs, allocating resources to replication can reduce the number of genotypes evaluated or the number of environments sampled. Empirical evidence is therefore needed to determine how replication interacts with TP size and composition under realistic breeding conditions.

Sweet sorghum [Sorghum bicolor (L.) Moench] is a multipurpose crop with potential value for grain, forage, stem biomass, sugar content, and bioenergy production. In Haiti, sorghumis relevant for crop diversification, food security, and the development of climate-resilient agricultural systems. Previous work has demonstrated the potential of GS for sorghum improvement in Haiti and highlighted the need to optimize GS for local breeding populations and environments (Muleta et al., 2019; Charles et al., 2024). However, empirical studies simultaneously evaluating replication intensity, TP size, TP composition, and model choice remain limited, particularly in small or resource-constrained breeding programs.

Here, we used an empirical multi-environment dataset from a CHIBAS sweet sorghum breeding population to quantify the relative contributions of different factors in determining genomic PA. Our objective was to identify practical design principles for improving GS in sorghum breeding programs operating under limited resources.

MATERIALS AND METHODS

Plant materials used in this study

The study used 250 sweet sorghum breeding lines derived from three CHIBAS sub-populations previously described by Charles et al. (2024). The first sub-population originated from recurrent phenotypic selection following a cross between CIRAD-ms3, an advanced selection from CIRAD, and ICSV 25280 from ICRISAT. This population was advanced through recurrent phenotypic selection to the S2 generation. The second sub-population consisted of multiple S0-derived families generated by crossing superior S0 plants from the CIRAD-ms3 × ICSV 25280 background to Coloudo Nevado, 00-SB-FSDT-427, IS23563, WILEY, CIR-1/OG2-4G-1G-M-M, PCR-2>723C-1-M-1, Papesek, and Dekabes (IS 23572). The third sub-population included lines derived from crosses between Papesek (Centa S3) and Dekabes (IS 23572).

Field Environments and Experimental design

The 250 genotypes were evaluated in 2017 in randomized complete block designs with four blocks per location in three environments in Haiti. The trials were conducted in the West region of Haiti (18.64526 N, 72.9985 W) on an Inceptisol. Environment 1 was planted in September 2017 and irrigated weekly throughout the growing season. Environments 2 and 3 were planted in

November 2017. Environment 2 was irrigated weekly; soil moisture was 77% one week after planting and did not fall below 50% during the season. Environment 3 was managed primarily under residual soil moisture, with rainfall increasing soil moisture to approximately 76% during the hard dough stage. The three environments differed in planting date and water availability, ranging from fully irrigated to residual-moisture conditions (Table S1). Plots consisted of four rows, with 0.70 m between rows and 0.25 m within rows. Each plot measured 2.8 m by 3.5 m. Further details about the environments are previously published (Charles et al., 2024).

Phenotype data collection

The population was phenotyped for four traits: (i) Stem weight (kg/ha) was measured by weighing leafless stems using a digital scale, with data collected from the two central rows of each plot (border plants excluded); (ii) grain yield (tons/ha) was estimated as 80% of panicle weight at physiological maturity, using the same plot sampling method; (iii) plant height (cm) was measured from the ground to flag leaf sheath on 10 plants selected randomly in the two central rows using a graduated polyvinyl chloride pipe; (iv) concentration of total soluble solids (^{o}Brix) was measured with an Atago brand refractometer graduated from 0 to 33 %. This measure was determined by adding a few drops of subsample sorghum juice from each plot on the slide of the refractometer.

Genotyping and marker filtering

The population was genotyped using genotyping-by-sequencing (GBS) at Kansas State University following the general approach of Elshire et al. (2011). DNA samples were digested with ApeKI and sequenced on an Illumina NextSeq500. The TASSEL-GBS pipeline version 4.3.7 was used for SNP calling, and sequence tags were aligned to the sorghum reference

genome Sbicolor_313_v3.0. A total of 160,066 raw SNPs were called. Markers were removed if they were non-biallelic, had a missing rate greater than 0.30, had minor allele frequency less than 0.01, or had heterozygosity greater than 0.20. Missing marker genotypes in the retained dataset were imputed using marker means. After filtering and imputation, 28,785 SNPs were used for genomic analyses.

Phenotypic adjustment and replication scenarios

In order to evaluate the effect of replication on GP accuracy, different scenarios were generated across the three environments. For each environment, each individual replicate was used in the following scenarios: only one replicate, six combinations of two replicates, and four combinations of three replicates. The complete dataset with four replicates was also included as the highest replication level.

Phenotypic data were adjusted for replicate effects using univariate linear models fitted with the base R *lm()* function. Best linear unbiased estimates (BLUEs) of genotype performance were obtained using the following model:

$$y = 1\mu + X_G b_G + X_R b_R + \epsilon$$

where y is the vector of phenotypic observations, μ is the overall mean, X_G and X_R are incidence matrices for genotype and replicate effects, b_G and b_R are vectors of fixed effects, and ϵ is the residual term assumed to follow $\varepsilon \sim N(0, I\sigma^2_\varepsilon)$. For scenarios with a single replicate, genotype performance corresponded to the single observed phenotypic value per genotype, as no

replication was available to estimate adjusted means. After fitting the model, we obtained the BLUEs for the genotypic effects which were later used in the GP model.

Genomic heritability estimates

Marker-based narrow-sense genomic heritability (h_g^2) was estimated for each trait, environment, and replication level using ridge regression best linear unbiased prediction (rrBLUP) implemented in the R package *rrBLUP*. Variance components were obtained from a mixed model fitted via restricted maximum likelihood (REML), where SNP effects were treated as random effects. Briefly, phenotypes were modeled as:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \boldsymbol{\epsilon}$$

where **y** is the vector of phenotypic values (BLUEs) for n individuals. μ is the overall mean, **Z** is an n*m design matrix containing the allele dosage at the m marker loci, **u** is an m*1 vector of m allele substitution effects, and $\boldsymbol{\epsilon}$ is the n*1 vector of residual effects.

Unlike the GP analysis, where marker effects were used to estimate genomic best linear unbiased predictions (GBLUPs), here the objective was variance decomposition rather than prediction. Therefore, genomic heritability was derived from the estimated variance components obtained from the rrBLUP mixed model as : $h_g^2 = p\sigma_u^2/(p\sigma_u^2 + \sigma_e^2)$;

where p is the number of markers, σ_u^2 is the marker effect variance, and σ_e^2 is the residual variance. Heritability was estimated separately for each environment and replication scenario, including single-replicate datasets, all pairwise and three-way replicate combinations, and the

full multi-replicate dataset. For multi-replicate scenarios, phenotypic values were averaged per genotype prior to model fitting.

Genomic prediction and predictive ability

GP was performed using rrBLUP to estimate GBLUPs. The mixed.solve function from the rrBLUP package was used to fit the following model with a fixed effect:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is the vector of BLUEs for n individuals. μ is the overall mean, $\mathbf{Z}$ is an n*m design matrix containing the allele dosage at the m marker loci, $\mathbf{u}$ is an m*1 vector of m allele substitution effects, and $\boldsymbol{\epsilon}$ is the n*1 vector of residual effects. The statistical model underlying the rrBLUP approach assumes that allele substitution effects and residual effects are normally distributed (Endelman, 2011):

$$u \sim N(0, I\sigma^2_u) \text{ and } \boldsymbol{\epsilon} \sim N(0, I\sigma^2_{\epsilon})$$

Phenotypic inputs consisted of BLUEs estimated separately for each replication scenario prior to cross-validation. Because BLUE estimation used the complete phenotypic dataset within each scenario, validation individuals contributed indirectly to variance estimation and phenotype adjustment. Consequently, the reported PAs are expected to be upwardly biased relative to fully independent prediction pipelines in which phenotype adjustment is performed separately within each training set. However, all GP scenarios evaluated in this study were affected equally by this dependency structure. Therefore, the analyses should be interpreted primarily as relative

comparisons among experimental design factors (replication intensity, TP size, genomic relatedness, and prediction model) rather than as unbiased estimates of operational PA.

PA was defined as the Pearson correlation between observed phenotypic values in the validation population (VP) and predicted genomic values. Because the observed validation values are BLUEs or single-replicate observations rather than true genetic values, we use the term "predictive ability" throughout, instead of "prediction accuracy." Since all scenarios shared the same phenotype-adjustment framework, relative differences among experimental design factors remain informative despite a possible upward bias in absolute PA. To evaluate factors affecting GP performance, three independent analyses were performed using the same set of phenotypic BLUEs and genome-wide SNP markers: (i) replication number, (ii) TP size, and (iii) TP composition.

Testing the effect of phenotyping replication intensity

The effect of replication number on GP accuracy was assessed using a repeated random cross-validation approach conducted independently within each environment and scenario. For each replication level, all available genotypes were included in the analysis.

Within each environment × replication combination, genotypes were randomly partitioned into training and VPs, with 80% of individuals assigned to the training set and the remaining 20% assigned to the validation set. This random partitioning procedure was repeated 100 times to account for sampling variability and obtain robust estimates of PA. For each iteration, marker effects were estimated using the TP and subsequently used to predict genomic values for individuals in the VP. PA was calculated as the Pearson correlation between observed phenotypic

BLUEs (or single-replicate observations for one-replicate scenarios) and predicted genomic values in the VP. This strategy was used to generate distributions of PA for all testing scenarios in this study.

Testing the joint effects of TP size and phenotyping replication

The full set of 250 genotypes was randomly partitioned into five equal sized folds (50 genotypes per fold). A random seed (20241015) was set prior to assignment, and the same fold assignments were used for all traits, environments, and replication scenarios to ensure comparability.

For each iteration, one fold was randomly selected as the validation set. Four TP sizes were evaluated: 50, 100, 150, and 200 individuals, corresponding to the use of one, two, three, or four folds from the remaining four folds, respectively. Training sets were constructed by randomly sampling the required number of folds from the non-validation folds without replacement. This procedure maintains a constant validation set size (~20% of the population) while varying training set size and ensures no overlap between training and validation. For each combination of TP size, replication level, environment, and trait, PA was evaluated over 100 independent iterations by repeating the random selection of validation fold and training folds.

Testing the joint effects of training-validation relatedness and replication number

To evaluate the effect of TP composition, population structure was described at multiple clustering resolutions using ancestry coefficients. The cross-entropy criterion identified K = 9 as the best-supported structure (Figure S1), and additional clustering resolutions were evaluated at K = 15, 20, 30, 40, 50, 60, 70, 80, 90, and 100. These clustering resolutions were used to generate leave-one-group-out cross-validation contrasts. For each K, one genetic group was

sequentially used as the VP while all remaining groups were used as the TP. Groups with fewer than two individuals were excluded from validation. This analysis is interpreted as a test of training-validation genomic relatedness and population-structure resolution rather than as a literal modification of the underlying population. Inter-group genomic relatedness was summarized from the genomic relationship matrix by calculating the mean genomic relationship coefficient among individuals belonging to different inferred groups at each K. Pairwise inter-group values were summarized to describe how genomic relatedness changed with clustering resolution. These levels of relatedness were evaluated jointly with different levels of replication number.

Comparison of genomic prediction models

In addition to rrBLUP, five alternative GP models were compared: GBLUP, BRR, BayesA, BayesB, and BayesC. rrBLUP assumes marker effects with a common Gaussian prior variance (Endelman, 2011). GBLUP models genetic covariance through the genomic relationship matrix (VanRaden, 2008). BRR is the Bayesian analog of ridge regression implemented in BGLR (Perez and de los Campos, 2014). BayesA allows marker-specific variances, BayesB uses a mixture prior with a proportion of markers assigned zero effect, and BayesC uses a mixture prior in which non-zero marker effects share a common variance (Meuwissen et al., 2001; Habier et al., 2011).

Bayesian models were run for 15,000 MCMC iterations with 3,000 burn-in iterations and thinning of 5. Model PA was estimated using repeated random 80:20 cross-validation with 100 repetitions per trait and environment, using all four replicates.

Statistical analysis of design-factor effects

The effects of replication number, TP size, training–validation relatedness, and their interactions were evaluated using the aligned rank transform framework (Wobbrock et al., 2011), because the assumptions of residual normality and homogeneity of variance were not consistently satisfied. Analyses were performed separately by trait and environment. For each analysis, aligned-rank ANOVA was used to test main effects and interactions, and explained-variance summaries were used to compare the relative contribution of each factor. Cross-validation repetitions were used to estimate the distribution of PA for each scenario, with scenario-level summaries treated as the primary inferential unit.

RESULTS

Replication increased estimates of genomic heritability

Marker-based h_g^2 increased with replication for all four traits in all three environments (Figure 1). Grain yield had lower genomic heritability than plant height, stem weight, and total soluble solids, consistent with the higher environmental sensitivity of yield (Falconer and Mackay, 2009). For grain yield, h_g^2 increased from 0.199 with one replicate to 0.370 with four replicates in environment 1, from 0.366 to 0.750 in environment 2, and from 0.157 to 0.516 in environment 3. Total soluble solids, plant height, and stem weight generally showed higher genomic heritability than grain yield, and also benefited from additional replication. These patterns indicate that replication improved the reliability of the phenotypic targets used for prediction.

Predictive ability improved with replication across traits and environments

PA increased with replication for all traits and environments (Figure 2). For total soluble solids, PA increased from 0.40 to 0.59 in environment 1, from 0.56 to 0.74 in environment 2, and from 0.52 to 0.63 in environment 3 when one and four replicates were compared. For plant height, PA increased from 0.59 to 0.71 in environment 1, from 0.40 to 0.59 in environment 2, and from 0.31 to 0.48 in environment 3. For stem weight, PA increased from 0.42 to 0.58 in environment 1, from 0.44 to 0.61 in environment 2, and from 0.37 to 0.58 in environment 3. Grain yield showed the largest relative response: PA increased from 0.22 to 0.41 in environment 1, from 0.27 to 0.53 in environment 2, and from 0.16 to 0.38 in environment 3. These results demonstrate how replication can have the greatest effect on the trait with more phenotyping error.

The benefit of increasing training population size was modest relative to increasing replication

Increasing TP size improved PA, but the benefit increased with greater replication (Figure 3). For grain yield, PA increased from 0.16 with 50 training individuals and one replicate to 0.41 with 200 training individuals and four replicates in environment 1. In environment 2, grain yield PA increased from 0.11 to 0.54 under the same contrast. In environment 3, it increased from 0.10 to 0.36 with 200 training individuals and three replicates. The aligned-rank analysis indicated that replication number was more important than TP size for grain yield PA across environments. Replication explained 88.0%, 61.7%, and 53.4% of the variation in environments 1, 2, and 3, respectively, while TP size explained 11.7%, 36.8%, and 39.1% (Table S2). Similar trends were observed in all other traits. Across traits, the response to TP size tended to plateau beyond approximately 150 individuals. The largest increases in PA were typically observed when

increasing replication from one to two or three replicates, with only marginal improvements from adding a fourth replicate.

The effect of increasing training-validation relatedness was modest compared to increasing replication

Mean inter-group genomic relatedness increased modestly as the number of inferred genetic groups increased from K = 9 to K = 100, with mean inter-group values ranging from approximately 0.35 to 0.41 (Figure S2). Leave-one-group-out cross-validation showed that PA generally increased when validation groups were more related to the TP and when phenotypes were estimated with more replicates (Figure 4). Although the increase was modest in terms of relatedness.

For grain yield, PA increased from 0.127 (K=9, 1 replicate) to 0.445 (K=90, 4 replicates) in environment 1; from 0.118 to 0.415 (K=40, 4 replicates) in environment 2; and from 0.044 to 0.238 (K=15, 4 replicates) in environment 3. Replication number explained more variation than TP composition in environments 1 (66% vs. 30%) and 2 (65% vs. 23%), but TP composition was more important in environment 3 (53% vs. 39%) (Table S3). For plant height, PA increased from 0.271 to 0.606 (K=70, 4 replicates) in environment 1; 0.226 to 0.515 (K=80, 4 replicates) in environment 2; and 0.208 to 0.474 (K=80, 4 replicates) in environment 3. Replication was more important in environments 2 (53% vs. 41%) and 3 (64% vs. 30%), with similar contributions in environment 1. For total soluble solids, PA increased from 0.153 to 0.500 (K=20, 3 replicates) in environment 1; 0.356 to 0.703 (K=60, 4 replicates) in environment 2; and 0.373 to 0.590 (K=40, 3 replicates) in environment 3. Replication was more important in environments 1 (58% vs. 38%) and 2 (63% vs. 34%), with similar contributions in environment 3. For stem weight, PA

increased from 0.087 to 0.579 (K=60, 4 replicates) in environment 1; 0.264 to 0.598 (K=60, 4 replicates) in environment 2; and 0.239 to 0.526 (K=60, 4 replicates) in environment 3. Replication was more important in environments 2 (56% vs. 39%) and 3 (65% vs. 29%), with similar contributions in environment 1 (Table S3).

The choice of model has little impact on predictive ability

GP models showed broadly similar PAs across traits and environments (Figure 5; Table S4). For most trait × environment combinations, no significant differences were detected among models (Tukey's HSD, $p < 0.05$). PAs generally ranged from 0.36 to 0.75, depending on the trait and environment. Bayesian models, BRR, and rrBLUP consistently exhibited comparable performance. GBLUP performed similarly in most cases, although it showed significantly lower PA for stem weight and total soluble solids in Environment 1. In Environments 2 and 3, all models showed equivalent PAs. These results indicate that model choice had a limited impact on genomic PA for the traits evaluated in this study.

DISCUSSION

This study used empirical field data from a sweet sorghum breeding population to evaluate how major design factors influence GP. The central result is that replication number was the strongest and most consistent determinant of PA across traits and environments. This finding is especially important because many GP studies emphasize marker density, model choice, TP size, and relatedness, whereas field-trial precision is sometimes treated as a secondary issue. In this dataset, the reliability of the phenotypic measurement strongly constrained PA, particularly for grain yield.

The effect of replication was reflected first in genomic heritability. For all traits, marker-based genomic heritability increased as more replicates were included. This pattern is expected because replication reduces plot-level error and produces more reliable genotype means (Bos, 1983; Wricke and Weber, 1986; Lorenz, 2013; Lourenço et al., 2020; Yan, 2021). The effect was strongest for grain yield, which is often among the most environmentally sensitive traits in cereal breeding. When genotype means are noisy, GP models may estimate marker effects against residual environmental variation rather than genetic signals (Kruijer et al., 2015). Increasing replication, therefore, improves the phenotype measurement and allows marker information to be used more effectively.

TP size also improved PA, consistent with theory and empirical studies in other crops (Lorenzana and Bernardo, 2009; Heffner et al., 2011; Lorenz, 2013; Rincent et al., 2017; Norman et al., 2018; Lozada et al., 2019; Zhu et al., 2021). However, the benefit of larger TPs depended on replication. Increasing the TP from 50 to 200 individuals had the largest effect when phenotype estimates were based on multiple replicates. This result has practical implications for breeding programs with limited resources: genotyping more individuals may not produce the expected benefit if the phenotypes used for training are too imprecise. Conversely, increasing replication without adequate training size may also limit prediction. The most efficient strategy is therefore a balanced allocation that provides enough individuals for marker-effect estimation and enough replication for reliable phenotypic targets.

The training-validation relatedness analysis confirmed the importance of population composition. PA increased when validation groups were more connected to the training set, in agreement with studies showing that relatedness between training and VPs strongly affects GP (Lorenz et al., 2011; Edwards et al., 2019; Werner et al., 2020; Fraslin et al., 2022).

However, this result should be interpreted with caution because relatedness was not manipulated independently. Changes in clustering resolution simultaneously altered validation-group size, group composition, and the degree of relatedness between training and validation sets. Consequently, the observed effects cannot be attributed solely to relatedness. Rather, they reflect the combined influence of relatedness and population partitioning.

The comparison among different models showed that model choice had a relatively minor effect compared with phenotyping and training design. This agrees with studies reporting that many GP models perform similarly for complex traits when marker density is high and TP size is modest (Merrick and Carter, 2021; Meher et al., 2022; Liu et al., 2025). Nonetheless, no model was consistently superior across all settings. For resource-limited breeding programs, these results suggest that investments in experimental design and phenotype quality may yield larger gains than switching among commonly used prediction models.

The study also highlights the value of empirical evaluation. Simulation studies are useful because they allow controlled manipulation of heritability, population size, genetic architecture, and relatedness (Zhong et al., 2009; Lorenz, 2013; Atefi et al., 2018). However, empirical field data include heterogeneous environments, incomplete precision, and operational constraints that define the real performance of GS in breeding programs. Results derived from simulations should be validated under field conditions, particularly when designing GS systems for small or emerging breeding programs.

A limitation of the present work is that PA was measured as the correlation between predicted genomic values and observed BLUEs or one-replicate phenotypes. This metric is appropriate for comparing scenarios, but it is not equivalent to the correlation with true genetic value. Because

replication improves the precision of the validation phenotype, part of the observed gain reflects a better measurement of the target trait. This is not a weakness from a breeding perspective, because breeders select based on estimated genotype performance, but it should be recognized when interpreting the magnitude of the reported effects. A second limitation is that the economic cost of replication, genotyping, and field-plot expansion was not explicitly modeled. Future work should combine PA with cost functions to identify optimum designs for specific breeding-program budgets.

CONCLUSION

In this empirical sweet sorghum breeding population, phenotyping replication was the most consistent driver of genomic PA across traits and environments. Additional replication increased genomic heritability and improved PA, with the largest gains observed for grain yield. TP size and training-validation relatedness also improved prediction, but their benefits were greatest when phenotype estimates were based on adequate replication. By contrast, commonly used GP models differed only modestly across most traits and environments. These findings indicate that resource-limited sorghum breeding programs should prioritize field-trial precision through adequate replication before expanding TP size, optimizing genomic relatedness, or exploring more complex GP models, although all three factors contributed to PA. Optimizing these design factors can make GS more effective and more cost-efficient in practical breeding systems.

Data availability

The datasets and analysis scripts generated in this study will be made available in a public repository upon publication.

Funding

This study was supported by the Ministry of Agriculture, Natural Resources, and Rural Development (MARNDR) in Haiti through the Technological Innovation Program in Agriculture and Agroforestry (PITAG).

Acknowledgments

The authors thank the CHIBAS field, laboratory, and data-management teams for their contributions to trial implementation, phenotyping, sample preparation, and data analysis.

Conflicts of interest

The authors declare no conflicts of interest.

Author contributions

J.R.C. and G.P. conceived the study and coordinated the breeding population development and field evaluations.

J.R.C. performed the phenotypic and genomic analyses.

B.R., G.M., and G.P. contributed to the genomic prediction strategy, interpretation of results, and manuscript development.

B.R., G.P., and T.T. revised the draft and the final version of the manuscript.

J.R.C. and G.P. wrote the initial draft of the manuscript. All authors reviewed, revised, and approved the final manuscript.

References

Alemu, A., J. Åstrand, O.A. Montesinos-López, J. Isidro Y Sánchez, J. Fernández-Gónzalez, et al. 2024. Genomic selection in plant breeding: Key factors shaping two decades of progress. Mol. Plant 17(4): 552–578. doi: 10.1016/j.molp.2024.03.007.

Atefi, A., A.A. Shadparvar, and N. Ghavi Hossein-Zadeh. 2018. Accuracy of Genomic Prediction under Different Genetic Architectures and Estimation Methods. Iran. J. Appl. Anim. Sci. 8(1): 43–52.

Bos, I. 1983. The optimum number of replications when testing lines or families on a fixed number of plots. Euphytica 32(2): 311–318. doi: 10.1007/BF00021439.

Cericola, F., A. Jahoor, J. Orabi, J.R. Andersen, L.L. Janss, et al. 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 (D.D. Fang, editor). PLOS ONE 12(1): e0169606. doi: 10.1371/journal.pone.0169606.

Charles, J.R., M.D. Dorval, J.B. Durone, L.F.V. Ferrão, R.R. Amadeu, et al. 2024. Genomic prediction of sweet sorghum agronomic performance under drought and irrigated environments in Haiti. Crop Sci. 64(3): 1595–1607. doi: 10.1002/csc2.21228.

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

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, et al. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species (L. Orban, editor). 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 J. 4(3): 250. doi: 10.3835/plantgenome2011.08.0024.

Endelman, J.B., G.N. Atlin, Y. Beyene, K. Semagn, X. Zhang, et al. 2014. Optimal Design of Preliminary Yield Trials with Genome-Wide Markers. Crop Sci. 54(1): 48–59. doi: 10.2135/cropsci2013.03.0154.

Falconer, D.S., and T. Mackay. 2009. Introduction to quantitative genetics. 4. ed., [16. print.]. Pearson, Prentice Hall, Harlow.

Fraslin, C., J.M. Yáñez, D. Robledo, and R.D. Houston. 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.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12(1): 186. doi:

10.1186/1471-2105-12-186.

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011. Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. Plant Genome 4(1): 65. doi: 10.3835/plantgenome2010.12.0029.

Kruijer, W., M.P. Boer, M. Malosetti, P.J. Flood, B. Engel, et al. 2015. Marker-Based Estimation of Heritability in Immortal Populations. Genetics 199(2): 379–398. doi: 10.1534/genetics.114.167916.

Liu, B., H. Liu, J. Tu, J. Xiao, J. Yang, et al. 2025. An investigation of machine learning methods applied to genomic prediction in yellow-feathered broilers. Poult. Sci. 104(1): 104489. doi: 10.1016/j.psj.2024.104489.

Lorenz, A.J. 2013. Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment. G3 GenesGenomesGenetics 3(3): 481–491. doi: 10.1534/g3.112.004911.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, et al. 2011. Genomic Selection in Plant Breeding. Advances in Agronomy. Elsevier. p. 77–123

Lorenz, A.J., and K.P. Smith. 2015. Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. Crop Sci. 55(6): 2657. doi: 10.2135/cropsci2014.12.0827.

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120(1): 151–161. doi: 10.1007/s00122-009-1166-3.

Lourenço, V.M., J.O. Ogutu, and H.-P. Piepho. 2020. Robust estimation of heritability and predictive accuracy in plant breeding: evaluation using simulation and empirical data. BMC Genomics 21(1): 43. doi: 10.1186/s12864-019-6429-z.

Lozada, D.N., R.E. Mason, J.M. Sarinelli, and G. Brown-Guedira. 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.

Meher, P.K., S. Rustgi, and A. Kumar. 2022. Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. Heredity 128(6): 519–530. doi: 10.1038/s41437-022-00539-9.

Merrick, L.F., and A.H. Carter. 2021. Comparison of genomic selection models for exploring predictive ability of complex traits in breeding programs. Plant Genome 14(3): e20158. doi: 10.1002/tpg2.20158.

Meuwissen, B.J., M.E. Goddard, and others. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819–1829.

Muleta, K.T., G. Pressoir, and G.P. Morris. 2019. Optimizing Genomic Selection for a Sorghum Breeding Program in Haiti: A Simulation Study. G3 GenesGenomesGenetics 9(2): 391–401. doi: 10.1534/g3.118.200932.

Norman, A., J. Taylor, J. Edwards, and H. Kuchel. 2018. Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. G3 GenesGenomesGenetics 8(9): 2889–2899. doi: 10.1534/g3.118.200311.

Perez, P., and G. de los Campos. 2014. Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198(2): 483–495. doi: 10.1534/genetics.114.164442.

Rincent, R., A. Charcosset, and L. Moreau. 2017. Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. Theor. Appl. Genet. 130(11): 2231–2247. doi: 10.1007/s00122-017-2956-7.

Satam, H., K. Joshi, U. Mangrolia, S. Waghoo, G. Zaidi, et al. 2023. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology 12(7): 997. doi: 10.3390/biology12070997.

Tan, B., D. Grattapaglia, G.S. Martins, K.Z. Ferreira, B. Sundberg, et al. 2017. Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. BMC Plant Biol. 17(1): 110. doi: 10.1186/s12870-017-1059-6.

Tayeh, N., A. Klein, M.-C. Le Paslier, F. Jacquin, H. Houtin, et al. 2015. Genomic Prediction in Pea: Effect of Marker Density and Training Population Size and Composition on Prediction Accuracy. Front. Plant Sci. 6. doi: 10.3389/fpls.2015.00941.

Technow, F., A. Bürger, and A.E. Melchinger. 2013. Genomic Prediction of Northern Corn Leaf Blight Resistance in Maize with Combined or Separated Training Sets for Heterotic Groups. G3amp58 GenesGenomesGenetics 3(2): 197–203. doi: 10.1534/g3.112.004630.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91(11): 4414–4423. doi: 10.3168/jds.2007-0980.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, et al. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92(1): 16–24. doi: 10.3168/jds.2008-1514.

Werner, C.R., R.C. Gaynor, G. Gorjanc, J.M. Hickey, T. Kox, 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.

Wetterstrand, K.A. 2020. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Genome.gov. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (accessed 30 May 2021).

Wobbrock, J.O., L. Findlater, D. Gergle, and J.J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, Vancouver BC Canada. p. 143–146

Wricke, G., and E. Weber. 1986. Quantitative genetics and selection in plant breeding. W. de Gruyter, Berlin ; New York.

Yan, W. 2021. Estimation of the Optimal Number of Replicates in Crop Variety Trials. Front. Plant Sci. 11: 590762. doi: 10.3389/fpls.2020.590762.

Zhang, H., L. Yin, M. Wang, X. Yuan, and X. Liu. 2019. 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.

Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J.-L. Jannink. 2009. Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics 182(1): 355–364. doi: 10.1534/genetics.108.098277.

Zhu, X., W.L. Leiser, V. Hahn, and T. Würschum. 2021. Training set design in genomic prediction with multiple biparental families. Plant Genome 14(3): e20124. doi: 10.1002/tpg2.20124.

Figure captions and figures

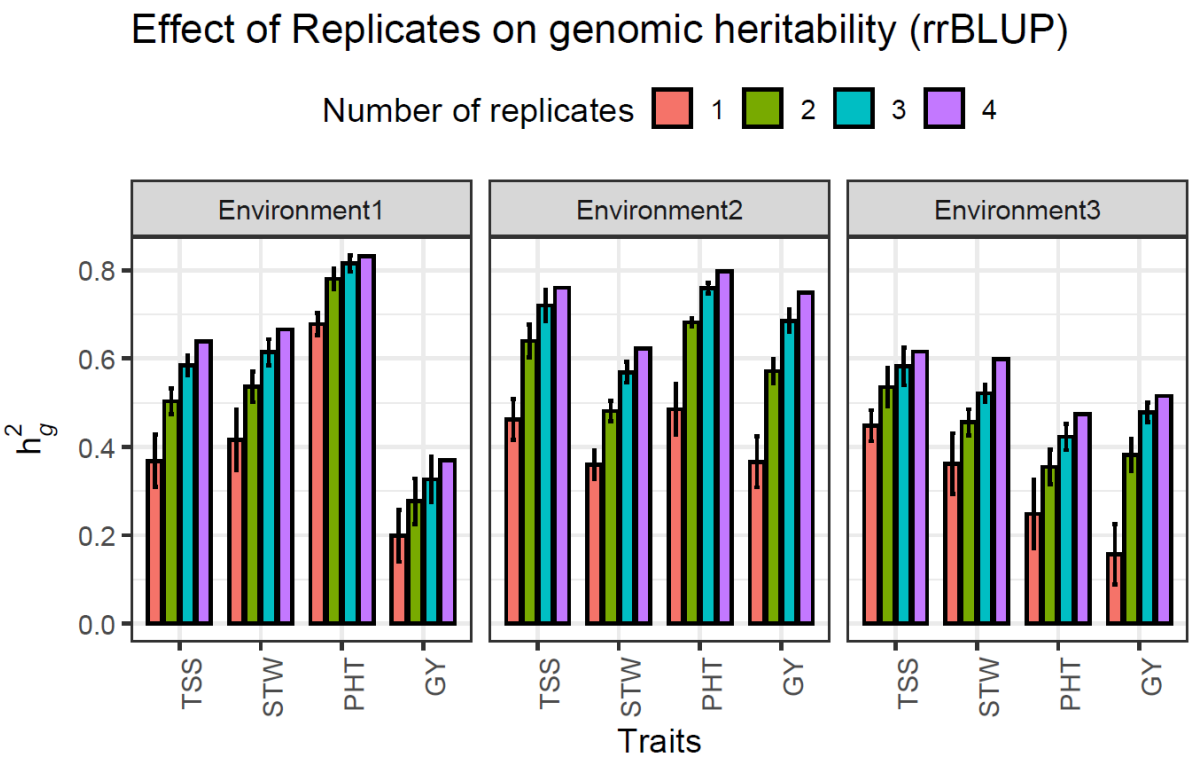

Figure 1. Effect of replication number on marker-based genomic heritability across traits and environments. Bars show genomic heritability estimates for grain yield (GY), plant height (PHT), stem weight (STW), and total soluble solids (TSS) in each environment at one, two, three, and four replicates. Error bars represent standard errors.

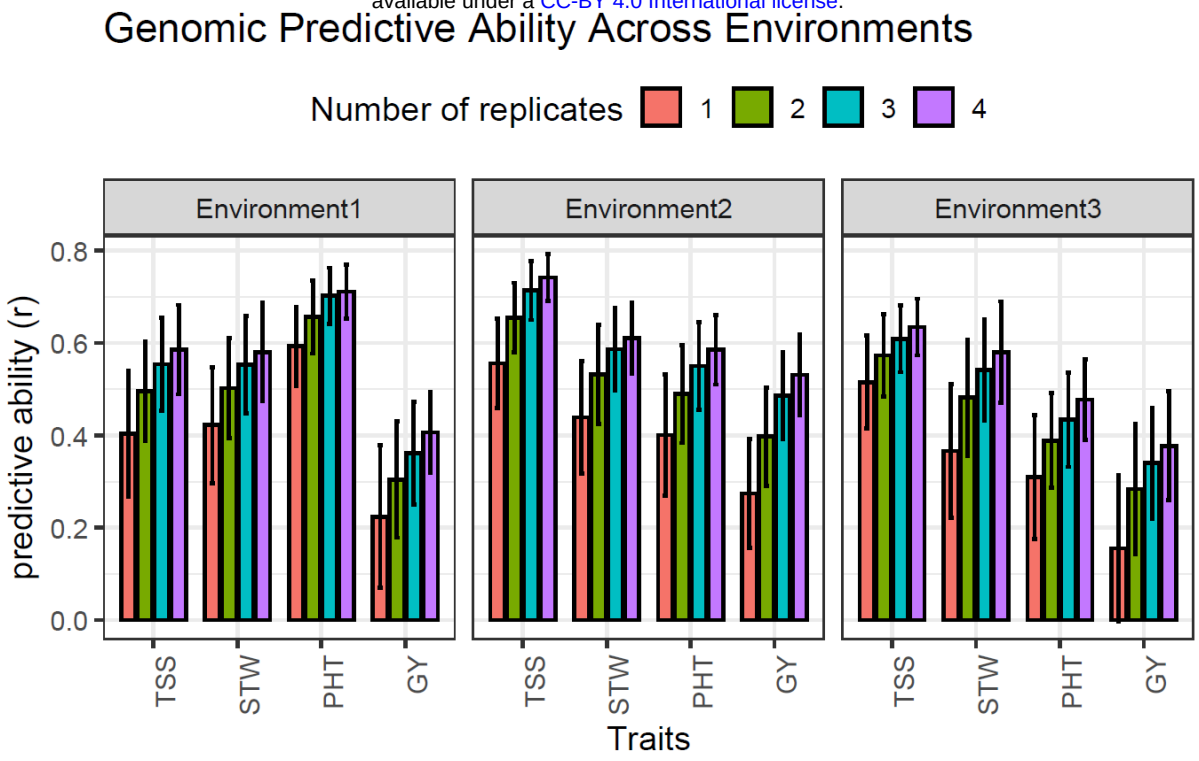

Figure 2. Effect of replication number on genomic predictive ability across traits and environments. Predictive ability is the Pearson correlation between observed genotype BLUEs or one-replicate phenotypes and predicted genomic values in validation sets. Bars show mean predictive ability across cross-validation repetitions for grain yield (GY), plant height (PHT), stem weight (STW), and total soluble solids (TSS); error bars show variation among repetitions.

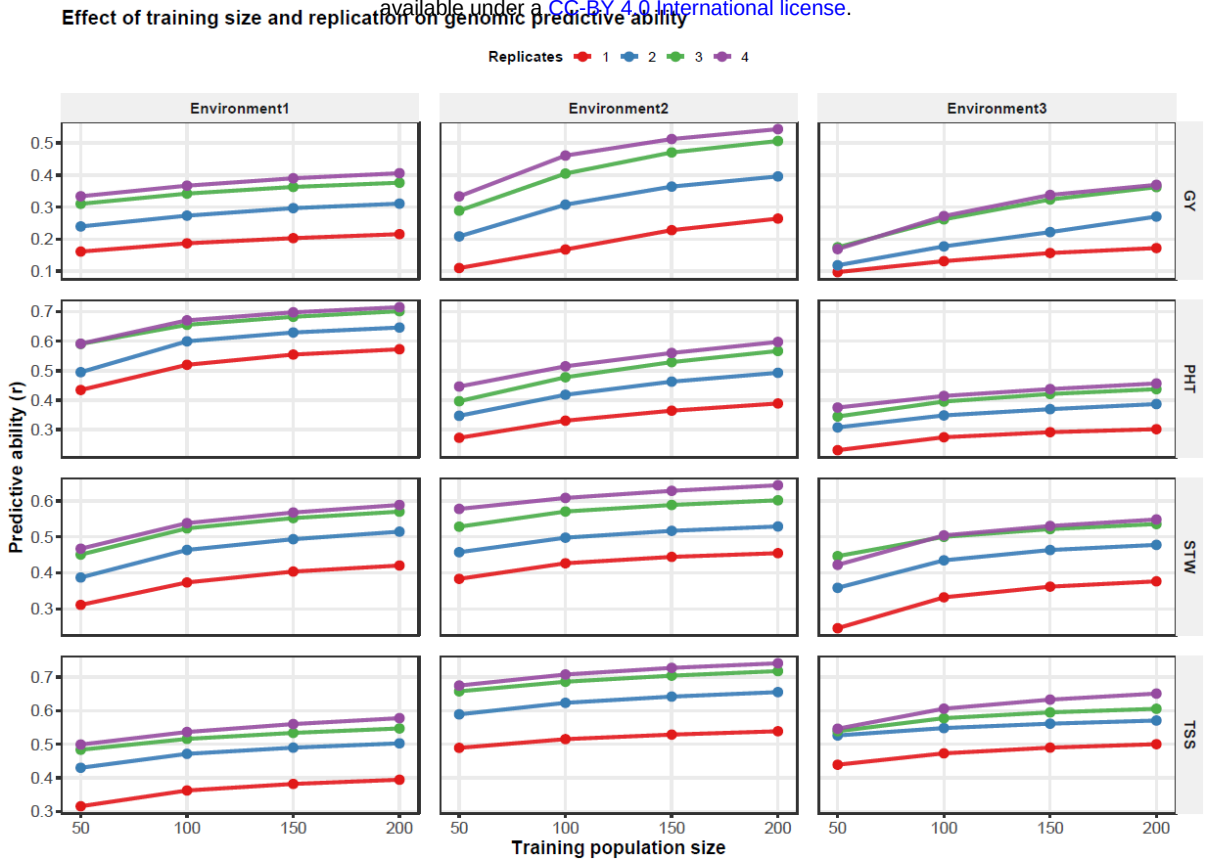

Figure 3. Combined effect of training population (TP) size and replication number on genomic predictive ability. Predictive ability is the Pearson correlation between observed genotype BLUEs or one-replicate phenotypes and predicted genomic values in validation sets. Lines show mean predictive ability as a function of TP size for each replication level across grain yield (GY), plant height (PHT), stem weight (STW), and total soluble solids (TSS) in different environments. The points represent the mean predictive ability obtained at each TP size.

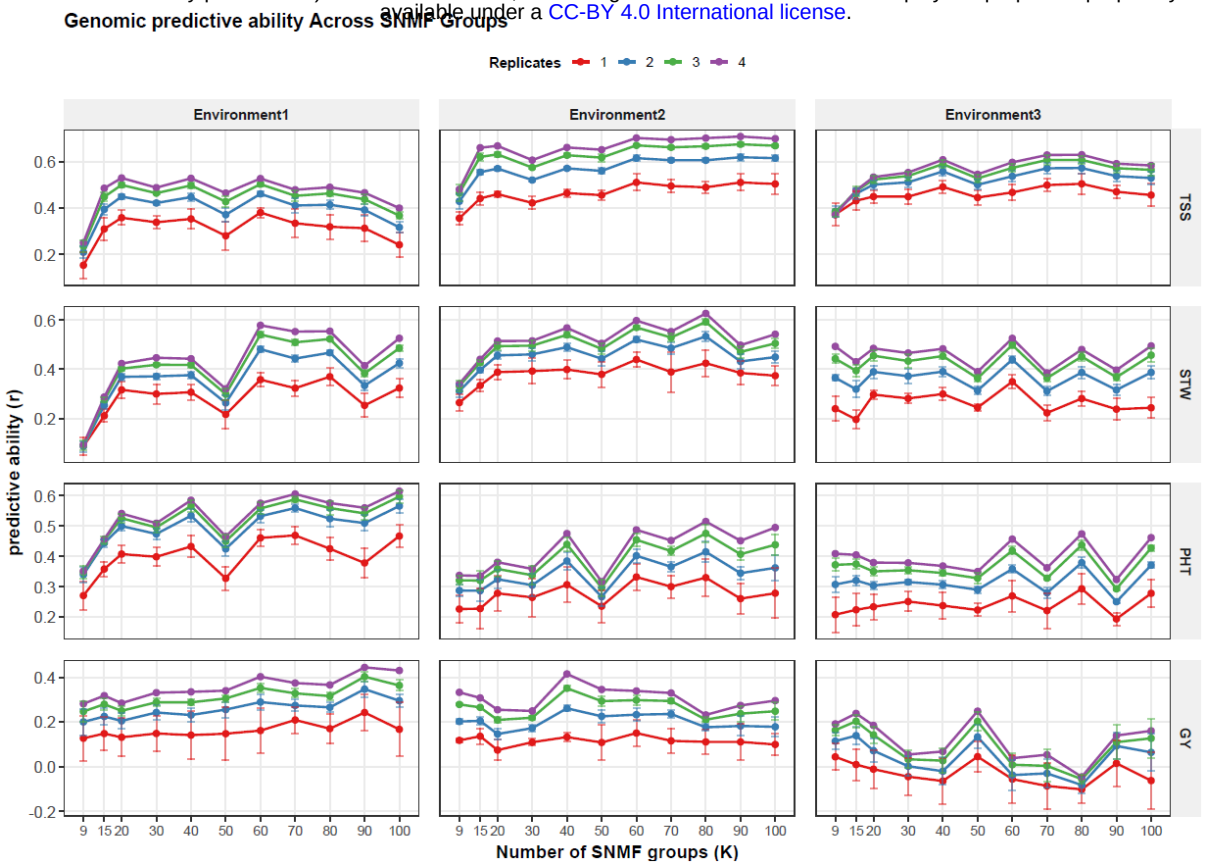

Figure 4. Combined effect of training-validation genomic relatedness and replication number on genomic predictive ability. Predictive ability is the Pearson correlation between observed genotype BLUEs or one-replicate phenotypes and predicted genomic values in validation sets. Points show predictive ability across sNMF clustering resolutions. Higher K values represent finer population-structure partitions. Error bars show variation among repetitions.

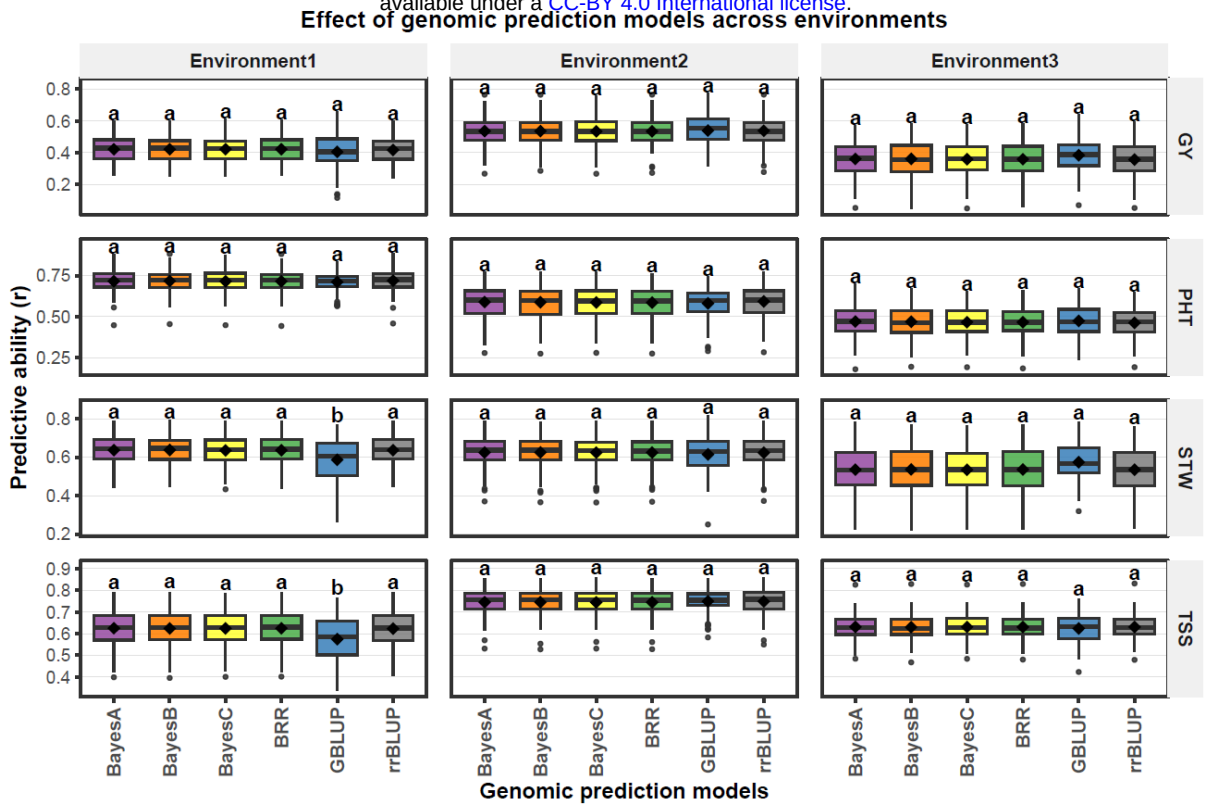

Figure 5. Genomic predictive ability (*r*) of six GP models across three environments for grain yield, plant height, stem weight, and total soluble solids (TSS). Boxplots show the distribution of predictive ability across cross-validation runs, and red diamonds indicate mean values. Different letters indicate significant differences among GP models within each trait × environment combination based on ANOVA followed by HSD mean grouping ($P < 0.05$).