frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 23 July 2024
DOI 10.3389/fpls.2024.1400000

OPEN ACCESS

EDITED BY
Leif Skot,
Aberystwyth University, United Kingdom

REVIEWED BY
Zitong Li,
Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia
João Ricardo Bachega Feijó Rosa,
RB Genetics & Statistics Consulting, Brazil

*CORRESPONDENCE
Fernando Silva Aguilar
fsilvaag@cenicana.org
Diego Jarquín
jhernandezjarqui@ufl.edu

RECEIVED 12 March 2024
ACCEPTED 13 June 2024
PUBLISHED 23 July 2024

CITATION
Garcia-Abadillo J, Adunola P, Aguilar FS, Trujillo-Montenegro JH, Riascos JJ, Persa R, Isidro y Sanchez J and Jarquín D (2024) Sparse testing designs for optimizing predictive ability in sugarcane populations. *Front. Plant Sci.* 15:1400000. doi: 10.3389/fpls.2024.1400000

Sparse testing designs for optimizing predictive ability in sugarcane populations

Julian Garcia-Abadillo[1,2], Paul Adunola[3], Fernando Silva Aguilar[4]*, Jhon Henry Trujillo-Montenegro[4], John Jaime Riascos[4], Reyna Persa[2], Julio Isidro y Sanchez[1] and Diego Jarquín[2]*

[1]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, Madrid, Spain, [2]Agronomy Department, University of Florida, Gainesville, FL, United States, [3]Horticultural Sciences Department, University of Florida, Gainesville, FL, United States, [4]Colombian Sugarcane Research Center, Cenicaña, Cali, Valle del Cauca, Colombia

Sugarcane is a crucial crop for sugar and bioenergy production. Saccharose content and total weight are the two main key commercial traits that compose sugarcane's yield. These traits are under complex genetic control and their response patterns are influenced by the genotype-by-environment (G×E) interaction. An efficient breeding of sugarcane demands an accurate assessment of the genotype stability through multi-environment trials (METs), where genotypes are tested/evaluated across different environments. However, phenotyping all genotype-in-environment combinations is often impractical due to cost and limited availability of propagation-materials. This study introduces the sparse testing designs as a viable alternative, leveraging genomic information to predict unobserved combinations through genomic prediction models. This approach was applied to a dataset comprising 186 genotypes across six environments (6×186=1,116 phenotypes). Our study employed three predictive models, including environment, genotype, and genomic markers as main effects, as well as the G×E to predict saccharose accumulation (SA) and tons of cane per hectare (TCH). Calibration sets sizes varying between 72 (6.5%) to 186 (16.7%) of the total number of phenotypes were composed to predict the remaining 930 (83.3%). Additionally, we explored the optimal number of common genotypes across environments for G×E pattern prediction. Results demonstrate that maximum accuracy for SA (ρ=0.611) and for TCH (ρ=0.341) was achieved using in training sets few (3) to no common (0) genotype across environments maximizing the number of different genotypes that were tested only once. Significantly, we show that reducing phenotypic records for model calibration has minimal impact on predictive ability, with sets of 12 non-overlapped genotypes per environment (72=12×6) being the most convenient cost-benefit combination.

KEYWORDS

genomic selection GS, genomic prediction GP, sparse testing designs, optimization, sugarcane breeding

1 Introduction

The world faces an urgent challenge to provide nutritional sustenance for a burgeoning 8 billion people (UNPF, 2023) amidst threats to food security, climate change, and finite resources (FAO, 2015; UNCCD, 2017, 2018). Bridging the gap between food production and population growth necessitates innovative agricultural solutions. Sugarcane emerges as a key player in this scenario, serving dual purpose: *i*) as a primary sugar source (a dietary staple), and *ii*) bioenergy feedstock production (Goldemberg, 2008; Waclawovsky et al., 2010). The versatility of sugarcane makes it an ideal crop for targeted breeding efforts to enhance both yield (sugar and biofuel) (Hoang et al., 2015; Mahadevaiah et al., 2021) and yield stability (Scortecci et al., 2012). However, sugarcane breeding is challenged by long breeding cycles, low genetic diversity, large genome size and clonal propagation, limiting efficient genetic improvement and yield potential (Souza et al., 2011; Hoang et al., 2015; Yadav et al., 2020; Mahadevaiah et al, 2021).

In plant breeding, identifying superior cultivars for agricultural demands remain critical. Sugarcane's narrow genetic base, primarily due to vegetative propagation, complicates breeding efforts (Roach, 1989; Raboin et al., 2008; Wei and Jackson, 2016). Multi-environmental trials (METs) are essential in sugarcane breeding, enabling the evaluation of genotype performance across environments and identifying stable or specifically adapted cultivars (Jackson and Hogarth, 1992; Abu-Ellail et al., 2020). However, METs are resource-intensive, particularly for clonally propagated crops, prompting the need for more efficient trial designs.

Genome-wide molecular prediction (GWP), genomic selection (GS) or simply genomic prediction (GP) has been transformative in plant breeding, enhancing genetic gain rates over the past two decades (Crossa et al., 2010, 2011, Jarquín et al., 2014; Ferrão et al., 2021; Voss-Fels et al., 2021; Mahadevaiah et al., 2021). GP uses genome-wide molecular markers to estimate breeding values of untested individuals (Meuwissen et al., 2001). By capturing the genetic variation in the genome, GP enables the prediction of complex traits that are otherwise challenging to assess using conventional breeding tools (Yadav et al., 2020; Ferrão et al., 2021). This approach facilitates early-stage breeding cycle decisions, reducing the need for extensive field trials and expediting cultivar development. In the context of sugarcane breeding, GP has been used to predict several traits of interest with moderate accuracy ranging from 0.2 to 0.45 (Deomano et al., 2020; Mahadevaiah et al., 2021; Islam et al., 2022; Islam et al., 2023), including both molecular and pedigree information (Deomano et al., 2020; Hayes et al., 2021) and modeling non-additive effects (Yadav et al., 2021). GP is particularly advantageous in predicting performance of individuals in new and partially tested environments through cross-validation (CV) schemes. This practice ensures that the predictive ability of GP models extends beyond the environments in which genotypes are bred, allowing accurate selection of individuals with potential high performance across a range of environmental and management conditions. Empirical studies suggest that incorporating G×E in GP models can streamline the breeding process, either by skipping certain stages or by reducing the number of genotypes tested in fields, thereby enhancing trial testing capacity (Crossa et al., 2017; Resende et al., 2018; Jarquín et al., 2020).

Sparse testing, the strategic selection of a subset of genotypes for evaluation in target environments, offers an innovative alternative to field trials for the entire population This approach leverages statistical techniques and predictive models, such as GP, to extrapolate the performance of the entire population from a selectively observed subset. Through model calibration and CV, sparse testing enables prediction of unobserved genotype-in-environment combinations. Consequently, it optimizes resource allocation, allowing breeders to focus on the most promising candidates from the target population of environments (TPEs). This methodology not only accelerates the breeding cycle but also the cost-effective identification of superior cultivars.

Sparse testing designs often mirror CV schemes CV1 and CV2 (Burgueño et al., 2012; Jarquín et al., 2017), encompassing a broad range of genotype-in-environment combinations. In the CV1 scenario, genotypes never tested at any environment are predicted, while CV2 involves predicting already observed genotypes in incomplete METs. This raises critical questions about the optimal design of sparse testing in METs, such as the balance between testing a few genotypes across multiple environments versus many genotypes in fewer environments; and the trade-off between prediction accuracy and selection intensity. Empirical evidence from crops such as maize (Jarquín et al., 2020; Atanda et al., 2021; Montesinos-López et al., 2023), wheat (Crespo-Herrera et al., 2021; Atanda et al., 2022) and soybean (Persa et al., 2023) has provided insights into optimizing sparse testing designs. Research by Jarquín et al. (2020) and Crespo-Herrera et al. (2021) revealed that GP models incorporating G×E can maintain robust predictive ability even with reduced training sets. These studies indicate that the highest recovery of predictive ability in scenarios with either non-overlapping or completely overlapping genotypes in training set combinations mostly depends on the species.

In this context, this study investigates the predictive ability of different sparse-testing allocation designs in sugarcane breeding. We simulated the allocation of finite breeding resources to optimize the prediction of un-phenotyped genotypes in METs. For that, we varied the number of overlapping genotypes and training set sizes, similarly to the approaches of Jarquín et al. (2020) and Crespo-Herrera et al. (2021). We considered three allocation scenarios: *i*) Non-Overlapping Genotypes (NOG), with a single observation per genotype across environments, *ii*) Overlapping Genotypes (OG) where all training genotypes are phenotyped across all environments, resembling the CV1 scenario, and *iii*) a combination of the NOG and OG schemes. To assess the contribution of capturing G×E in METs, three models were fitted for each sparse testing case: *i*) main effect of environment and genotype (M1); *ii*) main effect of environment, genotype and genomic markers (M2); and *iii*) main effect of environment, genotype, genomic markers, and the interaction between genomic markers and environment (M3). Our findings demonstrate effective resource optimization strategies for METs in sugarcane breeding, potentially increasing testing capacity by fivefold within a fixed

target set of genotype-in-environment combinations or reducing total phenotyping cost between 83% to 93%.

2 Materials and methods

2.1 Plant material and phenotyping

The dataset analyzed in this study comprises 220 genotypes (Jaimes et al., 2024) from the diverse panel of Centro de Investigación de la Caña de Azúcar, Cenicaña (Colombian Sugarcane Research Center). This panel includes 98 genotypes representing the phenotypic diversity of Cenicaña's germplasm bank, 58 genotypes selected for their resistance or susceptibility to prevalent diseases and pests in Colombia, 33 genotypes encompassing introductions and commercial varieties, and 31 wild species from *Saccharum officinarum, S. spontaneum, S. barberi, S. sinensi*, and *Erianthus* spp. The genotypes were planted across three locations in the Valle del río Cauca, Colombia: Balsora (Mayaguez sugarcane mill) in 2016 (E1) and 2017 (E2), Taula Mejía (La Cabaña sugarcane mill) in 2018 (E3) and 2019 (E4), and Porvenir (Manuelita sugarcane mill) in 2020 (E5) and 2021 (E6). Genotypes were planted at each location under a randomized incomplete block design. In Balsora and Taula Mejia 3 replications were planted, while in Porvenir, due to field limitations, only 2 replications were considered. The experimental unit at each location comprised a plot of 5 rows, each 5 meters long and spaced 1.65 m apart. Commercial checks, genotypes S29, S64, and S177, were replicated multiple times within each replicate-block, resulting in a total of 717, 720, and 576 experimental units for Balsora, Taula Mejia, and Porvenir, respectively.

Phenotypic data for sucrose accumulation (SA) and tons of cane per hectare (TCH), were collected during harvest (13 months after planting). SA was measured by shredding 10 stalks per experimental unit to obtain the juice per sample (Larrahondo and Torres, 1989). Then the sucrose content was quantified from the extracted juice through a near infrared (NIR) spectrophotometer (Larrahondo and Torres, 1989). For TCH, the measurement was taken after weighing all stalks per experimental unit. Both SA and TCH measurements were conducted for both plant cane and first ratoon at all locations.

2.2 Genotyping and quality control

DNA was extracted from each of the 220 genotypes following Dellaporta et al. (1983) protocol. Sequencing was conducted using both Genotyping-by-sequencing (GBS) and Restriction-site Associated DNA sequencing (RADSeq) on a Hiseq2000 Illumina System at a depth of 105X and 38X for GBS and RADSeq, respectively. Quality control was performed using FastQC (Andrews, 2010) and Trimmomatic (Bolger et al., 2014). Subsequently, the cleaned data were aligned and mapped to the CC 01–1940 monoploid reference genome (Trujillo-Montenegro et al., 2021), utilizing Bowtie2–2.5 (Langmead and Salzberg, 2012) with default parameters. Variant calling was conducted with NGSEP 4.0.2 (Tello et al., 2019) by filtering out markers that failed to meet the following quality control criteria: a Minor Allele Frequency (MAF) below 3%, Phred score below 30, distance between markers below 5 bp, presence of more than 50% of missing data, and an average depth below 5X, as well as markers with more than one allelic version. Post quality control, a finalized count of 22,324 single nucleotide polymorphism (SNP) markers remained for analyses. The markers were coded based on the scaled allele content with values ranging between 0 and 2 in steps of 0.2 {0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, and 2}.

2.3 Phenotypic adjustment

The statistical model was independently applied to raw data at each environment aiming to obtain the Best Linear Unbiased Predictor (BLUP) values of the genotypes considering the three checks assigned to each block as fixed effects for micro-environmental control (Belamkar et al., 2018; Xavier et al., 2022) using the linear predictor described in Equation 1.

$$y_{ikl} = \mu + Check + L_i + r_k + b_{l(r)} + \varepsilon_{ikl}$$
$$\begin{pmatrix} L \\ r \\ b_{(r)} \\ \varepsilon \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I\sigma^2_L & 0 & 0 & 0 \\ 0 & I\sigma^2_r & 0 & 0 \\ 0 & 0 & I\sigma^2_{b_{(r)}} & 0 \\ 0 & 0 & 0 & I\sigma^2_\varepsilon \end{bmatrix}\right) \quad (1)$$

where y_{ikl} denotes the phenotypic record of the i^{th} genotype in the l^{th} incomplete block within the k^{th} replication, μ is the overall mean, *Check* corresponds to the fixed effects of the checks, L_i is the random effect of the i^{th} genotype, r_k is the random effect of the k^{th} replication, $b_{l(r)}$ is the random nested effect of the l^{th} incomplete block within the k^{th} replication and ε_{ikl} is the corresponding error term. These terms are considered as independently and identically distributed outcomes, following a normal distribution with a covariance structure based on the identity matrix I and scaled by their corresponding variance component σ^2.

2.4 Allocation strategies

From the initial 220 genotypes, this study focused only on 186 Colombian cultivars. Thus, a total of 1,116 genotype-in-environment combinations (or phenotypes) resulted from the combination of the 186 Colombian genotypes observed at the six different environments. In practical scenarios, budgetary constraints imply that only a subset of these potential phenotypes can be realistically observed in the field. Therefore, strategies are essential in the allocation of genotypes across various environments. The goal is to create a calibration set comprising phenotypic records that maximize the accuracy of prediction models. This is key for obtaining accurate estimates of the performance of genotypes that are not observed in a set of environments. In this study, allocation optimization was executed by classifying the genotypes into one of the two following categories:

Non-overlapping Genotypes (Genotypes seen only in one environment): In this category, genotypes are randomly assigned to specific environments in such a way that they are unique and

observed only once across environments. For the case of study, this implies that each of the six environments would host 31 unique genotypes (186/6 = 31), guaranteeing that all genotypes are assigned to one environment only with no overlapping across environments.

Overlapping Genotypes (Genotypes seen in all the environments): Some genotypes that exhibit consistent performance across various environments or that are of particular interest for research purposes can be allocated in all six environments. This overlapping ensures that these genotypes are evaluated under diverse conditions, assessing their adaptability and stability across environments.

Let A and B represent, respectively, the number of non-overlapping and overlapping genotypes, and T the number of environments or trials. For a given 'A/B' design, the number of required phenotypes is (A+B)×T, the number of unique genotypes tested is A×T+B and the average number of replicates/phenotypes per genotype is (A+B×T)/(A+B). Assuming a fixed budget for 186 plots and uniform plot costs in all the environments, two extreme allocation strategies arise.

- Testing all the genotypes only once (31/0): In this scenario, 31 unique genotypes are assigned and phenotyped at each environment (A=31). Hence, there is no overlapping of genotypes between environments (B=0). The number of phenotypes is (31+0)×6=186, the number of unique tested genotypes is 186 and the number of phenotypes per genotype is (31+0×6)/(31+0)=1 phenotype/genotype.

- Testing same genotypes in all environments (0/31): Here, a subset of 31 unique genotypes from the original set of 186 is phenotyped in all the environments with no specific or non-overlapping genotypes at each environment (A=0) and all the 31 being overlapped genotypes (B=31). The number of phenotypes is still (0+31)×6=186 but the number of unique genotypes is reduced to (0×6)+31=31, and the number of phenotypes per genotype is increased to (0+31×6)/(0+31)=6 phenotypes/genotype.

In addition, we can construct mixed designs by adding one overlapping genotype and removing one non-overlapping genotype from each environment. For instance, the design 16/15 is composed of 16 non-overlapping genotypes per environment and 15 overlapping genotypes phenotyped in the six environments. In this case, the number of phenotypes remains the same (16+15)×6=186 but the number of unique genotypes is 16×6+15=111 and the average phenotypes per genotype is (16+15×6)/(16+15)=3.42.

Starting from the full non-overlapping design (31/0), nine designs resulted after increasing the number of overlapping or common genotypes (B) by four with respect to the previous design, except for the first case where only three genotypes were added to B. These designs are 31/0, 28/3, 24/7, 20/11, 16/15, 12/19, 8/23, 4/27, and 0/31. The set of phenotypes or genotype-in-environment combinations observed in fields, (A+B)×T=186, constitutes the calibration set for training models to predict the phenotypes of the remaining unobserved genotype-in-environment combinations (1,116-186=930 combinations). A visual representation of the designs is shown in Figure 1.

To test the allocation of resources in terms of predictive ability, not only regarding the proportion of non-overlapping and overlapping genotypes, but also with respect to different budget constraints, additional sample sizes were considered for composing training (168, 144, 120, 96, and 72 phenotypes). All calibration set designs are shown in Table 1, where the columns in the middle represent different designs (compositions) and the rows indicate the different calibration sizes. Ten random partitions or replicates were considered to evaluate the predictive ability of the different designs (combination of set size and composition). To ensure a fair comparison among sparse designs, the size of the prediction set was fixed to 930 (83.3%) for all these. This entails that some genotype-in-environment combinations were excluded from consideration in rows 2 to 6, neither included in either the calibration or the prediction set.

2.5 Predictive models

Three different GP models, based on random effects and modeled via covariance structures, were calibrated and used to predict the testing sets. The first model M1 is considered the base model, and it assumes independent and identically distributed (IID) outcomes among its components (Equation 2). The second model, M2, leverages molecular marker information to determine relationships between pairs of genotypes (Equation 3). While M1 and M2 are main effects models, the third model M3 introduces the G×E term (Equation 4). The model's performance was assessed using the Pearson's correlation between the observed and the predicted values within each environment. The traits SA and TCH were analyzed separately.

M1: Environment and genotype main effects.

$$y_{ij} = \mu + E_j + L_i + \varepsilon_{ij}$$
$$\begin{pmatrix} E \\ L \\ \varepsilon \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Z_E Z_E' \sigma_E^2 & 0 & 0 \\ 0 & Z_L Z_L' \sigma_L^2 & 0 \\ 0 & 0 & I\sigma_\varepsilon^2 \end{bmatrix} \right) \quad (2)$$

where y_{ij} is the adjusted phenotypic observation of the target trait in the j^{th} environment (j=1, 2,…, 6) for the i^{th} genotype (i=1, 2, …, 186), μ is the overall mean, E_j is the random effect of the j^{th} environment, L_i is the random effect of the i^{th} genotype, and ε_{ij} is the corresponding error term. E_j, L_i, and ε_{ij} are assumed to be independent and identically distributed outcomes from a normal density. Z_E (n×6) and Z_L (n×186) are the design matrices connecting phenotypes with environments and genotypes, respectively, and I (n×n) is the identity matrix, with n being the number of total observations (n=186×6). The variance component for each term is denoted by σ^2. In this model, the genotypes are assumed to be independent, and therefore, no information can be borrowed from phenotyped individuals to their non-phenotyped relatives.

M2: Environment, genotype, and genomic markers main effects.

$$y_{ij} = \mu + E_j + L_i + g_i + \varepsilon_{ij}$$
$$\begin{pmatrix} E \\ L \\ g \\ \varepsilon \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Z_E Z_E' \sigma_E^2 & 0 & 0 & 0 \\ 0 & Z_L Z_L' \sigma_L^2 & 0 & 0 \\ 0 & 0 & Z_L G Z_L' \sigma_g^2 & 0 \\ 0 & 0 & 0 & I\sigma_\varepsilon^2 \end{bmatrix} \right) \quad (3)$$

where all the terms in common with M1 have the same meaning, g_i is the genomic effect based on maker SNPs for the i^{th} genotype, and K (186×186) is the kinship matrix describing

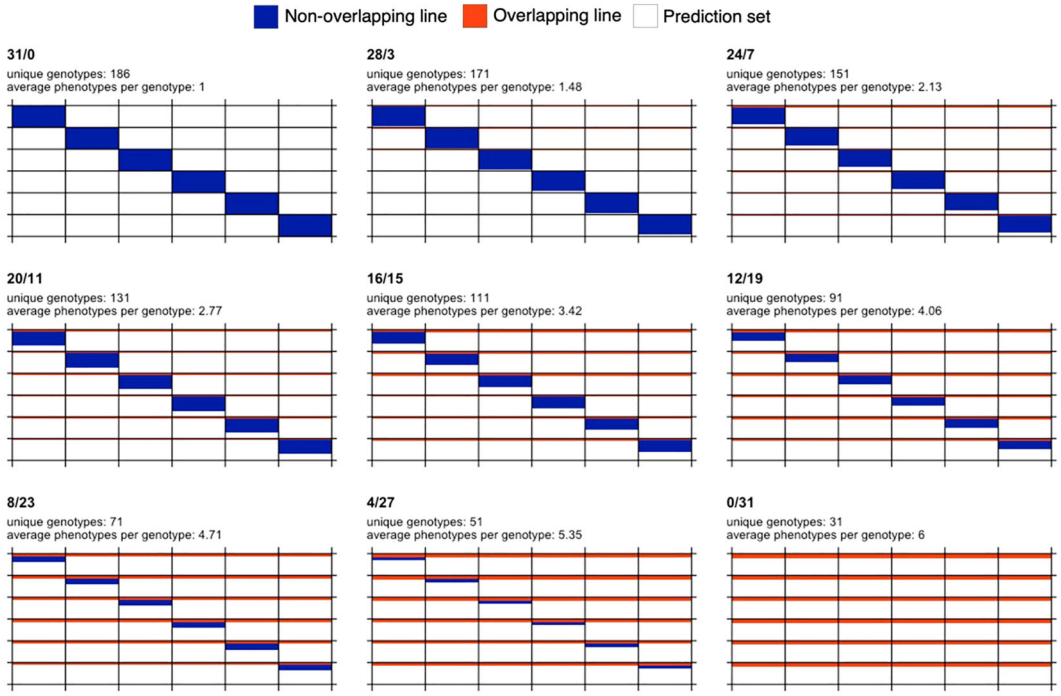

FIGURE 1
Graphical representation of sparse testing designs for a fixed budget and for a data set comprising 186 phenotypes and six environments. Each subplot represents a different design. The design name, number of unique genotypes and average phenotypes per genotype are shown at the top of each subplot. Rows and columns represent genotypes and environments, respectively. Blue color indicates that a genotype is observed in only one environment, while orange color indicates overlapping genotypes observed in all the six environments. White color is used to represent phenotypes not used in the calibration set and potentially forming the prediction set (155 phenotypes per environment or equivalently 930 phenotypes across environments).

genomic similarities between pairs of individuals (VanRaden, 2008). Both L_i and g_i terms capture information about the genotype but in the latter, genomic information is used to relate phenotyped and non-phenotyped individuals. The genotype effect L_i is not removed from the model to account for model miss-specification and imperfect information (genetic variability that cannot be explained through SNP markers only).

M3: Environment, genotype, and genomic markers main effects, and Genotype × Environment interaction.

$$y_{ij} = \mu + E_j + L_i + g_i + gE_{ij} + \varepsilon_{ij}$$
$$\begin{pmatrix} E \\ L \\ g \\ gE \\ \varepsilon \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Z_E Z_E' \sigo_E^2 & 0 & 0 & 0 & 0 \\ 0 & Z_L Z_L' \sigma_L^2 & 0 & 0 & 0 \\ 0 & 0 & Z_L G Z_L' \sigma_g^2 & 0 & 0 \\ 0 & 0 & 0 & Z_L Z_E \odot Z_L G Z_L' \sigma_{gE}^2 & 0 \\ 0 & 0 & 0 & 0 & I\sigma_\varepsilon^2 \end{bmatrix}\right) \quad (4)$$

where all the terms in common with M2 remain the same, and the added term gE_{ij} represents the effect of the interaction between the j^{th} environment and i^{th} genotype (Jarquín et al., 2014), ⊙ is the element-wise multiplication or Hadamard product operation between two matrices.

TABLE 1 Resource allocation designs.

# Calibration	Designs									# Prediction	# Unused
186 (16.7%)	31/0	28/3	24/7	20/11	16/15	12/19	8/23	4/27	0/31	930 (83.3%)	0 (0.0%)
168 (15.1%)		28/0	24/4	20/8	16/12	12/16	8/20	4/24	0/28	930 (83.3%)	18 (10%)
144 (12.9%)			24/0	20/4	16/8	12/12	8/16	4/20	0/24	930 (83.3%)	42 (23%)
120 (10.8%)				20/0	16/4	12/8	8/12	4/16	0/20	930 (83.3%)	66 (35%)
96 (8.6%)					16/0	12/4	8/8	4/12	0/16	930 (83.3%)	90 (48%)
72 (6.5%)						12/0	8/4	4/8	0/12	930 (83.3%)	114 (61%)

The first column indicates the size of the calibration set across environments (relative size with respect to the total potential number of combinations). The last two columns show, the size of the prediction set (constant) and the number (and proportion) of phenotypes discarded from the calibration set. All the designs in the same row share a common calibration set size while the designs in the same column share a common number of non-overlapping genotypes.

2.6 Cross validation and performance metrics

In the context of plant breeding, particularly in the GP domain, the evaluation of models often involves the use of CV schemes (Jarquin et al., 2021). One widely used method is the K-fold CV, which helps to assess the performance of predictive models. K-fold CV involves splitting the dataset into K subsets or folds. The process then iterates K times, each time using one of the K subsets as the prediction set while the remaining K-1 subsets are used for calibrating the predictive models. This approach provides a robust way to assess the model's performance. It returns predicted values for all the datapoints in a fold without any of them being present in the calibration set.

The process of partitioning the dataset into folds is not trivial and might include some constraints. In this study, we included features of two common CV schemes: CV2 and CV1. In the basic CV scheme, known as CV2, the dataset is randomly partitioned without any specific constraints. This means that the model may predict the phenotype of genotypes that have been observed in other environments. While this approach is informative, it doesn't guarantee that the model will predict unseen genotypes, as some genotypes may appear in both the calibration and the prediction sets due to random partitioning. In CV1, a specific constraint is applied: all phenotypes of the same genotype must be placed in the same fold. This ensures that when evaluating the model, none of the observations of a genotype in any environment are present in the calibration set. This constraint guarantees that the models will predict an unseen genotype, as all observations of that genotype are in the prediction test.

The allocation designs constitute the spectrum between CV2 and CV1, with 31/0 being the most extreme case that could occur on a CV2 assignment, and 0/31 being the CV1 scheme. The higher the number of overlapping genotypes, the closer to the CV1 scheme. All the designs, except those with 0 non-overlapping genotypes, are CV2-like schemes. In the context of predictive modeling, the CV1 scheme poses a greater challenge, especially for models that do not incorporate genomic information and treat genotypes as independent entities. This challenge arises because the calibration set does not provide any useful insights during the model training process for predicting a specific, new, and unseen genotype.

To evaluate the performance of predictive models in this study, the Pearson's correlation and the Mean Squared Error MSE were used to measure the association between the adjusted phenotypes and the predicted values. These were computed for each environment or trial, reflecting how well the models performed when predicting the genotypes in specific environmental conditions. Averaged correlations and MSE values from the 10 replications or partitions were obtained.

To further interpret the results of models' performance, we computed the percentage of variance explained by each term in the different fitted models after conducting a full dataset analysis (i.e., with no missing values).

2.7 Software

Genomic prediction analyses were computed in the R environment (R Development Core Team, 2023) and the models were fitted using the BGLR package (Pérez and De Los Campos, 2014). All models were fitted using 12,000 Markov Chain Monte Carlo iterations, incorporating a burn-in phase for the initial 2,000 iterations and a thinning factor of 5.

3 Results

3.1 Variance components

The percentage of variance explained by each term for SA and TCH is shown in Table 2. For both traits, the environmental component captured the largest proportion of variability. However, there are significant differences between traits, with 46%-48% of the total variance explained for SA while 80%-82% for TCH. Examining the decomposition of the genetic variance, we found that for SA the variance explained by genotype in model 1 (56.2% of the within-environments variance) is higher than the sum of the variance explained by genotype and genomic markers (12.9%

TABLE 2 Across and within environments percentage of explained variability by each model term for each trait.

Trait	Model	Across-environments variance (%)					Within-environments variance (%)			
		vE	vL	vg	vgE	vR	vL	vg	vgE	vR
SA	L+E	46.7	29.9			23.3	56.2			43.8
	L+E+G	48.1	6.7	17.9		27.3	12.9	34.5		52.6
	L+E+G+GE	47.7	6.2	18.0	15.4	12.7	11.9	34.4	29.5	24.2
TCH	L+E	82.2	6.2			11.6	34.9			65.1
	L+E+G	81.3	4.3	2.3		12.1	23.0	12.1		64.9
	L+E+G+GE	80.7	4.0	2.1	4.5	8.7	20.9	10.8	23.5	44.8

The first set of columns on the left side depict the percentage for all model terms: main effect of environments (vE), main effect of Genotype (vL), main effect of genomic markers (vg) and the interaction between genomic markers and environments (vgE), as well as the residual or unexplained variability (vR). The second set of columns on the right side depict the within-environment variance, i.e., the relative contribution of each term without considering the variance explained by the main effect of the environments.

10.3389/fpls.2024.1400000

+34.5%=47.4%) in model 2. A similar pattern was observed for TCH (34.9% in model 1 and 23.0%+12.1%=35.1% in model 2). Conversely, the variance explained by the interaction between genomic markers and environment is orthogonal to the other components, explaining the variability that cannot be captured by the main effects, thus reducing the residual variance.

3.2 Phenotypic correlation between environments

The phenotypic correlation values between environments for both SA and TCH traits are shown in Figure 2. A wide range of positive correlations was found between environments, with no high correlation values between environments sharing locations except in the case of SA in Balsora environments (r=0.822). Correlation values for Balsora environments with other locations were higher than those between Taula and Porvenir for SA. For TCH, we found lower correlations, with the maximum values observed between Taula 2019 and Porvenir environments (r=0.621 and r=0.588 for 2020 and 2021, respectively).

Figure 3 presents the accuracy achieved at each environment, as well as the overall mean accuracy and MSE values. The original values used to compute the means, as well as the values for scenarios with reduced calibration set are presented in Supplementary Table S1. The accuracy was generally higher when predicting SA compared to TCH for models M2 and M3 (M1 returned low results due to the lack of information to connect calibration and testing sets). As a general trend, the accuracy decreases as the number of overlapping genotypes increases. However, within the SA trait, the maximum accuracy was attained at the '28/3' allocation design. Furthermore, a local optimum was identified at the '12/19' allocation design, indicating that specific combinations of overlapping and non-overlapping genotypes can lead to improved predictive performance.

Regarding the models used in the study, it was found that M3, which accounts for G×E, generally outperformed the other models, but there were not significant differences with respect to M2, as stated in Supplementary Table S2. Additionally, as expected, the base model M1 consistently showed the lowest predictive accuracy across all scenarios, with the performance gap widening as the number of overlapping genotypes increased. Notably, the study consistently achieved the best predictions in specific environments, namely E1 (Balsora 2016), E2 (Balsora 2017), and E4 (Taula 2019), as can be noted in Supplementary Table S1. Figure 4 shows the average accuracy obtained across all calibration sizes and designs. In alignment with the results from Figure 3, we found that accuracy tends to decline as the number of non-overlapping genotypes decreases. This trend holds true for all models and designs for TCH, with a few exceptions for SA, specifically '12/19', '8/20', and '8/12' for models M2 and M3.

In contrast to the impact of decreasing the number of non-overlapping genotypes, reducing the calibration size (i.e., the number of plots) by allocating a lower number of overlapping genotypes appears to have a minimal effect on accuracy, as shown in the main diagonals ($i=j$) of Figure 4. Remarkably, in models M2 and M3, the accuracy for the '12/0' design surpasses that of all other designs within the '12/B' family, except for the one with the larger calibration size and, therefore, number of overlapping genotypes ('12/19'). This suggests that similar accuracy levels could be achieved by either decreasing the calibration size and using fewer overlapping genotypes or by maintaining the calibration size while evaluating more non-overlapping genotypes.

4 Discussion

The application of genomic prediction has found practical utility in sugarcane breeding. Recent studies have indicated moderate predictive ability for complex traits, including yield, fiber and sugar content, with great potential of increasing genetic gain (Deomano et al., 2020; Hayes et al., 2021; Yadav et al., 2021). Despite this, there is a need to find a balance between increasing genetic gain and reducing resource allocation in sugarcane breeding. Given fixed financial resources, it is often challenging to determine how to allocate a

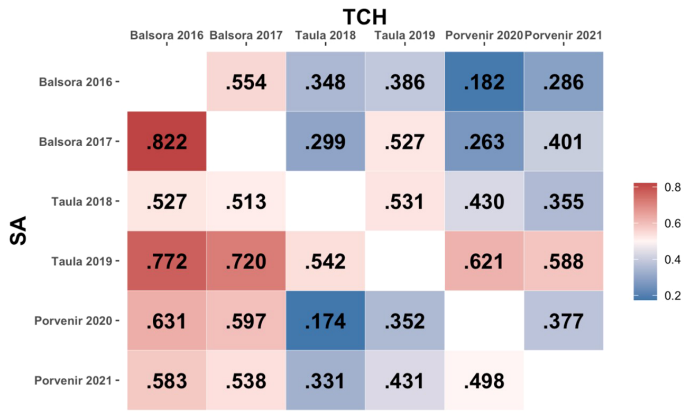

FIGURE 2
Phenotypic correlation between environments. The lower and upper triangular matrices represent, respectively, SA and TCH traits.

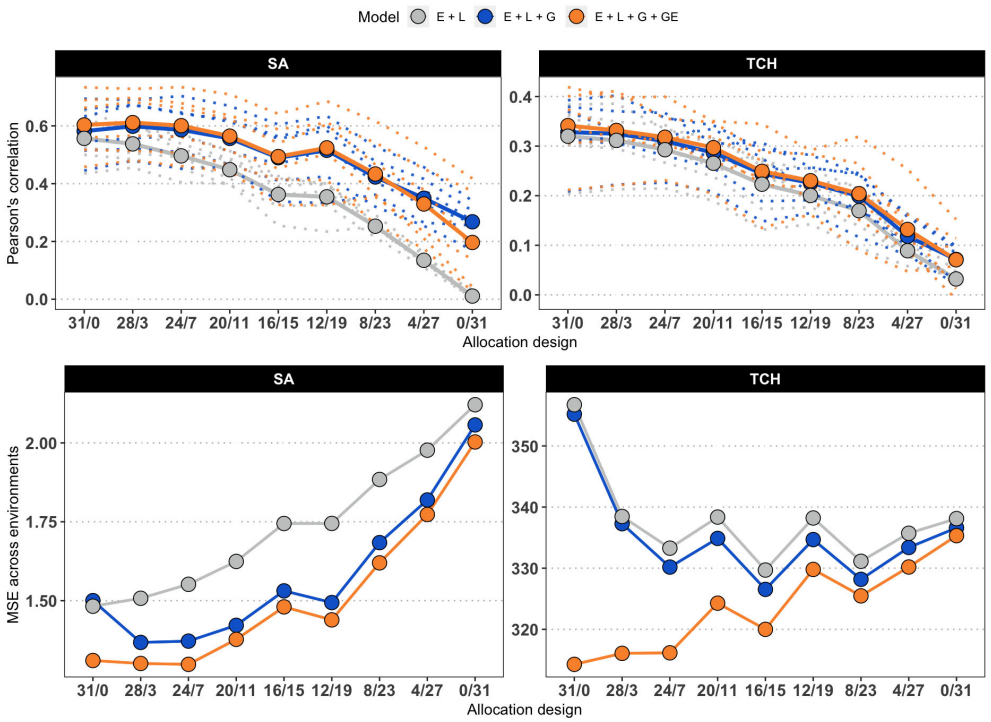

FIGURE 3
Average accuracy of models. Due to the increasing number of combinations, only the results for the allocation designs with the largest calibration set size (186) are shown for SA (left) and TCH (right). The *x*-axis represents the allocation designs, and the *y*-axis represents the accuracy measured as the Pearson's correlation (top panels) and Mean Squared Error (bottom panels) between predicted and observed values. Color denotes models. Dotted lines represent one specific environment. Solid lines and highlighted points represent the average accuracy across six environments.

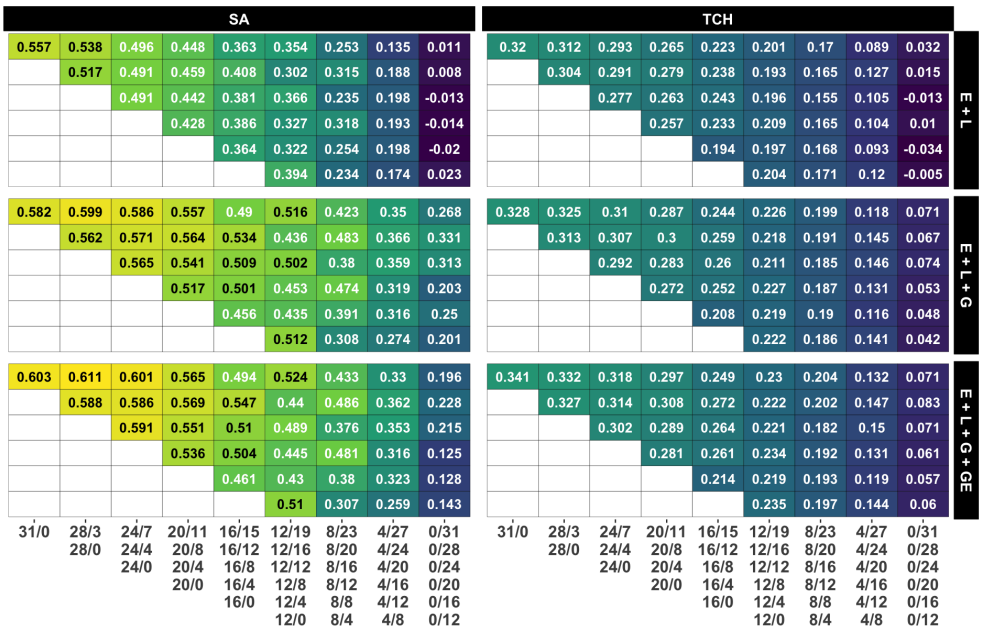

FIGURE 4
Average accuracy for all designs and calibration sizes. Each subplot represents the accuracy obtained by each combination of model and trait. Each grid within a subplot represents the accuracy for a specific design and calibration size, with yellow, green and blue colors representing high, medium and low accuracy values, respectively. Grids in the same row share the calibration size while grids in the same column share a common number of non-overlapping genotypes. The specific designs are presented below the columns.

predetermined number of genotypes for evaluation in targeted environments. To address this, optimization of resources in METs has become a routine practice. Among the strategies employed, the sparse testing allocation scheme with genomic prediction has been promising. This approach could identify the minimal number of candidate genotypes required for evaluation in METs and strategize their distribution in replicated or unreplicated field designs to achieve maximum genetic gains. In this study, we explored different sparse testing designs to determine the optimal calibration size and the trade-off threshold between prediction accuracy and selection intensity in sugarcane breeding.

One of the most notable trends is that CV1-like schemes ('0/B') pose significant challenges for M1. This phenomenon could be attributed to the fact that the predicted value of a genotype in each environment depends on the information of that exact genotype in other environments, which is not available in CV1 scenarios. By borrowing information from related genotypes through genomic information, M2 and M3 can extract patterns from the phenotypes of related genotypes and therefore, have a reasonable performance on CV1 scenarios. Similar results have been reported in a previous study by Atanda et al. (2022), where genomic and pedigree relationship between individuals can track segregating Quantitative Trait Loci, thus, explaining a large proportion of the genetic variance in a population.

On the other extreme, for the CV2-like schemes ('A/0') there were no notable differences in the predictive abilities of the models, even though they showed high predictive power (Figure 4). This is because the environmental effects are confounded with the genotype effect such that a single observation of a genotype in an environment is enough to predict its performance in other environments. Therefore, predictive ability strongly depends on the phenotypic information of tested individuals. The incorporation of genomic information explained a low genomic variance (SA: 17.9%, and TCH: 2.3%) in M2 (Table 2) and the model accounting for genotype by environment interaction in M3 captured a small information from related genotypes evaluated in correlated environments (SA: 15.4%, and TCH: 4.5%). This result implies a high correlation between the environments, given the observed small genotype differentiation across them. Other studies have also reported high genetic correlation between environments in sugarcane breeding programs (Deomano et al., 2020). As such, there is a need to redefine the target population of environments to avoid resource wastage during METs. These environments can be classified using important environmental factors and consolidated into mega-environments for efficient resource allocation (Resende et al., 2021).

The best accuracy is consistently obtained in scenarios with high 'A' value. The larger the number of different genotypes observed at least once, the higher the performance achieved. This suggests that G×E is small and, therefore, phenotyping a genotype in one environment provides useful insights for its prediction in the remaining environments. However, this might seem contradictory to the findings in Table 2, where a significant portion of the Within-environment variance was attributed to the G×E effect. It is crucial to interpret those results with caution because, while all the 1,116 available phenotypes were used to calibrate the model that estimates the variance components, the predictive models were trained using a maximum of 186 phenotypes.

Moreover, the slight gain in accuracy obtained by M3, which accounts for G×E, occurs in the scenarios with more non-overlapping genotypes. These findings indicate a potential cost saving opportunity can be achieved by allocating the genotypes to non-overlapping environments, that is, observing each genotype once across environment. This could decrease the cost of METs significantly in sugarcane breeding, as overlapping genotypes in some environments showed no gains in genomic prediction. This result is similar to the findings of Jarquín et al. (2020) in maize where marginal gains were observed even when overlapping genotypes were increased significantly. The plausible explanation is either the environments share strong environmental similarly or the traits exhibit low response to genotype-by-environment interaction. In that sense, Balsora, Taula and Porvenir locations are characterized by having the same soil taxonomy with soils from vertisol, inceptisol and mollisol orders, which have a fine, dry, and soils from deep to moderately deep (Carbonell et al., 2001), making them locations with similar behavior. However, even though the locations (E1 to E6) have similar soil profiles, Taula (E3 and E4) and Porvenir (E5 and E6) also share a lower drainage capacity when compared with Balsora (E1 and E2), affecting the performance of the planted varieties, impacting positively the correlations of SA and TCH among locations.

The breeding program at Cenicaña involves making decisions after 12 years of selection based on the phenotypic information collected during the early stages of selection (three stages). The METs provide information of genotype adaptability to the target population of environments of Colombia (CENICAÑA (Centro de Investigación de la Caña de Azúcar de Colombia), 1995). It is becoming important to integrate the allocation of phenotypes in plant breeding programs to optimize limited resources and increase genetic gain. Given the availability of genomic information during the breeding process, this can be leveraged to either evaluate more genotypes by increasing selection intensity or maximize the genetic gains with a fixed plot unit cost in sparse testing. As such, findings from this study in the case of sugar cane indicated that a cost saving opportunity could be achieved by overlapping a small number of genotypes in all environments while allocating the remaining genotypes to different environments. In addition, the use of a genomic prediction model that incorporates G×E (M3) allowed to capture genetic information among related genotypes in other environments to improve predictive ability.

The assumption of independence between genotypes held in M1 leads to biased models that perform poorly in comparison with the models that integrate genomic information to connect these. However, there is a clear trend to decrease accuracy with decreasing calibration set sizes. While the best accuracy is obtained with the full calibration set, there are designs with reduced calibration sizes that equal or even overpass the equivalent designs with larger calibration sizes. This is found especially in SA trait. Therefore, we can obtain similar accuracy values by reducing the budget or keeping the same budget but increasing the number of unique genotypes evaluated by

increasing the number of non-overlapping genotypes. This is similar to what has been obtained in maize, wheat, and soybean sparse testing allocation designs (Jarquín et al., 2020; Crespo-Herrera et al., 2021; Persa et al., 2023) but the difference we observed in sugarcane is that increasing the number of overlapping genotypes did not resulted in an increased predictive ability. This might also be due to the reduced training set sizes of our dataset compared with these studies. Interestingly, there are major differences in the average accuracy between traits. In the optimum scenarios, '31/0' and '28/3', the accuracy values for SA are 0.603 and 0.611, while for TCH are 0.341 and 0.332. SA is a more heritable trait with an extended genetic control while TCH trait is highly dependent on environmental conditions, specifically precipitation. Moreover, these differences might be caused by the phenotyping procedures.

5 Conclusions

In this study, we investigated the potential of genomic prediction for sparse testing resource allocation in sugarcane METs. It involved the utilization of different ratios of NOG/OG sugarcane clones and genome-based models, including G×E, to capture more genetic variability other than the main genomic effects. The obtained results indicated that genomic prediction models that incorporated G×E had the highest predictive response compared to other models in all allocation design scenarios. However, the advantage of the genomic model with G×E decreased with increasing NOG where highest predictive values were obtained. While this study focused on maximizing the genetic gain with a fixed cost per phenotype in sparse testing, the results showed that reducing the sample size of the genotypes assigned to environments (NOG) decreased the accuracy of genomic prediction. The trend decreased further with increasing overlapping genotypes evaluated across environments. This indicates that very few overlapping genotypes are needed across environments. This was attributed to a high environmental effect on the traits and moderate phenotypic correlation between environments. Generally, the results from this study showed that models including G×E can reduce resource allocation for phenotyping by up to 83% or increase the testing capacity by fivefold for multi-environmental trials in sugarcane. Therefore, sparse testing with genomic prediction is a promising strategy for maximizing genetic with fixed phenotyping cost in a breeding program. Premised on large environmental variance for the traits, we recommend the use of environmental factors to define TPEs to avoid investing limited resources in correlated environments without corresponding increase in genetic gain.

Data availability statement

The datasets and pipeline generated for this study can be found in the following link: https://uflorida-my.sharepoint.com/:f:/g/personal/jhernandezjarqui_ufl_edu/Ele_tW5RgC5PrfRHWnet5xsBzgOHCLzCZ_Yxz8SftLUv-Q?e=WhhHaq.

Author contributions

JG-A: Conceptualization, Formal analysis, Methodology, Visualization, Writing – original draft. PA: Writing – review & editing. FA: Conceptualization, Data curation, Resources, Writing – review & editing. JT-M: Resources, Writing – review & editing. JR: Conceptualization, Data curation, Resources, Writing – review & editing. RP: Writing – review & editing. JIS: Resources, Writing – review & editing. DJ: Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. JG-A was funded a UPM predoctoral grant as part of the program "Programa Propio I +D+i" financed by the Universidad Politécnica de Madrid. JIS was supported by the Beatriz Galindo Program (BEAGAL18/00115) from the Ministerio de Educación y Formación Profesional of Spain and the Severo Ochoa Program for Centers of Excellence in R&D from the "Agencia Estatal de Investigación" of Spain, grant SEV-2016SEV- -0672 (2017SEV- -2021)) to the CBGP. JIS was also supported by Grant PID2021-123718OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe, CEX2020-000999-S.

Conflict of interest

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

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1400000/full#supplementary-material

References

Abu-Ellail, F. F. B., Hussein, E. M. A., and El-Bakry, A. (2020). Integrated selection criteria in sugarcane breeding programs using discriminant function analysis. *Bull. Natl. Res. Cent* 44, 161. doi: 10.1186/s42269–020-00417–6

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data* (Cambridge, United Kingdom: Babraham Bioinformatics, Babraham Institute).

Atanda, S. A., Govindan, V., Singh, R., Robbins, K. R., Crossa, J., and Bentley, A. R. (2022). Sparse testing using genomic prediction improves selection for breeding targets in elite spring wheat. *Theor. Appl. Genet.* 135, 1939–1950. doi: 10.1007/s00122–022-04085–0

Atanda, S. A., Olsen, M., Crossa, J., Burgueño, J., Rincent, R., Dzidzienyo, D., et al. (2021). Scalable sparse testing genomic selection strategy for early yield testing stage. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.658978

Belamkar, V., Guttieri, M. J., Hussain, W., Jarquín, D., El-basyoni, I., Poland, J., et al. (2018). Genomic Selection in Preliminary Yield Trials in a Winter Wheat Breeding Program. *G3 Genes|Genomes|Genetics* 8, 2735–2747. doi: 10.1534/g3.118.200415

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

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

Carbonell, J., Amaya, A., Ortiz, B. V., Torres, J. S., Qunitero, R., and Isaacs, C. H. (2001). *Zonificación agroecológica para el cultivo de caña de azúcar en el valle del río Cauca* (Cali: Centro de Investigacion de la Cana de Azucar de Colombia).

CENICAÑA (Centro de Investigación de la Caña de Azúcar de Colombia) (1995). *El cultivo de la caña en la zona azucarera de Colombia*. Eds. C. Cassalett, J. Torres and C. e Isaacs (Cali, Colombia).

Crespo-Herrera, L., Howard, R., Piepho, H. P., Pérez-Rodríguez, P., Montesinos-Lopez, O., Burgueño, J., et al. (2021). Genome-enabled prediction for sparse testing in multi-environmental wheat trials. *Plant Genome.* 14, e20151. doi: 10.1002/tpg2.20151

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

Crossa, J., Pérez, P., de los Campos, G., Mahuku, G., Dreisigacker, S., and Magorokosho, C. (2011). Genomic selection and prediction in plant breeding. *J. Crop Improv.* 25, 239–261. doi: 10.1080/15427528.2011.558767

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Dellaporta, S. L., Wood, J., and Hicks, J. B. (1983). A plant DNA minipreparation: Version II. *Plant Mol. Biol. Rep.* 1, 19–21. doi: 10.1007/BF02712670

Deomano, E., Jackson, P., Wei, X., Aitken, K., Kota, R., and Perez-Rodriguez, P. (2020). Genomic prediction of sugar content and cane yield in sugar cane clones in different stages of selection in a breeding program, with and without pedigree information. *Mol. Breed.* 40, 38. doi: 10.1007/s11032–020-01120–0

FAO (2015). *Climate change and food security: risk and responses* (Rome, Italy: Food and Agriculture Organization of the United Nations), 122.

FAO (2018). *The Future of Food and Agriculture: Alternative Pathways to 2050* (Rome, Italy: Food and Agriculture Organization of the United Nations), 228.

Ferrão, L. F. V., Amadeu, R. R., Benevenuto, J., de Bem Oliveira, I., and Munoz, P. R. (2021). Genomic selection in an outcrossing autotetraploid fruit crop: lessons from blueberry breeding. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.676326

Goldemberg, J. (2008). The Brazilian biofuels industry. *Biotechnol. Biofuels* 1, 6. doi: 10.1186/1754-6834-1-6

Hayes, B. J., Wei, X., Joyce, P., Atkin, F., Deomano, E., Yue, J., et al. (2021). Accuracy of genomic prediction of complex traits in sugarcane. *Theor. Appl. Genet.* 134, 1455–1462. doi: 10.1007/s00122–021-03782–6

Hoang, N. V., Furtado, A., Botha, F. C., Simmons, B. A., and Henry, R. J. (2015). Potential for genetic improvement of sugarcane as a source of biomass for biofuels. *Front. Bioeng Biotechnol.* 3. doi: 10.3389/fbioe.2015.00182

Islam, M. S., Corak, K., McCord, P., Hulse-Kemp, A. M., and Lipka, A. E. (2023). A first look at the ability to use genomic prediction for improving the ratooning ability of sugarcane. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1205999

Islam, M. S., McCord, P., Read, Q. D., Qin, L., Lipka, A. E., Sushma, S., et al. (2022). Accuracy of genomic prediction of yield and sugar traits in saccharum spp. *Hybrids. Agric.* 12, 1436. doi: 10.3390/agriculture12091436

Jackson, P. A., and Hogarth, D. M. (1992). Genotype x environment interactions in sugarcane, 1. Patterns of response across sites and crop-years in north Queensland. *Aust. J. Agric. Res.* 43, 1447–1459. doi: 10.1071/AR9921447

Jaimes, H., Londono, A., Saavedra-Diaz, C., Trujillo-Montenegro, J. N., López-Gerena, J., Riascos, J. J., et al. (2024). Sequencing vs. amplification for the estimation of allele dosages in sugarcane (Saccharum spp.). *Appl. Plant Sci.*, E11574. doi: 10.1002/aps3.11574

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

Jarquin, D., De Leon, N., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I., et al. (2021). Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. *Front. Genet.* 11, 592769. doi: 10.3389/fgene.2020.592769

Jarquín, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., et al. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3 Genes Genomes Genet.* 10, 2725–2739. doi: 10.1534/g3.120.401349

Jarquín, D., Lemes da Silva, C., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., et al. (2017). Increasing genomic-enabled prediction accuracy by modeling genotype × Environment interactions in kansas wheat. *Plant Genome* 10. doi: 10.3835/plantgenome2016.12.0130

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357. doi: 10.1038/nmeth.1923

Larrahondo, A., and Torres, A. (1989). Evaluación y determinación del azúcar recuperable de la caña de azúcar. *Carta trimestral Cenicaña* 3, 12–14.

Mahadevaiah, C., Appunu, C., Aitken, K., Suresha, G. S., Vignesh, P., Mahadeva Swamy, H. K., et al. (2021). Genomic selection in sugarcane: current status and future prospects.Front. *Plant Sci.* 12. doi: 10.3389/fpls.2021.708233

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

Montesinos-López, O. A., Saint Pierre, C., Gezan, S. A., Bentley, A. R., Mosqueda-González, B. A., Montesinos-López, A., et al. (2023). Optimizing sparse testing for genomic prediction of plant breeding crops. *Genes (Basel)* 14, 927. doi: 10.3390/genes14040927

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

Persa, R., Canella Vieira, C., Rios, E., Hoyos-Villegas, V., Messina, C. D., and Runcie D and Jarquin, D. (2023). Improving predictive ability in sparse testing designs in soybean populations. *Front. Genet.* 14. doi: 10.3389/fgene.2023.1269255

Raboin, L. M., Pauquet, J., Butterfield, M., D'Hont, A., and Glaszmann, J. C. (2008). Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. *Appl. Genet.* 116, 701–714. doi: 10.1007/s00122-007-0703-1

R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org/.

Resende, M. D., Munoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., et al. (2018). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol.* 218, 1064–1074. doi: 10.1111/nph.15066

Resende, R. T., Piepho, H. P., Rosa, G. J. M., Silva-Junior, O. B., Silva, F. F., de Resende, M. D., et al. (2021). Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theor. Appl. Genet.* 134, 95–112. doi: 10.1007/s00122–020-03684-z

Roach, B. (1989). "Origin and improvement of the genetic base of sugarcane," in *Proceedings of the Australian Society of Sugar Cane Technologists- Annual Conference*. (Tweed Heads, NSW), 34–47.

Scortecci, C. K., Creste, S., Calsa, T., Xavier, M. A., Landell, M. G. A., Figueira, A., et al. (2012). Challenges, opportunities and recent advances in sugarcane breeding. *InTech.*, 267–296. doi: 10.5772/28606

Souza, G. M., Berges, H., Bocs, S., Casu, R., D'Hont, A., Ferreira, J. E., et al. (2011). The sugarcane genome challenge: strategies for sequencing a highly complex genome. *Trop. Plant Biol.* 4, 145–156. doi: 10.1007/s12042–011-9079–0

Tello, D., Gil, J., Loaiza, C. D., Riascos, J. J., Cardozo, N., and Duitama, J. (2019). NGSEP3: accurate variant calling across species and sequencing protocols. *Bioinformatics* 35, 4716–4723. doi: 10.1093/bioinformatics/btz275

Trujillo-Montenegro, J. H., Rodríguez Cubillos, M. J., Loaiza, C. D., Quintero, M., Espitia-Navarro, H. F., Salazar Villareal, F. A., et al. (2021). Unraveling the genome of a high yielding Colombian sugarcane hybrid. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.694859

UNCCD (2017). *Global Land Outlook* (Bonn, Germany: United Nations Convention to Combat Desertification (UNCCD), 340.

UNPF (2023). State of World Population 2023– 8 Billion Lives, Infinite Possibilities: The Case for Rights and Choices. Available online at: https://www.unfpa.org/swp2023.

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

Voss-Fels, K. P., Wei, X., Ross, E. M., Frisch, M., Aitken, K. S., Cooper, M., et al. (2021). Strategies and considerations for implementing genomic selection to improve

traits with additive and non-additive genetic architectures in sugarcane breeding. *Theor. Appl. Genet.* 134, 1493–1511. doi: 10.1007/s00122–021-03785–3

Waclawovsky, A. J., Sato, P. M., Lembke, C. G., Moore, P. H., and Souza, G. M. (2010). Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. *Plant Biotechnol. J.* 8, 263–276. doi: 10.1111/j.1467–7652.2009.00491.x

Wei, X., and Jackson, P. (2016). "Addressing slow rates of long-term genetic gain in sugarcane," in *Proceedings of the International Society of Sugar Cane Technologists: XXIX Congress*, Chiang Mai, Thailand. 480–484.

Xavier, A., Beavis, W., Specht, J., Diers, B., Mian, R., Howard, R., et al. (2022). Package 'SoyNAM'.

Yadav, S., Jackson, P. A., Wei, X., Ross, E. M., Aitken, K. S., Deomano, E. C., et al. (2020). Accelerating genetic gain in sugarcane breeding using genomic selection. *Agronomy* 10, 585. doi: 10.1007/s00122-021-03822-1

Yadav, S., Wei, X., Joyce, P., Atkin, F., Deomano, E., Sun, Y., et al. (2021). Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive genetic effects. *Theor. Appl. Genet.* 134, 2235–2252. doi: 10.1007/s00122–021-03822–1