University of São Paulo
"Luiz de Queiroz" College of Agriculture

Genomic prediction for soybean segregating populations: selection strategies and training set establishment

Leandro de Freitas Mendonça

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2019

Leandro de Freitas Mendonça
Agronomist

Genomic prediction for soybean segregating populations: selection strategies and training set establishment

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **ROBERTO FRITSCHE NETO**

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Mendonça, Leandro de Freitas

Genomic prediction for soybean segregating populations: selection strategies and training set establishment / Leandro de Freitas Mendonça. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2019.

61 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Seleção genômica 2. Ganhos de seleção 3. Seleção precoce 4. Matriz de correlação genômica 5. *Glycine max* I. Título

ACKNOWLEDGMENT

I thank God, the responsible all achievements in my life.

My wife Jéssika Angelotti Mendonça, that's with me since the beginning of this journey offering help, advice, encouragement and affection.

My family, for the love encouragement and unconditional support, especially my parents Romilda and Adelino, my sisters Cristiane and Fernanda.

The University of São Paulo - "Luiz de Queiroz" College of Agriculture for giving me technical and human resources for an excellent academic formation. Especially thanks for the Genetics Department.

The Allogamous Plant Breeding Laboratory, for the great professional and personal enrichment. To all the friends in the lab and to my advisor Professor Dr. Roberto Fritsche-Neto.

The Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Financial code 001 for granting.

The GDM seeds, to provide the data and granting the research. Specially thanks for Gaspar Malone, Nizio Giasson and Jose Clavijo.

Finally, I thank all people not mentioned that, in a way, contributed to the accomplishment of this work.

CONTENTS

RESUMO 6
ABSTRACT 7
1. INTRODUCTION 9
2. GENOMIC PREDICTION ENABLES EARLY BUT LOW-INTENSITY SELECTION IN SOYBEAN SEGREGATING PROGENIES 11
ABSTRACT 11
2.1. INTRODUCTION 11
2.2. MATERIALS AND METHODS 13
2.2.1. Population synthesis 13
2.2.2. Phenotypic data 14
2.2.3. Genomic data 17
2.2.4. Prediction analysis 17
2.2.5. Validation 19
2.3. RESULTS 20
2.3.1. Population structure 20
2.3.2. Phenotypic parameters 21
2.3.3. Predictions using independent populations 21
2.3.4. Predictions using cross-validation scenarios 23
2.3.5. Selection coincidence and predicted genetics gains 24
2.4. DISCUSSION 27
2.4.1. Predictive abilities 27
2.4.2. Impact on selection 29
REFERENCES 34
SUPPLEMENTARY TABLES 38
3. THE ACCURACY OF DIFFERENT STRATEGIES FOR BUILDING TRAINING SETS FOR GENOMIC PREDICTIONS IN SOYBEAN SEGREGATING POPULATIONS 41
ABSTRACT 41
3.1. INTRODUCTION 41
3.2. MATERIAL AND METHODS 44
3.2.1. Genotype information 44

3.2.2. Phenotype information 45
3.2.3. The panel of elite inbred lines 46
3.2.4. Evaluation of the TS scenarios 47
3.3. RESULTS 49
3.3.1. Population structure 49
3.3.2. Prediction abilities and optimization of training sets 50
3.3.3. Selection coincidences 52
3.4. DISCUSSION 53
REFERENCES 58

RESUMO

Predição genômica para populações segregantes de soja: estratégias de seleção e estabelecimento da população de treinamento

Novas cultivares de soja são geradas a partir de cruzamentos bi-parentais, seguido de etapas de seleção e avanço de homozigose, cuja ordem de número de gerações varia de acordo com o método de melhoramento adotado. Nas etapas iniciais, a pouca quantidade de sementes por progênie, além da grande quantidade de indivíduos inviabiliza testes à campo com boa acurácia seletiva. Nesse contexto, a seleção genômica vem como método preditivo alternativo à simples amostragem aleatória nessas etapas. Sendo assim, o objetivo desta pesquisa for explorar aspectos relevantes ligados à aplicação de predição genômica nas etapas iniciais de um programa de melhoramento de soja. Os resultados mostram boa capacidade preditiva (acima 0.4) para os caracteres estudados (produtividade, altura de plantas e maturidade), mostrando ser possível aplicar seleção genômica já em F2 e obter ganhos de seleção. Além disso, demostrou-se que é possível obter capacidades preditivas equivalentes a um set de treinamento com irmãos completos, compondo-o apenas com linhagens avançadas no programa de melhoramento, possibilitando a criação de populações de treinamento performantes sem a necessidade de avaliação previa de progênies da mesma família, o que possibilita a criação de sets de treinamento estáveis ao longo ao longo dos anos e aplicáveis em distintas famílias.

Palavras-chave: Seleção genômica, Ganhos de seleção, Seleção precoce, Matriz de correlação genômica, *Glycine max*

ABSTRACT

Genomic prediction for soybean segregating populations: selection strategies and training set establishment

New soybean cultivars are generated from bi-parental crosses, followed by selection and homozygosis increasement stages, which the order of number of generations can vary according to the breeding method adopted. In the initial steps, the low quantities of seeds per progeny and the large number of individuals to be tested, makes it impossible to obtain a high-quality evaluation on field. In this context, genomic selection comes as an alternative predictive method, instead of simple random sampling. Therefore, the objective of this research is to explore relevant aspects related to the application of genomic prediction in the initial stages of a soybean breeding program. The results show good prediction ability (above 0.4) for traits tested evaluated (yield, plant height and maturity), showing that it is possible to apply genomic selection already in F2 and obtain selection gains. In addition, it has been shown that it is possible to obtain predictive abilities equivalent to a full-sibs training set, establishing it only with advanced progenies of the breeding program, allowing the generation high predictive training populations without prior evaluation of within-family progenies, which allows the creation of stable training sets over the years and applicable in different families.

Keywords: Genomic selection, Selection gains, Early selection, Genomic relationship matrix, *Glycine max*

1. INTRODUCTION

One of the most important goals of breeding it to develop and release new varieties that should enables greater financial returns to the farmers. Breeders usually have some tools to make it easy and increase the efficiency of this process. One of these tools is call genomic selection (GS). The GS is a genetic-statistical approach capable to predict the performance of an individual, based in its molecular marker profile (Meuwissen et al. 2001). The first step is to obtain a training population, that must be genotyped, phenotype and related with the prediction set. Next, a linear or non-linear regression is usually used to estimate the alleles effect of each marker. Therefore, this vector can be used to predict the genomic estimated breeding value of a population (prediction set) that was just genotyped (Bernardo and Yu 2007).

The use of GS is recommend in phases that the phenotypic evaluation and selection is inefficient (Jannink et al. 2010). The early steps of soybean breeding is a great example of it, once there are usually lots of new progenies to test and a reduced number of seeds (not enough for multi-location testing), what makes impossible a high-quality field evaluation. In this sense, GS could be applied to early select the best progenies, saving resources by non-evaluation of low potential ones in later steps.

The key point in the use of GS is the balance between cost of genotyping and prediction ability. Currently advances in the sequencing technology have reduced the cost of genotyping (Muir et al. 2016), what increase the advantages of applying GS. In addition, the prediction ability of GS usually competes against low values of heritability in this phase for quantitative traits, such as yield. For that reason, in this research we investigated relevant aspects related to the use of GS in early steps of soybean breeding. In the first chapter, we focus on test different model's components and investigate the impact of selection intensity in the selection gains; and in the second, we explore different strategies to compose a stable and highly predictive training set.

REFERENCES

Bernardo, R., and J. Yu. 2007. Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci. 47(3): 1082. doi: 10.2135/cropsci2006.11.0690.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9(2): 166–177. doi: 10.1093/bfgp/elq001.

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819–29. http://www.ncbi.nlm.nih.gov/pubmed/11290733 (accessed 22 September 2017).

Muir, P., S. Li, S. Lou, D. Wang, D.J. Spakowicz, L. Salichos, J. Zhang, G.M. Weinstock, F. Isaacs, J. Rozowsky, and M. Gerstein. 2016. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 17(1): 53. doi: 10.1186/s13059-016-0917-0.

2. GENOMIC PREDICTION ENABLES EARLY BUT LOW-INTENSITY SELECTION IN SOYBEAN SEGREGATING PROGENIES

ABSTRACT

In soybean, new commercial lines are commonly obtained from bi-parental crosses, and the selection is performed as homozygosity increases. However, it is difficult to select for quantitative traits in the early steps of breeding, due to the high heterozygosity level and a vast number of new progenies, which sometimes lead breeders to randomly select for these traits in this phase. Therefore, we aimed to assess the impact of genomic selection in early generations of a soybean breeding program. Working on germplasm derived from two different maturity regions in Brazil, genotyped in F_2 and phenotyped in $F_{2:4}$ for grain yield, plant height, maturity rating and days to maturity, we compared the composition of different training populations, models with and without the genotype × environmental (G×E) interaction effect, and two types of relationship measurements (genetic similarity and Euclidian distance). Results showed superior performance of the Euclidian distance kernel over the standard VanRaden kernel in major scenarios tested. In general, G×E models did not obtain superior performance compared with mean principal models, and the training population composed only of the nearest progenies had the highest prediction ability. The best models achieved prediction abilities between 0.40 and 0.56, thereby enabling application of a low-intensity selection in F_2. As a result, half of the progenies could be discarded without missing a great part the good ones. Our results show that through genomic prediction, it is possible to select for quantitative traits in the early steps of breeding, which might increase the efficiency of the program in the advanced phases.

Keywords: Efficient selection, Quantitative traits, Selection coincidence, Molecular markers

2.1. Introduction

Current biotechnological advances, such as genotyping-by-sequencing (GBS) (Elshire et al. 2011), have expressively reduced genotyping costs. Thus, auxiliary breeding tools, such as genomic selection (GS), have grown in importance for breeding purposes. This is partly due to the ability of GS to reduce the number of in-field evaluated progenies and, simultaneously, increase the number of empirically assessed genotypes. Thus, it is possible to explore more diversity and increase the chances of finding superior recombinants for quantitative traits (Jannink et al. 2010).

GS consists of the prediction of the genomic estimated breeding value (GEBV), of individuals based on genomic data (Meuwissen et al. 2001). First, the breeder should obtain the training population, which is a group of phenotyped and genotyped individuals used to

train the prediction models. Later, using information from the training, it is possible to predict the GEBV of another related population, that has not been phenotyped. The correlation between true values and the predicted values gives the predictive accuracy (PA) of the model (Bernardo and Yu 2007).

One of the more utilized prediction method is the genomic best linear unbiased prediction (G-BLUP) (Bernardo 1994; Zhang et al. 2007), which is traditionally performed using a genomic relationship matrix (VanRaden 2007). Besides G-BLUP, genomic models can be fitted through different relationship matrices, also known as kernels, for example, the Gaussian kernel matrix (GK), which estimates the genomic relationship based on the Euclidian distance (Schaid 2010). In previous articles (Cuevas et al. 2016b; Sousa et al. 2017), GK has shown superior performance compared with the standard relationship matrix. In addition, both the matrices can be combined with genotype × environmental (G×E) interaction models. Currently, some models deal well with these effects (Burgueño et al. 2012; Cuevas et al. 2016a; Montesinos-López et al. 2016), but in cases with so many progenies and sites unbalanced, those models can be computationally intensive and result in biased parameter estimation. Hence, considering the two-step approach, the G×E interaction can be modeled in the field trial analysis model, and the overall genetic effect for progeny can be used on the GS model (Damesa et al. 2017).

Most of the reported results in soybean genomic selection derive from studying homozygous populations (Jarquín et al. 2014; Zhang et al. 2016; Duhnen et al. 2017), considering different kernels and G×E interaction. However, this does not fulfill the whole set of situations where prediction might be performed. Hence, little information is available about the performance of early stages of breeding, predicting segregating progenies and populations.

In soybean, new elite lines are commonly obtained from bi-parental crosses, generating segregating populations where the selection is performed as homozygosity increases (Bilyeu et al. 2010). However, early selection for quantitative traits has low success because of a weak correlation between the performances of the F_2 and advanced generations, mainly for traits with low heritability (Bernardo 1991). Hence, the phenotype at this point is not representative of yield potential for homozygous lines. Additionally, there are usually lots of progenies per cross, making in-field evaluation of all individuals infeasible. In this sense, the genomic prediction at early stages might increase the possibility of finding superior materials.

The selection coincidence, that is, the proportion of coincidence entries to be selected by different methods, and the percentage selected, which gives the proportion of this

selection, are important components to be considered when comparing different selection schemes (Bhering et al. 2015). Furthermore, Blondel et al. (2015) showed that the same PA could be obtained from different rankings. Thus, the use of a high percentage of selection can lead to a selection of lines with low performance, depending on how the GS ranked the progenies. Conversely, Resende et al. (2017) verified that the superiority of GS is enhanced over a high percentage of selection compared with phenotypic selection in *Eucalyptus*. Indeed, the ideal proportion of individuals selected depends on the step of the breeding program. For example, in early steps, a low selection intensity is recommended to avoid the early discard of elite materials, but it is not always possible to be applied, due to size of breeding population.

Since soybean is profoundly affected by the photoperiod length (Carpentieri-Pípolo et al. 2002), breeder works with specific groups of maturation in different latitudes zones or mega-environments. This physiological barrier makes germplasm interchange difficult between programs of different maturations zones, increasing the genetic distance among sets of germplasm adapted to different latitudes (Lado et al. 2016). Given that this may be a problem for predictions that involve progenies of different regions, the genetic relationship between the training and target population is a key factor of success (Bernardo and Yu 2007; Isidro et al. 2015; Fristche-Neto et al. 2018).

We aimed to define the best strategy for formation and optimization of training population and selection intensity for the traits grain yield (GY), plant height (PH), maturity rating (MR), and days to maturity (DM), in soybean segregating progenies, considering different kernels (VanRaden and Gaussian), ways to model G×E interaction, and the inclusion of lines that's comes from a different mega-environments.

2.2. Materials and Methods

2.2.1. Population synthesis

In Brazil, soybean breeding zones are divided into five mega-environments (M1 to M5) starting south and going north, with changes mainly in latitude, altitude, temperature, rain distribution and soils. Therefore, two datasets, one from M2 and other from M4, that's are the two main soybean production zones, were used. Maturity groups of M2 range from V to VII, while those in M4 range from VII to IX. These datasets represent the germplasm derived from two different breeding programs with distinct genetic backgrounds. Seven bi-

parental crosses were selected, to compose seven F_2 populations in each dataset. Those crosses included 11 distinct parents for M2, and 13 distinct parents for M4 (Supplementary Table S1). For each cross, 125 F_2 plants were generated, and all single F_2 plants were advanced on field to $F_{2:4}$ without selection. In total 875 $F_{2:4}$ progenies per dataset were derived across all seven bi-parental populations. In addition, for each cross, a group of 30 F_2 plants was randomly sampled among the 125 form the independent population for validation.

2.2.2. Phenotypic data

The traits evaluated were grain yield (weight of all seeds harvested in one plot converted to t ha^{-1}), plant height (avenge of 5 plants in the same plot in cm), maturity rating (measured on field taking checks with MR known as reference), and days to maturity (difference between sowing date and physiological maturity in days). Each trait was evaluated in a distinct number of sites within each dataset (Table 1). All evaluations were performed in the $F_{2:4}$ generation during the 2016/17 summer season. The experimental design was an augmented blocks design with 45 progenies and five common checks per block. All progenies were replicated once on each location. Each plot consisted of two rows five meters in length and spaced 0.5 m apart. All field evaluation was performed by GDM®, at company experiment stations.

Table 1. Evaluated sites for dataset M2 and M4 with its geographic coordinates in 2016/17 summer season.

Mega-environment	Trait	Site	Coordinates
M2	GY	Bela Vista – PR	22°26`S / 51°20`W / 400m
		Rolandia – PR	23°16`S / 51°27`W / 620m
		Floresta – PR	23°37`S / 52°03`W / 350m
		Toledo – PR	24°40`S / 53°48`W / 530m
		Santa Teresinha do Itaipu – PR	25°25`S / 54°25`W / 270m
	PH	Bela Vista – PR	22°26`S / 51°20`W / 400m
		Rolandia – PR	23°16`S / 51°27`W / 620m
		Palotina – PR	24°21`S / 53°55`W / 360m
	MR	Bela Vista – PR	22°26`S / 51°20`W / 400m
		Rolandia – PR	23°16`S / 51°27`W / 620m
		Palotina – PR	24°21`S / 53°55`W / 360m
		Toledo – PR	24°40`S / 53°48`W / 530m
	DM	Bela Vista – PR	22°26`S / 51°20`W / 400m
		Rolandia – PR	23°16`S / 51°27`W / 620m
		Palotina – PR	24°21`S / 53°55`W / 360m
M4	GY	Lucas do Rio Verde – MT (1)	13°09`S / 55°51`W / 430m
		Lucas do Rio Verde – MT (2)	13°03`S / 55°58`W / 420m
		Campo Novo do Parecis – MT	13°35`S / 57°53`W / 550m
		Sorriso – MT	12°25`S / 55°40`W / 380m
	PH	Campo Novo do Parecis – MT	13°35`S / 57°53`W / 550m
		Sorriso – MT	12°25`S / 55°40`W / 380m
	MR	Campo Novo do Parecis – MT	13°35`S / 57°53`W / 550m
		Sorriso – MT	12°25`S / 55°40`W / 380m
	DM	Campo Novo do Parecis – MT	13°35`S / 57°53`W / 550m
		Sorriso – MT	12°25`S / 55°40`W / 380m

GY: grain yield; PH: plant height; MR: maturity rating; DM: days to maturity (DM); M2: mega-environment 2 of Brazil; M4: mega-environment 4 of Brazil.

We applied a mixed linear model using the ASReml-R software package (Butler et al. 2009) to estimate (1) the genotypic value of progenies for each site and (2) a main genotypic value across all sites plus the G×E interaction effect, described by the following models:

$$y = X\boldsymbol{r} + Z_1\boldsymbol{g} + Z_2\boldsymbol{b} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{y}$ is a vector of observed phenotypic data (GY, PH, MR, or DM) per site; $\boldsymbol{r}$ is a vector of checks, considered as fixed; $\boldsymbol{g}$ is a vector of progeny effects, considered random, where $\boldsymbol{g} \sim N(\boldsymbol{0}, I\sigma_g^2)$; $\boldsymbol{b}$ is a vector of block effects, considered as random, where $\boldsymbol{b} \sim N(0, \sigma_b^2)$; $\boldsymbol{\varepsilon}$ is a vector of residuals, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, I\sigma^2)$, and $\boldsymbol{X}$, $\boldsymbol{Z_1}$, and $\boldsymbol{Z_2}$ are incidence matrices that relate the independent variables to the dependent variable $\boldsymbol{y}$.

$$\boldsymbol{y} = X\boldsymbol{r} + Z_1\boldsymbol{g} + Z_2\boldsymbol{s} + Z_3\boldsymbol{b} + Z_4\boldsymbol{ge} + \varepsilon \qquad (2)$$

where $\boldsymbol{y}$ is a vector of observed phenotypic data (GY, PH, MR, or DM) of all sites; $\boldsymbol{r}$ is a vector of checks, considered as fixed; $\boldsymbol{g}$ is a vector of the progeny effects, considered as random, where $\boldsymbol{g} \sim N(\boldsymbol{0}, I\sigma_g^2)$; $\boldsymbol{s}$ is a vector of sites, considered as random, where $\boldsymbol{s} \sim N(\boldsymbol{0}, I\sigma_s^2)$; $\boldsymbol{b}$ is a vector of the blocks within sites, considered as random, where $\boldsymbol{b} \sim N(\boldsymbol{0}, I\sigma_b^2)$; $\boldsymbol{ge}$ is a vector of G×E interaction, considered as random, where $\boldsymbol{ge} \sim N(\boldsymbol{0}, I\sigma_m^2)$, and $\boldsymbol{\varepsilon}$ is a vector of residuals, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, R)$. $\boldsymbol{X}$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$, $\boldsymbol{Z_3}$, and $\boldsymbol{Z_4}$ are incidence matrices that relate the independent variables to the dependent variable $\boldsymbol{y}$. The heterogeneous error variances were modeled through the $\boldsymbol{R}$ matrix by specifying site-specific residual variances.

Using variance components estimated from both models, we obtained the phenotypic accuracy (*AC*) (Deon et al. 2007) and heritability, according to the following expressions:

$$AC = \left[1 - \frac{1}{1 + b.\left(\frac{CVg}{CVe} \right)^2} \right]^{1/2} \qquad (3)$$

where $\boldsymbol{b}$ is the number of sites; $\boldsymbol{CVg}$ is the genetic coefficient of variation, and $\boldsymbol{CVe}$ is the residual coefficient of variation;

$$H^2 = \frac{\sigma^2 g}{\sigma^2 g + \frac{\sigma^2 ge}{s} + \frac{\sigma^2}{rs}} \qquad (4)$$

where H^2 is the broad-sense heritability on an entry-mean basis; $\boldsymbol{\sigma^2_g}$, $\boldsymbol{\sigma^2_{ge}}$, and $\boldsymbol{\sigma^2}$ are genotypic, G×E interaction and residual variance, respectively; $\boldsymbol{s}$ is the number of sites, and $\boldsymbol{r}$ is the number of replications in each site.

2.2.3. Genomic data

The DNA of all F_2 plants was extracted using a Kleargene™ spin plates kit (384-well format), from LGC Genomics®. All DNA sample were sent to Biotechnology Resource Center of Cornell University, where, genotyping-by-sequencing (GBS) was performed (Elshire et al. 2011). Single-end sequence reads were produced on a NextSeq 500 machine with 1×90, and HiSeq2500 with 1×100, respectively. The number of reads per sample was 0.5 million, on average.

Single nucleotide polymorphisms (SNPs) were called using Tassel 5.0 software (Bradbury et al. 2007) and Gmax_275_v2 (Schmutz et al. 2010) as reference genome. The Tassel pipeline used was the 'discovery' pipeline described in Glaubitz et al. 2014. GBS provided 195k SNPs before filtering of markers for quality control. Quality control was carried out using the Synbreed-R package (Wimmer et al. 2012). Tri-allelic and unmapped SNPs were removed from the marker dataset. Additionally, SNPs with a call rate less than 70%, and minor allele frequencies less than 5% were removed. Beagle software (Browning and Browning 2007), implemented in the Synbreed-R package was used for imputation of missing data. After filtering, 1,330 SNPs in the M2 dataset remained, and 1,363 SNPs in the M4 dataset remained. There were 900 SNPs in common between the M2 and M4 datasets.

2.2.4. Prediction analysis

Since the primary aim of the genomic selection in this study is to better rank the F_2 progenies, early discarding those with low performance and selecting those with best performance, we used the genotypes of F_2 and the phenotypes of $F_{2:4}$ progenies. Furthermore, we employed 16 prediction scenarios based on combinations of training population, target population, genomic kernel, and G×E interaction. Moreover, we applied the validation within and between datasets (M2 and M4) and, also, based on a joined dataset (M2 + M4). The kernels were $G = (XX'/p)$ (VanRaden 2007), where $\boldsymbol{X}$ is the centralized matrix of markers, and $\boldsymbol{p}$ is the number of markers; and GK, where $\boldsymbol{GK}(x_i x_j) = \exp(-h d^2_{ij}/\text{median}(d^2_{ij}))$; $\boldsymbol{h} = 1$, and $\boldsymbol{d_{ij}}$ is the Euclidean distance between the i^{th} and j^{th} individuals (Crossa et al. 2010). Additionally, to deal with the G×E interaction, we applied (5) the single-environment main genotypic effect model (SM) and (6) the multi-environment single variance G×E deviation model (MD) (Sousa et al. 2017) for predictions (Table 2).

Table 2. Prediction scenarios proposed with its combinations of training and target population, kernel, and model.

Ref.	Training population	Training population size	Target population	Kernel	Model
A	M2	687	M2	G	SM
B	M2	684	M2	GK	SM
C	M2	687	M2	G	MD
D	M2	687	M2	GK	MD
E	M2M4	1329	M2	G	SM
F	M2M4	1329	M2	GK	SM
G	M4	642	M2	G	SM
H	M4	642	M2	GK	SM
A	M4	642	M4	G	SM
B	M4	642	M4	GK	SM
C	M4	642	M4	G	MD
D	M4	642	M4	GK	MD
E	M2M4	1329	M4	G	SM
F	M2M4	1329	M4	GK	SM
G	M2	687	M4	G	SM
H	M2	687	M4	GK	SM

M2: mega-environment 2 of Brazil; M4: mega-environment 4 of Brazil; G: VanRaden kernel; GK: Euclidian distance kernel; SM: single-environment main genotypic effect model; MD: multi-environment single variance G×E deviation model.

The SM was as follows:

$$\boldsymbol{y} = \mu 1 + Z_1\boldsymbol{g} + \varepsilon \qquad (5)$$

where $\boldsymbol{y}$ is the vector of BLUPs obtained by model (2) (GY, PH, MR, or DM) considering all sites; $\boldsymbol{\mu}$ is the general mean; $\boldsymbol{g}$ is the vector of GEBVs, considered as random, where $\boldsymbol{g}\sim N(\boldsymbol{0}, \sigma_g^2\boldsymbol{K})$, where $\boldsymbol{K}$ is $\boldsymbol{G}$ or $\boldsymbol{GK}$ matrix, and $\boldsymbol{\varepsilon}$ is the residual, where $\boldsymbol{\varepsilon}\sim N(\boldsymbol{0}, I\sigma^2)$. $\boldsymbol{Z_1}$ is the incidence matrix that relates the effects of independent vectors to the dependent $\boldsymbol{y}$ vector.

The MD was given as:

$$\boldsymbol{y} = \mu 1 + Z_E\boldsymbol{\beta_E} + Z_1\boldsymbol{g} + Z_2\boldsymbol{ge} + \varepsilon \qquad (6)$$

where $\boldsymbol{y}$ is the stacked vector of BLUPs obtained by model (1) (GY, PH, MR, or DM) for each site; $\boldsymbol{\mu}$ is the general mean; $\boldsymbol{\beta_E}$ is a random vector of environmental effects, where $\boldsymbol{\beta_E}\sim N(\boldsymbol{0}, I\sigma_e^2)$; $\boldsymbol{g}$ is the vector of GEBVs, considered as random where $\boldsymbol{g}\sim N(\boldsymbol{0}, \boldsymbol{K})$; $\boldsymbol{\varepsilon}$ is the residual term where $\boldsymbol{\varepsilon}\sim N(\boldsymbol{0}, I\sigma^2)$, and $\boldsymbol{ge}$ is the vector of random effects of the interaction where $\boldsymbol{ge}\sim N(\boldsymbol{0}, [Z_1KZ_1']^{o}[Z_EZ_E'] \sigma^2_{ge})$, where $\boldsymbol{K}$ is $\boldsymbol{G}$ or $\boldsymbol{GK}$ matrix, and ^{o} indicates the

Hadamard product. $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$ and $\boldsymbol{Z_E}$ are the incidence matrices that relate the effects of independent vectors to the dependent $\boldsymbol{y}$ vector.

The SM and MD were fitted with Bayesian generalized linear regression (BGLR) R package (Perez 2014). The genomic estimated breeding values were obtained as a posterior means from the BGLR, with 100,000 iterations and a burn-in of 10,000.

2.2.5. Validation

Two types of validations were undertaken: 1) *n-fold cross-validation*: all progenies are randomly divided into *n* groups, applying prediction and take several estimations of PA, making possible deviation measurement; 2) *independent population*: estimation of PA using a group of individuals previously selected specifically for validation (Đorđević et al. 2019).

For *n-fold cross-validation*, we used a 10-fold approach, whereby the population was randomly divided into 10 groups, and 9 of these (training) were used to predict the remaining group (target), until all groups were predicted. This approach was performed five times. On each fold, the Pearson's correlation between the phenotypic observation and the genomic estimated breeding values of target group, was used as prediction accuracy estimator. For the MD, that consider multiple sites, the observations of lines in target were removed of all sites. This procedure is known as cross-validation 1, described in Sousa et al. (2017). We estimated the mean and confidence interval for prediction accuracy with 95% of confidence (standard deviation divided by square root of observations times Z score).

In the independent population, we used the training and target population previous split, as described in the Population synthesis topic. Therefore, just one scenario of training-target population was considered, without confidence interval measurement. Similarly, Pearson's correlation between the phenotypic observation and the genomic estimated breeding values of target was used as prediction accuracy estimator.

2.2.6. Comparison of scenarios

We obtained the PA, selection coincidence, as described by Matias et al. 2017, and the predicted mean of the improved population (population mean + genetic gain) for 5, 10, 25, and 50% of percentage selected; and based on that, we attained the selection gain (%). Additionally, using 10,000 random samples of the phenotype vector, we checked what

proportion of these random re-samplings resulted in a mean superior to improved population. We called this parameter the probability of random success. In scenarios with two mega-environments, to makes those means comparable, we fixed the degree of shrinkage using the values derived from a unique complete field model.

In addition, we compared the rankings, looking for the selection coincidence in each percentage of selection tested. When we randomly divide a ranking into two groups (selected and not selected), the probability of correctly allocating an individual to a given group is the proportion of this group related to the size of the population. Thus, in a random selection, it is expected that the percentage of selection used is equal to the likelihood to correctly select some individuals in this subset. For example, aiming to select the best 10%, we randomly choose 10% of individuals. In this case, it is expected that we obtained only 10% of the truly best 10%. Therefore, if the selection coincidence was higher than the percentage of selection used, the GS is better than the random selection.

2.3. Results

2.3.1. Population structure

We have provided a population structure analysis using principal components (Figure 1). As expected, progenies of the same mega-environments grouped together, and its parents spread around their respectively clouds. Moreover, it is possible to note that the cloud of M2 is too smaller than M4, showing that the southern group is more concise, what may explain the stronger adverse effect on PA in M2 including M4 data than the opposite. In addition, the two extra lines used to compose M4 populations may increases the diversity compared to M2, what contribute to the big M4's cloud.

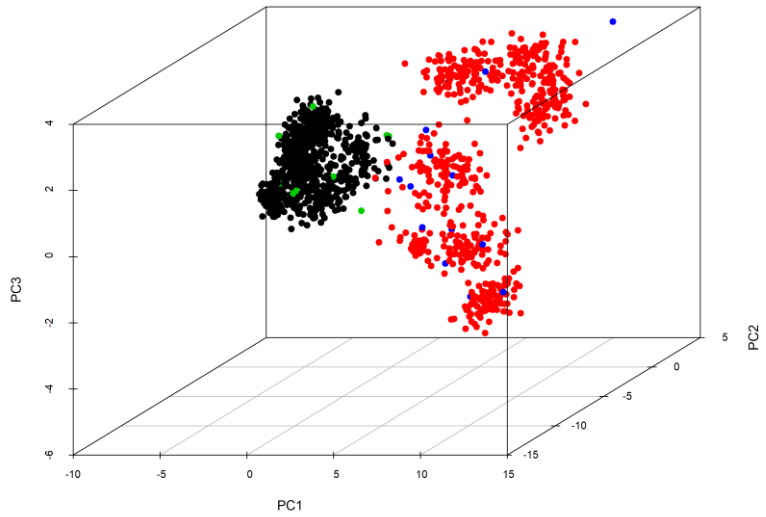

Figure 1. Principal components analysis plot for all individuals in datasets M2 and M4. M2 progenies is shown in black; M2 parents in green, M4 progenies in red; and M4 parents in blue.

2.3.2. Phenotypic parameters

After applying the described quality control steps to the datasets 763 and 713 $F_{2:4}$ progenies remained for M2 and M4 datasets, respectively. The likelihood ratio test (LRT) was conducted to verify significant differences by site and in the joint analysis. Significant differences were found for all traits in all sites in both datasets, except for GY which only showed significant differences between genotypes at one site of M2 and two sites of M4. However, in the joint analysis for each dataset, significant differences between progenies were found for GY. Among the traits, GY showed small values of H^2 and AC, as expected, given its high reaction to genotype x environmental interaction. Owing to the vast number of progenies evaluated, the experimental design used did not include replications within sites, just among sites and so it may have impaired *H*, as well as the LRT significance. (Supplemental Table S2 and S3).

2.3.3. Predictions using independent populations

Comparing the PA of all models, it was possible to observe consistent results among all traits (Figure 2). Overall, the best PA was obtained when the training population comprised only progenies from the same dataset (M2–M2; M4–M4). Instead, the lowest PA

was obtained when the training population and target population were made up using different datasets (M2–M4; M4–M2). Moreover, the joined dataset (M2 + M4) for training did not improve the PA. Furthermore, this parameter decreased in some cases. The inclusion of M4 to predict M2 (M2M4–M2) caused a stronger adverse effect on PA in comparison to the inclusion of M2 to predict M4 (M2M4–M4).

Regarding the kernels, both resulted in satisfactory PAs, but matrix GK was superior in the majority of tested scenarios (Figure 2). This superiority was enhanced in the model MD compared with the SM. Concerning the models dealing with G×E, there was no improvement in PA using the MD.

Overall, dataset M2 yielded better predictions than M4. For GY in M2, the models SM and MD, using matrix GK and only M2 progenies in training, were equivalent. In M4, the SM performed well for both relationship matrices, but the trait PH received the lowest performance. Considering the traits MR and DM, the results in M2 and M4 datasets were very similar, with the GK matrix providing the best performance.

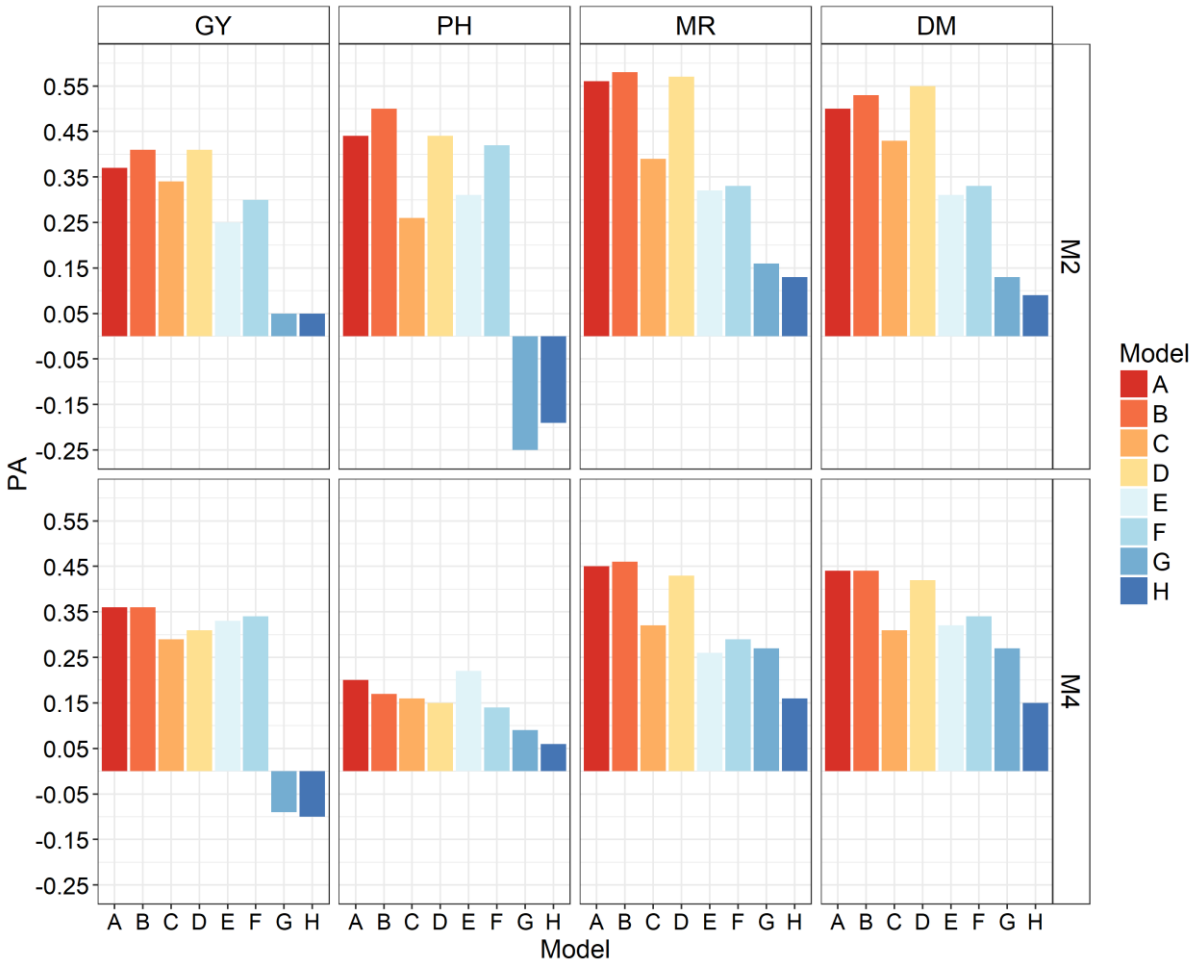

Figure 2. Prediction ability estimated using the validation strategy independent population for all scenarios tested in datasets M2 and M4. The legend is according to the reference letter described in Table 2.

2.3.4. Predictions using cross-validation scenarios

The PAs obtained by cross-validation confirmed the results shown in the independent validation (Figure 3). Models that included only progenies of the same dataset were the best (M2–M2; M4–M4). The inclusion of M4 to predict M2 (M2M4–M2) also yielded worse predictions than the opposite (M2M4–M4). G matrix presented a similar performance to GK, but GK remained superior in most scenarios, especially in M2 and with model MD. There were no relevant differences between the models SM and MD. For all traits in both datasets, GK–SM with the training population containing only progenies of the same dataset, were the best models, except for trait PH in M4 and trait DM in M2, in which, GK–MD yielded higher values.

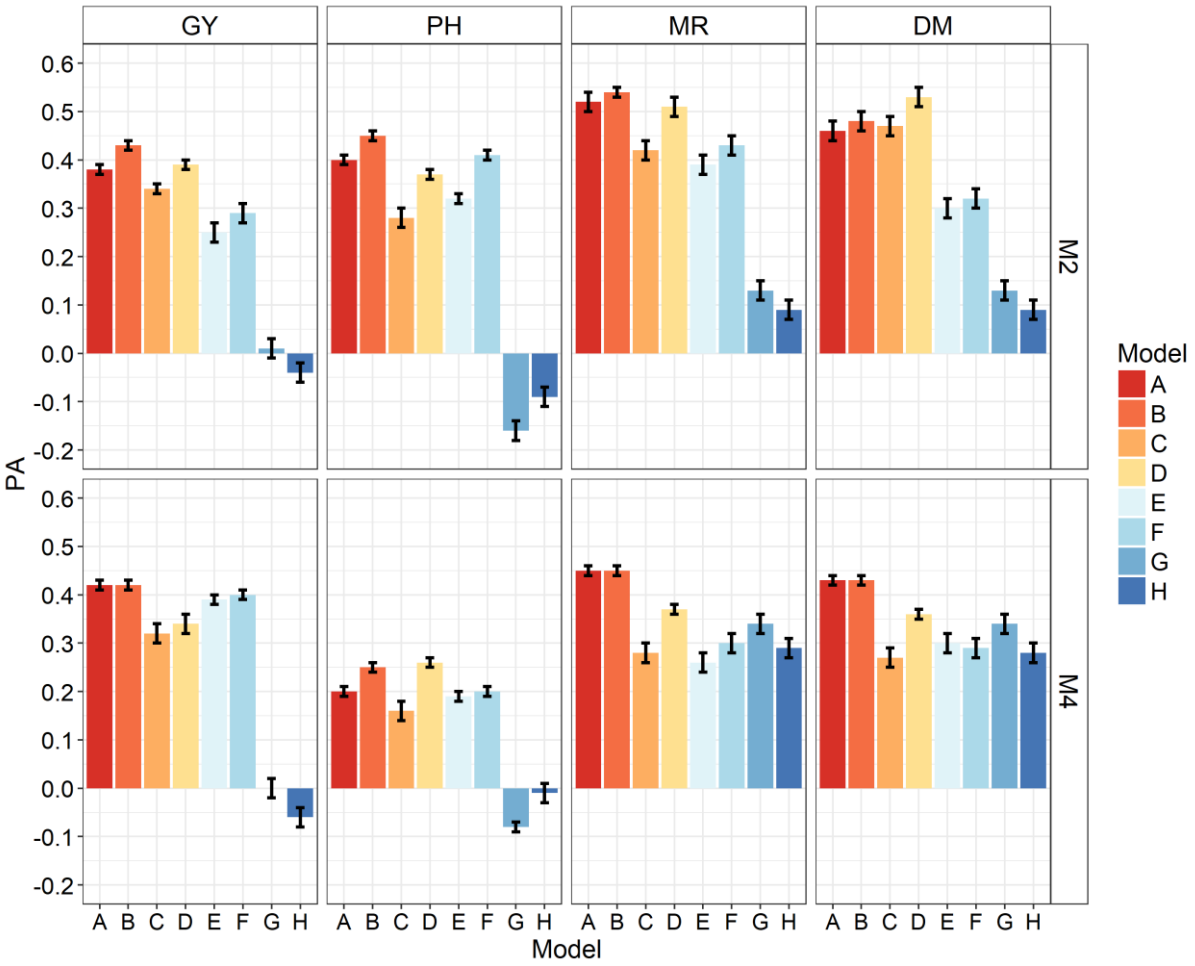

Figure 3. Prediction ability and confidence interval using the validation strategy 10-fold cross-validation for all scenarios tested in datasets M2 and M4. The legend is according to the reference letter described in Table 2.

2.3.5. Selection coincidence and predicted genetics gains

As mentioned, the scenarios that resulted in the best outcomes were GK–SM with training population composed only of progenies of the same dataset (M2–M2; M4–M4). Therefore, we chose this scenario to study the usefulness of GS in identifying the best F_2 individual (early selection).

It should be noted that the mean of real best progenies is always higher than the mean of the improved population by GS, both would be equal if GS and conventional rankings were equal (Table 3 and 4). In the last line of each trait presented, the average is shown. We used this value to estimate the selection gains (%). Since the usual selection in F_2 can be considered random for all traits, it is reasonable to assume that the expected mean in F_4 will be the mean of all F_4 progenies. Thus, all methods showed better results than the standard

approach regarding genetic gains. Even though in some cases, the objective of a soybean breeding program might be to reduce the value of traits, only by standardization, all inferences were made, as the aim was to increase them. In Table 3 and 4, the column selection coincidence (%) shows the coincidence of selection between GS and current selection on each percentage selected. By simulations, we produced the column probability of random success, which shows the likelihood of randomly attaining a higher mean than improved population by GS, using a random selection (Table 3 and 4).

Table 3. Parameters for all traits evaluated considering the prediction scenario M2–M2 using matrix GK and model SM.

Trait	PS (%)	Mean (real best)	IP	SG (%)	SC (%)	PRS
GY	5	4.62	4.32	3.35	10	0.24
	10	4,57	4.28	2.39	20	0.31
	25	4.46	4.31	3.11	40	0.27
	50	4.35	4.25	1.67	60	0.38
	100	4.18	4.18	-	-	-
PH	5	101.55	95.21	5.34	10	0.16
	10	99.63	94.36	4.40	35	0.22
	25	96.98	93.24	3.16	42	0.28
	50	94.34	92.51	2.36	72	0.32
	100	90.38	90.38	-	-	-
MR	5	66.62	64.69	4.20	40	0.10
	10	65.91	64.10	3.25	45	0.12
	25	64.49	63.69	2.59	58	0.15
	50	63.51	62.94	1.39	74	0.24
	100	62.08	62.08	-	-	-
DM	5	126.50	123.72	2.60	40	0.09
	10	125.38	123.15	2.12	45	0.11
	25	123.31	123.29	2.24	60	0.13
	50	121.96	121.36	0.64	67	0.25
	100	120.59	120.59	-	-	-

PS – percentage selected; IP – mean of improved population; SG – selection gain; SC – selection coincidence; PRS - probability of random success. GY – grain yield; PH – plant height; MR – maturity rating; DM – days to maturity.

Furthermore, it was possible to relate the parameter's value with the PA in the M2 dataset. For example, GY had the lowest PA, as well as the lowest SC and highest PRS. Conversely, despite the low performance in selection coincidence, the selection gains were satisfactory, achieving more than 2%. Usually, the selection coincidence decreases with reduction of percentage of selection, and in restrictive scenarios (e.g., 5% and 10%), traits GY

and PH got low values, indicating the disadvantage of using a low percentage of selection in low heritable traits (Table 4).

Table 4. Parameters for all traits evaluated considering the prediction scenario M4–M4 using matrix GK and model SM.

Trait	PS (%)	Mean (real best)	IP	SG (%)	SC (%)	PRS
GY	5	4.40	3.94	4.23	0	0.29
	10	4.30	3.97	5.03	23	0.24
	25	4.16	3.93	3.97	46	0.31
	50	4.02	3.88	2.65	60	0.40
	100	3.78	3.78	-	-	-
PH	5	100.17	88.41	2.36	0	0.37
	10	98.80	89.54	3.67	17	0.31
	25	95.77	87.46	1.26	30	0.42
	50	92.13	86.96	0.68	57	0.44
	100	86.37	86.37	-	-	-
MR	5	85.43	82.75	1.22	0	0.41
	10	84.75	82.65	1.10	0	0.41
	25	83.86	82.50	0.92	21	0.41
	50	83.22	82.57	1.00	69	0.40
	100	81.75	81.75	-	-	-
DM	5	116.49	113.56	0.95	0	0.41
	10	115.69	113.45	0.85	0	0.42
	25	114.70	113.25	0.68	21	0.41
	50	114.05	113.35	0.76	69	0.41
	100	112.49	112.96	-	-	-

PS – percentage selected; IP – mean of improved population; SG – selection gain; SC – selection coincidence; PRS - probability of random success; GY – grain yield; PH – plant height; MR – maturity rating; DM – days to maturity.

The inability of GS to select the best progenies under a low percentage selected was evident in M4, as all traits obtained 0% of selection coincidence for a 5% of percentage selected. Regardless of this trend, selection gains were positive, showing that even though the best progenies were not selected, some positive gain can be achieved. Traits MR and DM presented greater PA than GY, whereas, those traits had lower selection coincidence than GY (Figure 2 and Table 4). Likewise, for M2 with a high percentage of selection, the selection coincidence showed similar values across the traits.

2.4. Discussion

The use of GS at the beginning of a breeding scheme may be a powerful tool, once progenies with high potential performance can be earlier identified and selected. The success of this approach depends mainly on two factors: cost and prediction ability. The marker-assisted selection, for example, has been applied with successes in soybean breeding programs in early steps, due to the reduced cost of genotyping and high accuracy (Zhang et al. 2016; Kadam et al. 2016; Patil et al. 2016). Naturally, for quantitative traits, marker-assisted selection cannot well capture all effects involved, what leads to the use of GS approach, even with a more expensive genotyping cost (Campbell et al. 2015), due to the high number of markers demanded. However, currently, advances in sequencing area are leading to reductions in the genotyping costs (van Nimwegen et al. 2016), which makes the scenario more favorable to GS. In addition, the prediction abilities of GS in this step competes against low values of heritability, which could allow the application of GS even under intermediate accuracies. In this sense, it will be discussed the previous results in terms of prediction ability, impact on selection, and cost-efficiency.

2.4.1. Predictive abilities

GS in soybean has been reported as a useful tool. For instance, using the United States Department of Agriculture (USDA) Soybean Germplasm Collection, Jarquin et al. (2016) found a PA of 0.79 for GY. Duhnen et al. (2017) analyzed two inbred populations and revealed an average correlation of 0.60 for GY. Also, Matei et al. (2018) obtained a maximum of 0.70 for the same trait when working with recombinant inbred lines and varieties. Differently, our results (Figure 2 and 3) were inferior to the literature examples described above. The reasons for that is the use of few segregating populations of elite lines, what lead to a low variability and implies in high heterogeneity within lines and less variability among lines, reducing the heritability. However, it does not mean that GS cannot be used in segregating populations, hence, suitable estimates of selection gains and selection coincidences were obtained in our studies.

The GK matrix yielded a better PA than G matrix in majority scenarios, what agrees with the findings published by Sousa et al. (2017) that tested the same matrices on two maize hybrid populations. The authors mentioned the non-linear kernel ability to capture not only additive but epistatic effects (e g., additive–additive interaction), as the reason for its

superiority. In our case, it was noticeable that the superiority of GK was intensified in the MD. In these models, each progeny is predicted in all sites. Thus, it could be inferred, due to the type of effect capitate by GK, that more non-additive effects are taken into account in the prediction model compared with the SM, for which, most of those effects are structured in the field model. Cuevas et al. (2016b) observed similar results, working with a non-linear kernel matrix based on Euclidian distant in G×E interaction models for maize.

Despite the superiority of GK in the model MD, its computational time was expressively greater than by the SM, especially when using a cross-validation system. Moreover, while the PA did not differ much between the SM and MD, in this case, the use of the MD was not affordable. Actually, in the field model, the G×E interaction effect was low (data are not shown), what in M4 was caused by a similar environmental characteristics among sites. Nevertheless, in M2 the broad adaptations of those material due to many years of breeding reduced the complexity G×E interaction. Therefore, no crossover interaction is expected among the sites. Additionally, as reported by Muranty et al. (2015), the PA is strongly affected by the phenotypic performance and the H^2, which was quite low considering single sites (Supplemental Table S2 and S3), due to the experimental design used (without replications). It also partly explains the low prediction accuracy by the model MD versus the model SM.

Schulz-Streeck et al. (2012) stated that combining information from related populations may increase the PA. It is also known that soybean breeding germplasm has a narrow genetic basis (Zhu et al. 2003). However, the inclusion of progenies of different datasets in the training population reduced the PA. Therefore, checking the populations' structure using principal components analysis (Figure 1), there was a distinct separation of the progenies of both mega-environments, showing that despite the narrow genetic base, the high level of independent artificial selection increased the distance between the datasets. Lado et al. (2016) also determined that the predictions among or combining MEs lead to losses in PA compared to within mega-environments. In their case, the mega-environments were statically determinate, identifying groups of environments with small G×E interaction.

Another possible reason for the performance of PA when combining M2 and M4 is the low overlapping of markers between both datasets. Almost 32% of the total markers could not be used, potentially avoiding estimation of an important LD block. However, this hypothesis loses validity because Contreras-Soto et al. (2017) noted 900 markers were sufficient to capture the QTL effects, according to the number of LD blocks in soybean. The phenotypic overlapping could be a problem too, once lines of M2 were not evaluated in M4,

and vice-versa. In this case, the model cannot capture G X E interaction between mega-environments, what may reduce PA in the combined scenario.

The reduction in PA by combining populations assents the results of earlier research that the relationship between TS and breeding population is more important than the size (de Roos et al., 2009; Isidro et al., 2015; Lorenz and Smith, 2015). Since soybean is a self-pollinating plant, LD is large (Lam et al. 2010). Thus, combining datasets from different populations might be an attractive way to increase genetic variability, LD, amount of information, and the PA, as well. However, Zhong et al. (2009) demonstrated that several factors could affect this parameter, among them, the relationship is the most critical.

Interestingly, M4 had a more complex structure compared with M2 (Figure 1). A distinct division of progenies from those groups was apparent. The number of parents used to build them, and the considerable genetic diversity between the parents might explain this phenomenon. Thus, the substantial difference in genetic variability probably justifies the negative impact on PA for M2 when M4 is included in the training population, resulting in some negative correlations.

2.4.2. Impact on selection

One focus of GS studies is to improve the selection coincidence through different methods since high correlations could provide a more favorable selection. However, PA is a simple correlation, and so different rankings can be obtained with the same PA. Moreover, in a ranking context, some groups of individuals are more critical, for example, those ranked on the top. In this sense, the disordered tops can still have a high PA, which would lead to an elusive ability to select the best genotypes. These outcomes were also noted by Blondel et al. (2015), showing that PA can be poorly correlated with ranking. Thus, the focus of GS, similar to the traditional selection, must be to rank the individuals appropriately.

The superiority of GS over random selection was verified for all traits in M2, as all selection coincidences obtained were higher than the respective percentage selected (Figure 4A). In general, traits DM and MR had high selection coincidence, illustrating a positive association between PA and selection coincidence. In contrast, despite the high PA, the GS for M4 was worse than random selection in some scenarios of percentage selected (10–30%) (Figure 4B). It agrees with the results documented by Blondel et al. (2015) that PA is not the only measure to consider when assessing the usefulness of GS. It is evident when checking the difference between selection coincidence and percentage of selection, where positive

values were found for M2 and contrary in M4 for traits DM and MR, despite the great PA performance.

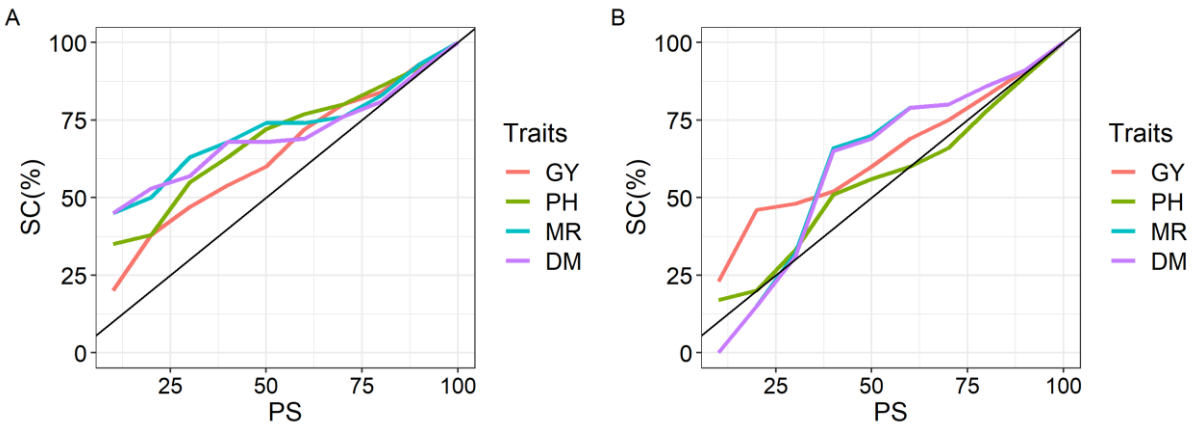

Figure 4. Selection coincidence × percentage selected for M2 (A) and M4 (B). Values over the black line indicate that the ranking of GS selects better than random sampling.

Similar results among all traits analyzed were verified, indicating that GS may have a limited ability to select individuals at the top of the ranking, although, it can be helpful to discard low-potential progenies under high percentage of selection. On that basis, we assessed the performance of GS to select the best by field performance at a specific percentage selected. For GY, using 60% as the percentages of selection, we select 100% of the 10 best progenies and more than 90% of the best 40 ones. The results for trait PH were similar. For traits MR and DM, 50% of percentage of selection was enough to identify the majority of the best progenies (Figure 5). Instead, the selection coincidence was less satisfactory in M4 than M2, as expected. Despite a percentage selected < 50%, acceptable results could be obtained. For PH, the trait with the lowest PA, even under low selection intensities, it was not possible to achieve good results (Figure 6).

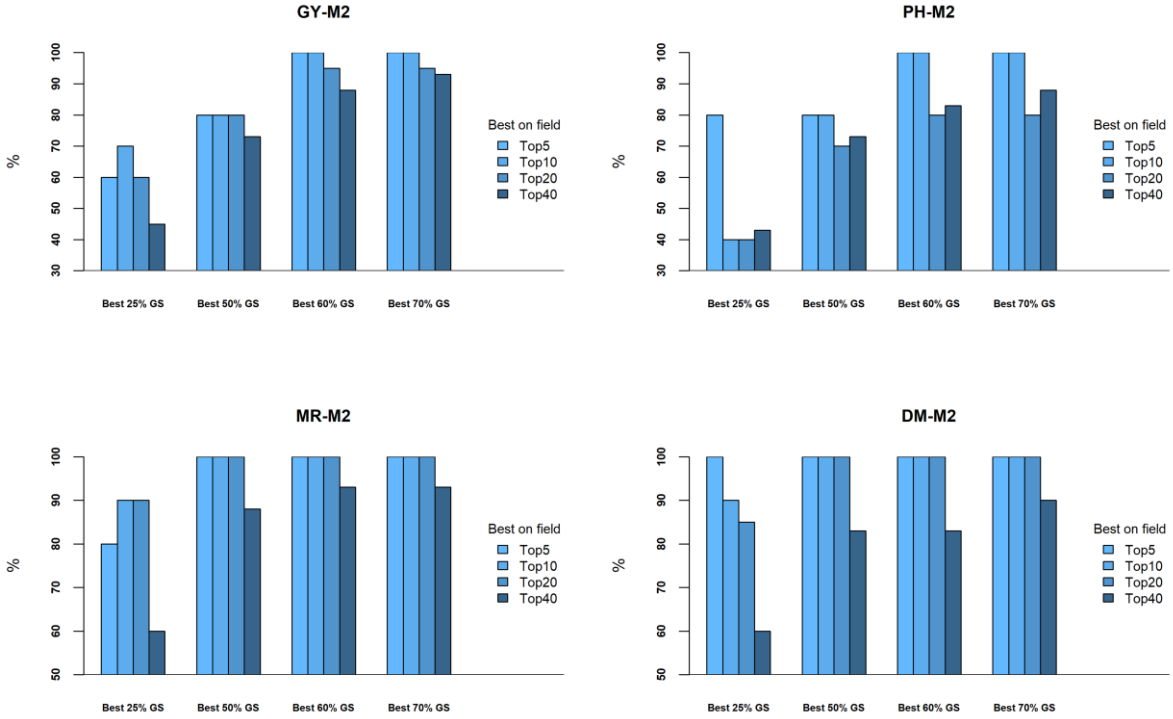

Figure 5. Selection profile of the best progenies on the field according to specific percentage of selection values in M2 for all traits tested. The *x*-axis shows different percentage of selection values by GS, and the *y*-axis shows the percentage of best progenies on field that these percentage of selection catch.

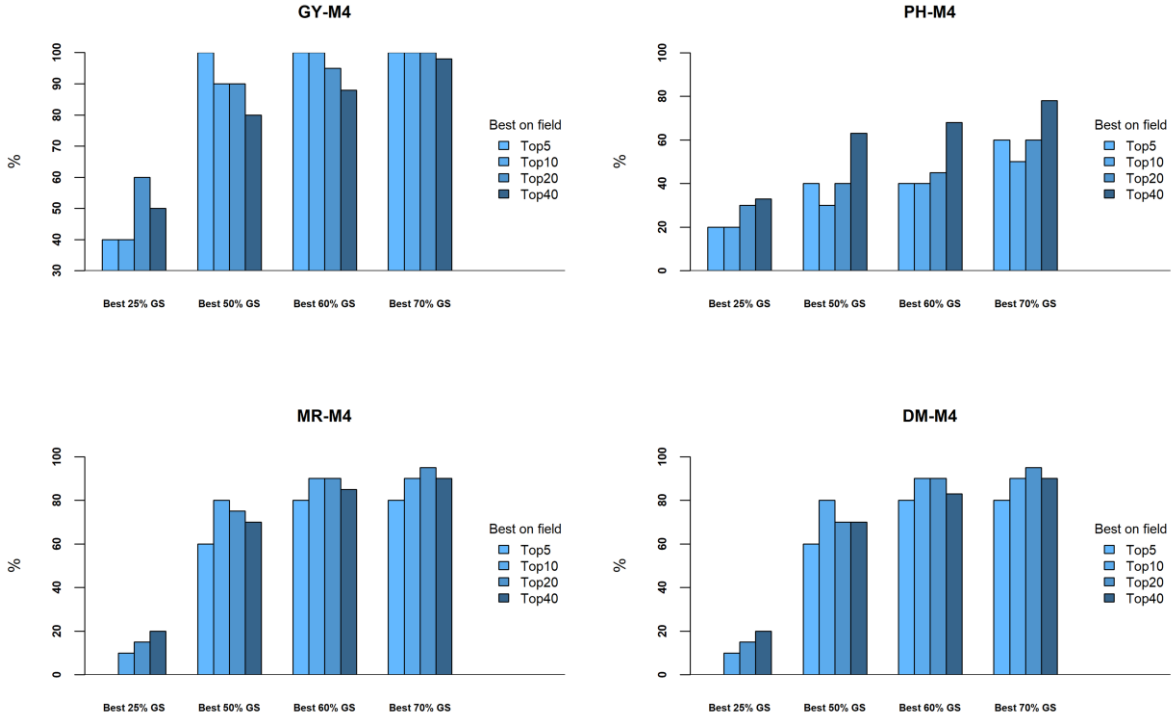

Figure 6. Selection profile of the best progenies on the field according to specific percentage of selection values in M4 for all tested traits. The *x*-axis shows different percentage of selection values by GS, and the *y*-axis shows the percentage of best progenies on field that these percentage of selection catch.

The use of a high percentage of selection may lead to select some weak progenies, but the best ones will not be wrongly discarded early on, as would happen under a lower percentage of selection. Thus, GS enables discarding half of the progenies in the early stages, reducing the costs with seeds multiplication and field evaluation, and increasing the probability of success. Moreover, F_2 is not a generation used to select, being formed by a random sample. Furthermore, as the number of individuals in this generation can be increased, depending on genotyping price and GS applied, this generation can be a determinant to improve the genetic gains achievable at later stages.

Therefore, GS allows breeders to conduct an initial screening selection for soybean in F_2, two generations before the traditional approach. Some breeding programs also perform modified single seed decedent until F_4, but although this strategy increases the homozygosis level and, consequently, probably offers the best performance of GS that we report, it also increases the time to generate a variety, due to additional steps required to increase the seeds quantities per progeny, and thereby perform testing in multi-locations. Nevertheless, the

general idea is to apply GS in the first segregating population step, where it is not possible to accurately predict phenotype, due to the reduced number of repetitions of the same genotype (e.g., a single plant). A great selection in this step increases the efficiency of the breeding process since fewer investments would be employed on progenies with low potential. The impact of this early selection can be interpreted in two ways: reducing the number of plots in later generations, while these were already selected, the same selection gains can be obtained with less experimental efforts; or keeping the same number of plots, thereby exploring a better population, which may increase selection gains.

In addition, a brief economic analysis is provided. According data providers, the cost of genotyping one sample (e.g., a F_2 plants) is around US$12,00, using Ion technology of ThermoFisher® (Galindo-González et al. 2015). This is above the cost to advance a single F_2 plant until a $F_{2:4}$ progeny (harvest F_3's and save F_4 seeds), that's around US$13.00, and above the field evaluation of $F_{2:4}$ progeny, that cost US$84,00 (considering 6 locations without repetitions). Therefore, considering just the initial steps of breeding (F_2 to $F_{2:4}$ evaluation), some scenarios are possible. We compare three strategies, considering one thousand F_2 plants and ignoring natural losses (e.g. death of plants, not enough seeds production), one without genomic selection and two considering a discard in F2 by GS (Table 5). Both GS strategies save resources compared to the conventional strategy, due to the reduced number of progenies evaluated in the most expensive step. Moreover, a selection intensity over 85% can be used to equal GS and non-GS costs. Among time, the expectation is an increase of GS advantage, once genotyping costs tends to decrease (Muir et al. 2016).

Table 5. Cost of three possible strategies for the beginning of soybean breeding program without and with GS in two selection intensity scenarios**.**

Step	Cost	(A) without GS		(B) 75% SI with GS		(C) 50% SI with GS	
		n	$	n	$	n	$
F2	$ 1.00	1000	1,000.00	1000	1,000.00	1000	1,000.00
Genotyping	$ 12.00	0	0.00	1000	12,000.00	1000	12,000.00
F3	$ 12.00	1000	12,000.00	750	9,000.00	500	6,000.00
F4	$ 84.00	1000	84,000.00	750	63,000.00	500	42,000.00
Total			97,000.00		85,000.00		61,000.00

n - number of individuals; $ - cost in US$; SI – selection intensity; GS – genomic selection.

In summary, we provided an interesting insight among the performance of GS, random sampling and selection based on traits measured in the field. It is important to remember that like GS, the field performance is also a prediction of the true genetic value of a

variety, and both have a bias. Therefore, for each step of a breeding program, the prediction method with less error must be chosen.

In conclusion, our findings suggest GS is a useful prediction approach in the early stages of soybean breeding. It is not recommended combining progenies from distinct mega-environments, to perform predictions because the high genetic divergence of these progenies negatively affects the PA. Moreover, the strategy using GK matrix and model SM with low intensity of selection lends the best trade-off between percentage of selection and selection coincidence.

REFERENCES

Bernardo, R. 1991. Correlation between testcross performance of lines at early and late selfing generations. Theor. Appl. Genet. 82(1): 17–21. doi: 10.1007/BF00231272.

Bernardo, R. 1994. Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Sci. 34(1): 20. doi: 10.2135/cropsci1994.0011183X003400010003x.

Bernardo, R., and J. Yu. 2007. Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci. 47(3): 1082. doi: 10.2135/cropsci2006.11.0690.

Bhering, L.L., V.S. Junqueira, L.A. Peixoto, C.D. Cruz, and B.G. Laviola. 2015. Comparison of methods used to identify superior individuals in genomic selection in plant breeding. Genet. Mol. Res. 14(3): 10888–10896. doi: 10.4238/2015.September.9.26.

Bilyeu, K.D., M.B. Ratnaparkhe, and C. Kole. 2010. Genetics, genomics and breeding of soybean. Science Publishers.

Blondel, M., A. Onogi, H. Iwata, and N. Ueda. 2015. A Ranking Approach to Genomic Selection (A de la Fuente, Ed.). PLoS One 10(6): e0128570. doi: 10.1371/journal.pone.0128570.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23(19): 2633–2635. doi: 10.1093/bioinformatics/btm308.

Browning, S.R., and B.L. Browning. 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. 81(5): 1084–1097. doi: 10.1086/521987.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 52(2): 707–719. doi: 10.2135/cropsci2011.06.0299.

Butler, D.G., B.R. Cullis, A.R. Gilmour, and B.J. Gogel. 2009. ASReml-R reference manual mixed models for S language environments. Train. Ser. QE02001: 149.

Campbell, N.R., S.A. Harmon, and S.R. Narum. 2015. Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Mol. Ecol. Resour. 15(4): 855–867. doi: 10.1111/1755-0998.12357.

Carpentieri-Pípolo, V., L.A. de Almeida, and R.A. de S. Kiihl. 2002. Inheritance of a long juvenile period under short-day conditions in soybean. Genet. Mol. Biol. 25(4): 463–469. doi: 10.1590/S1415-47572002000400016.

Contreras-Soto, R.I., M.B. de Oliveira, D. Costenaro-da-Silva, C.A. Scapim, and I. Schuster. 2017. Population structure, genetic relatedness and linkage disequilibrium blocks in cultivars of tropical soybean (Glycine max). Euphytica 213(8): 173. doi: 10.1007/s10681-017-1966-5.

Crossa, J., G. De Los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2): 713–724. doi: 10.1534/genetics.110.118521.

Cuevas, J., J. Crossa, O. Montesinos-Lopez, J. Burgueno, P. Perez-Rodriguez, and G. de Los Campos. 2016a. Bayesian Genomic Prediction with Genotype × Environment Interaction Kernel Models. G3 (Bethesda).: g3.116.035584. doi: 10.1534/g3.116.035584.

Cuevas, J., J. Crossa, V. Soberanis, S. Perez-Elizalde, P. Perez-Rodriguez, G. de Los Campos, O.A. Montesinos-Lopez, and J. Burgueño. 2016b. Genomic Prediction of Genotype x Environment Interaction Kernel Regression Models. Plant Genome 9(3): 1–20. doi: 10.1534/g3.116.035584.

Damesa, T.M., J. Möhring, M. Worku, and H.P. Piepho. 2017. One step at a time: Stage-wise analysis of a series of experiments. Agron. J. 109(3): 845–857. doi: 10.2134/agronj2016.07.0395.

Deon, M., V. De Resende, and J.B. Duarte. 2007. PRECISÃO E CONTROLE DE QUALIDADE EM EXPERIMENTOS DE AVALIAÇÃO DE CULTIVARES 1.

Đorđević, V., M. Ćeran, J. Miladinović, S. Balešević-Tubić, K. Petrović, Z. Miladinov, and J. Marinković. 2019. Exploring the performance of genomic prediction models for soybean yield using different validation approaches. Mol. Breed. 39(5): 74. doi: 10.1007/s11032-019-0983-6.

Duhnen, A., A. Gras, S. Teyssèdre, M. Romestant, B. Claustres, J. Daydé, and B. Mangin. 2017. Genomic selection for yield and seed protein content in Soybean: A study of breeding program data and assessment of prediction accuracy. Crop Sci. 57(3): 1325–1337. doi: 10.2135/cropsci2016.06.0496.

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

Fristche-Neto, R., D. Akdemir, and J.-L. Jannink. 2018. Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. Theor. Appl. Genet. 131(5): 1153–1162. doi: 10.1007/s00122-018-3068-8.

Galindo-González, L., D. Pinzón-Latorre, E.A. Bergen, D.C. Jensen, and M.K. Deyholos. 2015. Ion Torrent sequencing as a tool for mutation discovery in the flax (Linum usitatissimum L.) genome. Plant Methods 11(1): 19. doi: 10.1186/s13007-015-0062-x.

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline (NA Tinker, Ed.). PLoS One 9(2): e90346. doi: 10.1371/journal.pone.0090346.

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot, and M.E. Sorrells. 2015. Training set optimization under population structure in genomic selection. Theor. Appl. Genet. 128(1): 145–158. doi: 10.1007/s00122-014-2418-4.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9(2): 166–177. doi: 10.1093/bfgp/elq001.

Jarquín, D., K. Kocak, L. Posadas, K. Hyma, J. Jedlicka, G. Graef, and A. Lorenz. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics 15(1): 740. doi: 10.1186/1471-2164-15-740.

Jarquin, D., J. Specht, and A. Lorenz. 2016. Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions. G3: Genes|Genomes|Genetics 6(8): 2329–2341. doi: 10.1534/g3.116.031443.

Kadam, S., T.D. Vuong, D. Qiu, C.G. Meinhardt, L. Song, R. Deshmukh, G. Patil, J. Wan, B. Valliyodan, A.M. Scaboo, J.G. Shannon, and H.T. Nguyen. 2016. Genomic-assisted phylogenetic analysis and marker development for next generation soybean cyst nematode resistance breeding. Plant Sci. 242: 342–350. doi: 10.1016/j.plantsci.2015.08.015.

Lado, B., P.G. Barrios, M. Quincke, P. Silva, and L. Gutiérrez. 2016. Modeling Genotype × Environment Interaction for Genomic Selection with Unbalanced Data from a Wheat Breeding Program. Crop Sci. 56(5): 2165. doi: 10.2135/cropsci2015.04.0207.

Lam, H.-M., X. Xu, X. Liu, W. Chen, G. Yang, F.-L. Wong, M.-W. Li, W. He, N. Qin, B. Wang, J. Li, M. Jian, J. Wang, G. Shao, J. Wang, S.S.-M. Sun, and G. Zhang. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42(12): 1053–1059. doi: 10.1038/ng.715.

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.

Matei, G., L.G. Woyann, A.S. Milioli, I. de Bem Oliveira, A.D. Zdziarski, R. Zanella, A.S.G. Coelho, T. Finatto, and G. Benin. 2018. Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. Mol. Breed. 38(9): 117. doi: 10.1007/s11032-018-0872-4.

Matias, F.I., G. Galli, I.S. Correia Granato, and R. Fritsche-Neto. 2017. Genomic Prediction of Autogamous and Allogamous Plants by SNPs and Haplotypes. Crop Sci. 0(0): 0. doi: 10.2135/cropsci2017.01.0022.

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819–29. http://www.ncbi.nlm.nih.gov/pubmed/11290733 (accessed 22 September 2017).

Montesinos-López, A., O.A. Montesinos-López, J. Crossa, J. Burgueño, K.M. Eskridge, E. Falconi-Castillo, X. He, P. Singh, and K. Cichy. 2016. Genomic Bayesian Prediction Model for Count Data with Genotype × Environment Interaction. G3: Genes|Genomes|Genetics 6(5): 1165–1177. doi: 10.1534/g3.116.028118.

Muir, P., S. Li, S. Lou, D. Wang, D.J. Spakowicz, L. Salichos, J. Zhang, G.M. Weinstock, F. Isaacs, J. Rozowsky, and M. Gerstein. 2016. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 17(1): 53. doi: 10.1186/s13059-016-0917-0.

Muranty, H., M. Troggio, I. Ben Sadok, M. Al Rifaï, A. Auwerkerken, E. Banchi, R. Velasco, P. Stevanato, W.E. van de Weg, M. Di Guardo, S. Kumar, F. Laurens, and M.C.A.M. Bink. 2015. Accuracy and responses of genomic selection on key traits in apple breeding. Hortic. Res. 2: 15060. doi: 10.1038/hortres.2015.60.

van Nimwegen, K.J.M., R.A. van Soest, J.A. Veltman, M.R. Nelen, G.J. van der Wilt, L.E.L.M. Vissers, and J.P.C. Grutters. 2016. Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. Clin. Chem. 62(11): 1458–1464. doi: 10.1373/clinchem.2016.258632.

Patil, G., T. Do, T.D. Vuong, B. Valliyodan, J.-D. Lee, J. Chaudhary, J.G. Shannon, and H.T. Nguyen. 2016. Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. Sci. Rep. 6(1): 19199. doi: 10.1038/srep19199.

Perez, P. 2014. BGLR : A Statistical Package for Whole Genome Regression and Prediction. Genetics 198(2): 483–495. doi: 10.1534/genetics.114.164442.

Resende, R.T., M.D. V Resende, F.F. Silva, C.F. Azevedo, E.K. Takahashi, O.B. Silva-Junior, and D. Grattapaglia. 2017. Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. Heredity (Edinb). 119(4): 245–255. http://dx.doi.org/10.1038/hdy.2017.37.

de Roos, A.P.W., B.J. Hayes, and M.E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183(4): 1545–1553. doi: 10.1534/genetics.109.104935.

Schaid, D.J. 2010. Genomic similarity and kernel methods I: Advancements by building on mathematical and statistical foundations. Hum. Hered. 70(2): 109–131. doi: 10.1159/000312641.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, Q. Song, J.J. Thelen, J. Cheng, D. Xu, U. Hellsten, G.D. May, Y. Yu, T. Sakurai, T. Umezawa, M.K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X.-C. Zhang, K. Shinozaki, H.T. Nguyen, R.A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R.C. Shoemaker, and S.A. Jackson. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463(7278): 178–183. doi: 10.1038/nature08670.

Schulz-Streeck, T., J.O. Ogutu, Z. Karaman, C. Knaak, and H.P. Piepho. 2012. Genomic Selection using Multiple Populations. Crop Sci. 52(6): 2453. doi: 10.2135/cropsci2012.03.0160.

Sousa, M.B. e, J. Cuevas, E.G. de O. Couto, P. Perez-Rodriguez, D. Jarquin, R. Fritsche-Neto, J. Burgueno, and J. Crossa. 2017. Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype x Environment Interaction. G3 Genes|Genomes|Genetics 7(6): 1995–2014. doi: 10.1534/g3.117.042341.

VanRaden, P.M. 2007. Genomic Measures of Relationship and Inbreeding. Interbull Bull. (37): 111–114. doi: 10.1007/s13398-014-0173-7.2.

Wimmer, V., T. Albrecht, H.-J. Auinger, and C.-C. Schön. 2012. synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28(15): 2086–7. doi: 10.1093/bioinformatics/bts335.

Zhang, J., Q. Song, P.B. Cregan, and G.-L. Jiang. 2016. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycine max). Theor. Appl. Genet. 129(1): 117–30. doi: 10.1007/s00122-015-2614-x.

Zhang, Z., R.J. Todhunter, E.S. Buckler, and L.D. Van Vleck. 2007. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. J. Anim. Sci. 85(4): 881–885. doi: 10.2527/jas.2006-656.

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, Y.L., Q.J. Song, D.L. Hyten, C.P. Van Tassell, L.K. Matukumalli, D.R. Grimm, S.M. Hyatt, E.W. Fickus, N.D. Young, and P.B. Cregan. 2003. Single-nucleotide polymorphisms in soybean. Genetics 163(3): 1123–1134. doi: 10.1534/genetics.105.043877.

SUPPLEMENTARY TABLES

Supplementary Table S1. Parent combinations used to compose F_2 populations.

F_2 population	M2*	M4*
1	P03 (5.2) × P08 (6.0)	P15 (8.6) × P21 (7.2)
2	P04 (6.4) × P11 (6.5)	P18 (7.6) × P24 (7.1)
3	P09 (6.3) × P10 (6.4)	P19 (8.4) × P25 (8.4)
4	P07 (6.2) × P10 (6.4)	P14 (7.6) × P22 (7.6)
5	P01 (6.4) × P06 (5.9)	P17 (7.8) × P23 (8.5)
6	P05 (6.0) × P06 (5.9)	P16 (7.8) × P24 (7.1)
7	P02 (6.2) × P11 (6.5)	P13 (8.2) × P20 (7.4)

*maturity group is between parentheses; M2: mega-environment 2 of Brazil; M4: mega-environment 4 of Brazil

Supplementary Table S2. Phenotype parameters for M2 progenies.

Trait	Site	Mean	SD	LRT	h^2	AC
GY	Bela Vista do Paraiso - PR	4.29	0.24	*ns*	0.17	0.41
	Rolandia - PR	4.87	0.32	*	0.28	0.53
	Floresta - PR	3.16	0.18	*ns*	0.07	0.26
	Toledo - PR	4.66	0.23	*ns*	0.20	0.45
	Santa Teresa do Itaipu - PR	3.95	0.23	*ns*	0.16	0.40
	All sites	4.19	0.36	*	0.51	0.77
PH	Bela Vista do Paraiso - PR	78.20	4.21	*	0.56	0.75
	Rolandia - PR	92.10	3.99	*	0.30	0.55
	Palotina - PR	100.52	4.35	*	0.36	0.60
	All sites	90.29	7.15	*	0.81	0.87
MR	Bela Vista do Paraiso - PR	62.72	1.20	*	0.79	0.89
	Rolandia - PR	62.66	0.93	*	0.36	0.60
	Palotina - PR	61.33	1.03	*	0.90	0.95
	Toledo - PR	61.85	0.88	*	0.79	0.88
	All sites	62.15	0.77	*	0.81	0.91
DM	Bela Vista do Paraiso - PR	128.23	1.53	*	0.79	0.89
	Rolandia - PR	121.77	1.07	*	0.35	0.59
	Palotina - PR	131.56	1.34	*	0,91	0.95
	All sites	120.60	6.88	*	0.64	0.83

* - significant at 0.01 of probability at likelihood ratio test; *ns* - not significant, GY – grain yield; PH – plant height; MR – maturity rating; DM – days to maturity; SD – standard deviation; LRT – likelihood ratio test (genotypic effect); h2 – heritability; AC – phenotypic accuracy.

Supplementary Table S3. Phenotype parameters for M4 progenies.

Trait	Site	Mean	SD	LRT	h^2	AC
GY	Lucas do Rio Verde 1 - MT	3.84	0.36	*ns*	0.17	0.41
	Lucas do Rio Verde 2 - MT	3.84	0.40	*	0.37	0.60
	Campo Novo do Parecis - MT	4.15	0.42	*	0.35	0.59
	Sorriso - MT	3.25	0.20	*ns*	0.13	0.37
	All sites	3.77	0.31	*	0.53	0.77
PH	Campo Novo do Parecis - MT	91.90	7.68	*	0.61	0.78
	Sorriso - MT	81.79	6.02	*	0.54	0.73
	All sites	86.77	6.87	*	0.73	0.86
MR	Campo Novo do Parecis - MT	81.63	1.79	*	0.44	0.67
	Sorriso - MT	81.72	1.43	*	0.26	0.51
	All sites	81.70	1.29	*	0.75	0.88
DM	Campo Novo do Parecis - MT	112.33	1.94	*	0.45	0.67
	Sorriso - MT	112.47	1.55	*	0.27	0.53
	All sites	112.42	1.29	*	0.74	0.89

* - significant at 0.01 of probability at likelihood ratio test; *ns* - not significant; GY – grain yield; PH – plant height; MR – maturity rating; DM – days to maturity; SD – standard deviation; LRT – likelihood ratio test (genotypic effect); h2 – heritability; AC – phenotypic accuracy.

3. THE ACCURACY OF DIFFERENT STRATEGIES FOR BUILDING TRAINING SETS FOR GENOMIC PREDICTIONS IN SOYBEAN SEGREGATING POPULATIONS

ABSTRACT

The design of the training set is a key factor in the success of the genomic selection approach. Considering the highly dynamic nature of lines inclusion in soybean breeding programs, generating a training set that endures across the years and regions is challenging. Therefore, we aimed to define the best strategies for building training sets to apply genomic selection in soybean segregating populations for traits with different genetic architectures. For that, we used two data sets for grain yield (GY) and maturity group (MG) from two different soybean breeding regions in Brazil. Moreover, five training set schemes were tested. In addition, we included a training set formed by an optimization algorithm based on a predicted error variance. The predictions achieved good values for both traits; reaching 0.5 in some scenarios. The best strategy changes according to the trait. Although the best performance was achieved with the use of full-sibs in the MG, for GY, full-sibs and a set of advanced lines were equivalent. For both traits, there was no improvement of predictive ability by training set optimization. Furthermore, the use of advanced lines from the same breeding program is recommended as a training set for GY, once, the training set is continuously renewed, closely related to the breeding populations, and no additional phenotyping is needed. On the other hand, to improve prediction accuracies for the MG, it is necessary to use training sets with less genetic variability but with more segregation resolution.

Keywords: Maturity, Efficient breeding, Selection gains, Population structure, Genetic architecture

3.1. Introduction

The genomic prediction approach is a tool for predicting phenotypes from molecular markers that allows breeders to select superior individuals without field evaluations (Meuwissen et al. 2001). To make this possible, marker values must be estimated using a phenotyped and genotyped population, known as the training set (TS). Then, marker effects are used to predict a new population genotyped. In the early stages of a soybean breeding program—in an F_2 generation for example—there are many progenies, a condition that makes it impossible to test all of them under field conditions. Moreover, the F_2 phenotype does not adequately represent the performance of its offspring in high homozygous levels because of the high heterozygote rate and low heritability for quantitative traits (Costa et al. 2008).

Usually, a random sample of the progeny from each cross is used to form the F_2 population. Therefore, genomic selection (GS) could be used to replace this random sampling process. Thus, applying GS in this step could be advantageous if a high quantity of F_2 progeny can be evaluated without the evaluation of its offspring in the field. Furthermore, only F_2 progeny with high yield potential in later stages can be selected early.

Regarding the performance of GS, the TS consists of the highest cost in the process; it demands a well-phenotyped, genotyped, and large population (Bernardo and Yu, 2007). Consequently, one of the greatest challenges in GS is to build an affordable TS that can accurately predict several generations (Heffner et al. 2009). Building an affordable TS is especially challenging in a breeding program because of the addition of new parents every year, and the selection could increase the distance between breeding populations and the TS after a number of generations. Thus, it is necessary to refresh the TS over time.

Concerning the composition and structure of the TS, some of the relevant aspects observed were its relative nature and population structure. In this sense, the easiest way is not always the best. In maize, Fristche-Neto et al. (2018) demonstrated the high impact of the tester on the population structure, and how this creates a disinclination for building a TS and apply predictions in other populations. The critical point is that the effect tester is too strong and can create a high correlation with the alleles' performance and the tester; this reduces the accuracy of the predictions for new populations. Therefore, the test cross—one of the most-used breeding designs for maize—is the worst mating design to use as a TS. Full-sibs and half-sibs have great potential as TS populations. Full-sibs share half of the additive variance and one-fourth of the dominant variance, and half-sibs share one-fourth of the additive variance. Moreover, they have similar population structures and share many coincidental combinations of alleles; this phenomenon increases the accuracy due to the more trustworthy estimations of the marks' effects (Riedelsheimer et al. 2013).

Another critical factor to consider is the trait targeted for prediction. In soybeans, the two most important traits are grain yield (GY) and maturity groups (MG). GY determines the profitability of the crop because the main commercial product of soybeans is their grains. Hence, farmers are always searching for soybean varieties with high GY. MG is a classification that relates the photoperiod sensitivity of soybeans with the number of days the soybeans take to complete a cycle. In the classification used currently, soybeans are divided into 13 MGs (Hartwig 1973). In practice, the MG determines the region in which the soybean will be sown. Thus, different latitudes zones require soybeans with different MGs. For example, in the USA, the most-sown groups are II–IV (Mourtzinis and Conley 2017); in

Brazil, groups V–VIII are predominant (Alliprandini et al. 2009). Furthermore, the variability within zones is significant; for example, in Brazil there are precocious varieties in some specific latitude regions that are able to yield two crops (usually maize after soybean) within the same year (Zdziarski et al. 2018).

In genomic terms, GY and MG have different genetic architectures. GY is highly polymorphic and affected by environmental conditions, whereas MG is an oligogenic trait controlled by a few genes (*E series*) (Langewisch et al. 2014; Jiang et al. 2014). These differences may affect the predictions and the strategies for building the TS. According to Hayes et al. (2010), genomic prediction performance is affected by the number of loci that control the trait, and high accuracies are observed for traits with few loci and large effects. Despite the different heritably between both traits, Matei et al. (2018) obtained a successful prediction for GY and MG; they achieved a value greater than 0.6 for predictive ability (PA) in the context of population structure. In some scenarios, MG was active above 0.8, showing the high potential of GS for this trait. In addition, Smallwood et al. (2019) showed that the use of MG to divide the populations can increase the accuracy of predictions when they used lines with similar maturities as the TS. This result indicates greater success in a more closely related TS with the prediction set, since a similar MG indicates that lines originated from the same breeding region.

New soybean crosses are performed every year, and the focus is to predict early steps of breeding without full- or half-sibs evaluated in high homozygous levels of those. In this case, two scenarios are possible: using advanced lines that come from previous crosses or using a panel of elite lines. Usually breeding programs have both, but the adoption of a TS must be performed carefully once the accuracy is reduced when relatedness among progenies decreases (Würschum et al. 2017). Considering the necessity of applying GS in the F_2 of soybeans and based on the description above, a high accuracy is not expected if the TS and breeding population are composed of genotypes from different origins. However, considering the strong linkage disequilibrium (LD) (Lam et al. 2010) and the narrow genetic basis in soybeans (Zhu et al., 2003), it could be possible to build a TS using representative lines and their progenies from related parents to predict offspring obtained from different new crosses. Additionally, as breeding programs usually have extensive historical data of parental crossing blocks, these can be used as the TS for the F_2 offspring. This strategy was reported by He et al. (2016), who found that a TS comprised of several winter wheat lines and filtered for phenotype quality resulted in increased accuracy for predicting new recombinant inbred lines (RILs).

Breeding programs can be disadvantaged, in context of GS, by the rapid rate of change in their breeding populations' structure; this characteristic can make the TS unstainable. However, breeding programs must evaluate many RILs every year, so they can obtain a robust TS using algorithms for the optimization of training sets (OTS). These algorithms assist the user in choosing appropriate progenies with predefined population sizes based on the genetic structure and genomic information of the prediction set, and they can maintain or increase the accuracy level by reducing the size of the TS. The algorithm—using only the genotypic information of the individuals—applies principal components to the markers' matrix to make an approximate prediction error variance (PEV) and select those individuals in the TS that best fit the prediction set (Isidro et al. 2015). Therefore, the algorithm uses the genotypic information of the whole germplasm to identify which ones will compose the best TS. Moreover, it is possible to build a personalized TS to predict different genetic backgrounds, as observed in the breeding populations of distinct latitude regions and MGs.

Therefore, we aim (a) to define the best strategies for building a TS to apply genomic selection in soybean segregating populations; (b) to understand the traits' genetic architecture that may affect the choice of the best TS strategies and; (c) to verify if it is possible to improve prediction accuracy using algorithms in the criteria for building a TS. We discuss the implications of each strategy concerning PA and selection coincident (SC).

3.2. Material and methods

3.2.1. Genotype information

Two datasets that represent different breeding germplasms, called M2 and M4, from distinct regions (South and Midwest) in Brazil were used. These represent the two most important soybean production regions in Brazil, that manage maturity groups from V to VII in M2 and VII to IX in M4. For each, we selected seven bi-parental crosses to compose seven F_2 populations. Those crosses involved 11 lines for M2 and 13 lines for M4 (Table 1). For each bi-parental cross, we selected 125 F_2 plants; we obtained 875 F_2 plants per dataset by using this method. The DNA of all F_2 progenies were extracted using 384-Kleargene Spin Plates kit, from LGC Genomics®. All samples were sent to the Genomic Diversity Facility at Cornell University to perform GBS. SNP Calling was performed using Tassel 5.0 (Bradbury et al. 2007); this resulted in approximately 195000 SNPs. QC was performed using Synbreed-

R (Wimmer et al. 2012) with eliminations of triallelic SNPs and scaffolds; 70% of the call rate and 5% of the minimum allele frequency (MAF). Data input was performed using Beagle (Browning and Browning 2007) in the Synbreed-R. The final number of markers used was 1370.

Table 1 – Parental combinations used to produce F_2 populations.

F_2 population	M2*	M4*
1	P03 (5.2) × P08 (6.0)	P15 (8.6) × P21 (7.2)
2	P04 (6.4) × P11 (6.5)	P18 (7.6) × P24 (7.1)
3	P09 (6.3) × P10 (6.4)	P19 (8.4) × P25 (8.4)
4	P07 (6.2) × P10 (6.4)	P14 (7.6) × P22 (7.6)
5	P01 (6.4) × P06 (5.9)	P17 (7.8) × P23 (8.5)
6	P05 (6.0) × P06 (5.9)	P16 (7.8) × P24 (7.1)
7	P02 (6.2) × P11 (6.5)	P13 (8.2) × P20 (7.4)

*maturity group is between parentheses; M2: mega-environment 2 of Brazil; M4: mega-environment 4 of Brazil

3.2.2. Phenotype information

All F_2 plants were advanced to F_4 without selection; this process resulted in 875 $F_{2:4}$ progeny. GY in t ha^{-1} and MG were evaluated in 5 locations for M2 and 4 locations for M4 during the 2016 and 2017 summer seasons (Table 2). The experimental design was an augmented blocks design with 45 progenies and five common checks per block. All progenies were replicated once on each location. Each plot consisted of two rows five meters in length and spaced 0.5 m apart. All field evaluation was performed by GDM®, at company experiment stations.

Table 2 – Sites evaluated for each dataset.

Dataset	Locations	Coordinates
M2	Bela Vista – PR	22°26’S / 51°20’W / 400m
	Rolandia – PR	23°16’S / 51°27’W / 620m
	Floresta – PR	23°37’S / 52°03’W / 350m
	Toledo – PR	24°40’S / 53°48’W / 530m
	Santa Teresinha do Itaipu – PR	25°25’S / 54°25’W / 270m
M4	Lucas do Rio Verde – MT (1)	13°09’S / 55°51’W / 430m
	Lucas do Rio Verde – MT (2)	13°03’S / 55°58’W / 420m
	Campo Novo do Parecis – MT	13°35’S / 57°53’W / 550m
	Sorriso – MT	12°25’S / 55°40’W / 380m

We used a mixed models procedure with restricted maximum likelihood / best linear unbiased predictor (REML/BLUP) using ASReml-R (Butler et al. 2009) to predict the breeding values of these progenies, according to the following model:

$$y = X_1\boldsymbol{r} + X_2\boldsymbol{g} + Z_1\boldsymbol{s} + Z_2\boldsymbol{b} + Z_3\boldsymbol{ge} + \varepsilon$$

where $\boldsymbol{y}$ is the vector of phenotype data in all sites; $\boldsymbol{r}$ is the vector of checks added to the mean, considered as fixed; $\boldsymbol{g}$ is the vector of the progenies' effect, considered as fixed; $\boldsymbol{s}$ is the vector of locations, considered as random, where $\boldsymbol{s} \sim N(0, \sigma_s^2)$; $\boldsymbol{b}$ is the vector of the blocks within sites, considered as random, where $\boldsymbol{b} \sim N(0, \sigma_b^2)$; $\boldsymbol{ge}$ is the vector of the G×E interaction, considered as random where $\boldsymbol{ge} \sim N(0, \sigma_i^2)$; $\boldsymbol{\mathcal{E}}$ is the experimental error, where $\boldsymbol{\mathcal{E}} \sim N(0, \boldsymbol{R})$ where $\boldsymbol{R}$ is the diagonal residual variance matrix. $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$, and $\mathbf{Z_3}$ are incidence matrixes that relate the effects of independent vectors of each matrix to the dependent **y** vector. The structure of the R matrix was a heterogenic diagonal, with a residual variance component for each location.

3.2.3. The panel of elite inbred lines

In order to enhance the information from the TS, we used two groups of advanced—one from each region—as a panel of elite lines (PEL). The PELs were composed of 324 elite lines in M2 and 249 elite lines for M4. The obtention of the genomic information followed the same procedure of progeny data. These lines were not the same lines that originated from the segregated populations, but they have the same genetic base. For 7 years (2011–2017), the lines for M2 were evaluated for GY and MG at 40 locations; for M4, 33 locations were evaluated over the same time period. All field evaluation was performed by GDM®, at company experiment stations. The mixed model equations (REML/BLUP) using ASReml-R (Butler et al. 2009) were used to obtain the BLUEs, according to the following equation:

$$y = X\boldsymbol{g} + Z_1\boldsymbol{s} + Z_2\boldsymbol{ge} + Z_3\boldsymbol{b} + \varepsilon$$

where $\boldsymbol{y}$ is the vector of phenotype data in all sites; $\boldsymbol{g}$ is the vector of lines added to the mean, considered as fixed; $\boldsymbol{s}$ is the vector of the sites (a combination of location and year), considered as random, where $\boldsymbol{s} \sim N(0, \sigma_s^2)$; $\boldsymbol{ge}$ is the vector of G×E interaction, considered as random, where $\boldsymbol{ge} \sim N(0, \sigma_i^2)$; $\boldsymbol{b}$ is the vector of the nested effect of block in sites, considered

as random, where $\boldsymbol{b} \sim N(0, \sigma_b^2)$; $\boldsymbol{\mathcal{E}}$ is the experimental error, where $\boldsymbol{\mathcal{E}} \sim N(0, \boldsymbol{R})$ where $\boldsymbol{R}$ is the diagonal residual variance matrix. $\mathbf{X}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$, and $\mathbf{Z_3}$ are incidence matrixes that relate the effects of independent vectors of each matrix with the dependent $\mathbf{y}$ vector.

3.2.4. Evaluation of the TS scenarios

In order to predict the performance of the $F_{2:4}$ progenies, we performed five scenarios for the TS. In A, the TS comprised a part of the $F_{2:4}$ progenies using 10-fold cross-validation, in which there were progenies from the same bi-parental populations (full-sibs) in the TS and validation set (VS). This scenario may not fit with a real application of GS in the early stages of a soybean breeding program once we have all the F_2 evaluated in $F_{2:4}$, but it can be used to check the potential PA of the TS in cases in which some crosses are repeated.

B and C are scenarios that represent the use of different segregating populations for the TS (not full-sibs as in A, but some half-sibs). To simulate these conditions, in B, the TS comprised all parent lines and their $F_{2:4}$ progenies, except for one parent line and its offspring that was used for the VS. For C, we considered six of seven crosses and their progenies for the TS, and we considered the remaining cross as the VS. This approach was applied for each line and cross, respectively. These situations tested if the use of $F_{2:4}$ progenies that come from new crosses can be predicted by preexisting $F_{2:4}$ progenies.

In D, the PEL, which was not part of the previous three scenarios, was used as the TS to predict all $F_{2:4}$ progenies. The parents of the populations used for validation are included in the PEL. For E, with the same system of cross-validation as in A, we combined A and D and used the PEL and a portion of the $F_{2:4}$ progenies as the TS to predict the other part of the $F_{2:4}$ progenies. In this case, only $F_{2:4}$ progenies participate in cross-validation, and the PEL was wholly included in the TS for all interactions (Figure 1).

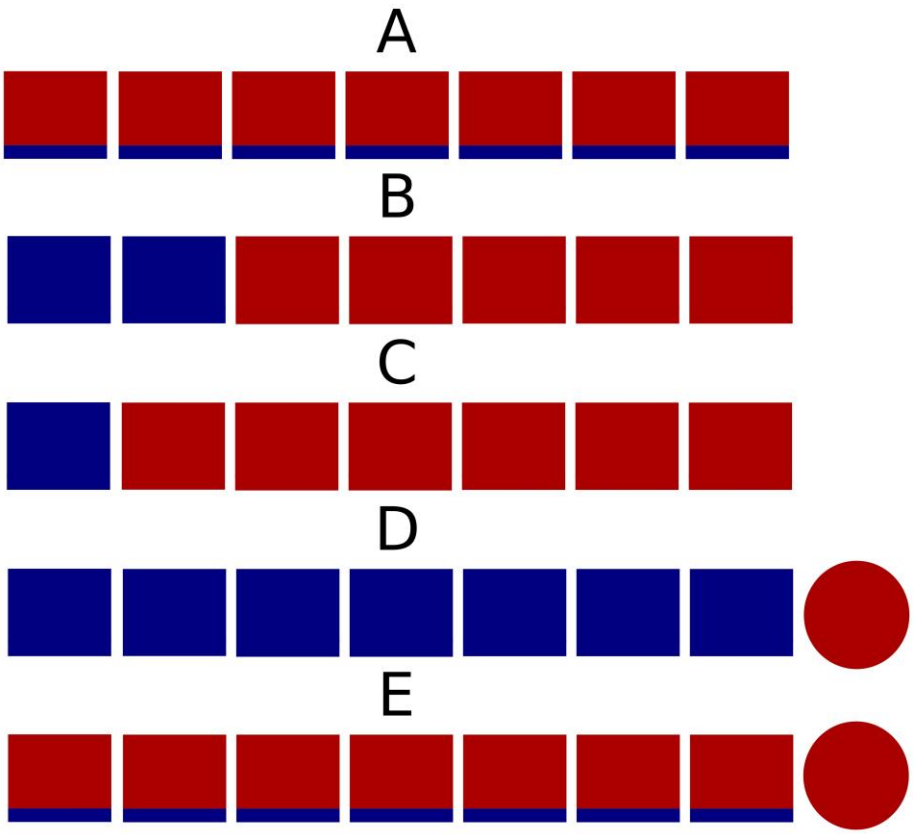

Figure 1. Scheme of the TS scenarios evaluated. Squares represent $F_{2:4}$ populations and circles represent panels of elite lines. Red and blue areas represent the TS and VS, respectively. Because some lines are involved in more than one cross, B shows two families as the VS.

In order to fit the prediction model, the G-BLUP approach was used by ASReml-R (Butler et al. 2009) with the following equation:

$$y = \mu\mathbf{1} + Z_1\boldsymbol{g} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is the vector of best liner unbiased estimator (BLUE); $\boldsymbol{\mu}$ is the general mean; $\boldsymbol{g}$ is the vector of the estimated genomic breeding values, considered as random, where $\boldsymbol{g} \sim N(\boldsymbol{0}, \boldsymbol{G})$; $\boldsymbol{\mathcal{E}}$ is the residual, where $\boldsymbol{\mathcal{E}} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$ where $\boldsymbol{I}$ is an identity matrix and σ^2 is the residual variance. $\boldsymbol{Z_1}$ is the incidence of matrixes that relate the effects of independent vectors to the dependent variable, **y**. The genomic relationship matrix is $\boldsymbol{G} = (\boldsymbol{XX'/p})$ (VanRaden 2007), where $\boldsymbol{X}$ is the matrix of markers and p is the number of markers. One model was performed for each trait.

Additionally, to improve PEL's prediction performance, we used an algorithm for the optimization of TS (OTS) to select a subset of PEL that was related more closely with the VS.

The algorithm was the Selection of Training Populations by Genetics algorithm, R-package STPGA (Akdemir et al. 2015), with predefined population sizes (80%, 60%, 40%, and 20% of total lines in the PEL). The algorithm uses principal components from the marker matrix to estimate the prediction error variance (PEV). Subsequently, the best lines were selected for TS based in the predefined population size.

We compared the five scenarios of TS for both traits (GY and MG) on the basis of their prediction ability (Pearson correlations of the prediction values and real performance of the VS); response to selection (prediction ability divided by time demand of each strategies, with fixed values for selection intensity and additive variance); and the selection coincidence (SC), using 10%, 25%, and 50% as the percentages of selection to generate the improved populations.

3.3. Results

3.3.1. Population structure

In order to check the relationship between PEL and the seven bi-parental populations in M2 and M4, a PCA analysis was performed. As expected, each bi-parental cross was grouped together. However, mixed areas were observed due to the coincidental parents in some crosses, especially in M2. The less coincidental occurrence of parents in M4 resulted in a more precise separation of families. In both cases, PEL was grouped among the progenies, which is an excellent indicator of the proper performance of those for predictions. This closeness was less evident in M2 than M4, in which three of the seven populations showed more substantial distances (Figures 2 and 3).

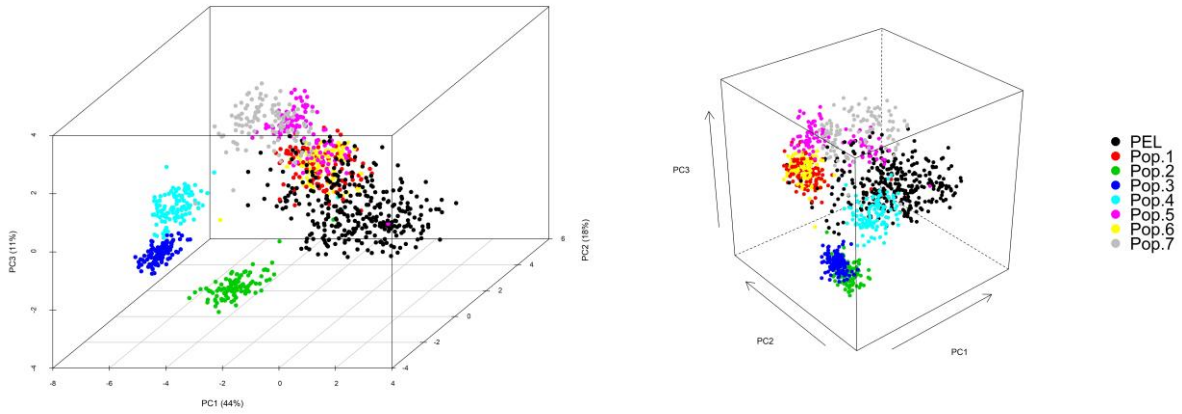

Figure 2. PCA analysis of M2 populations and their PEL.

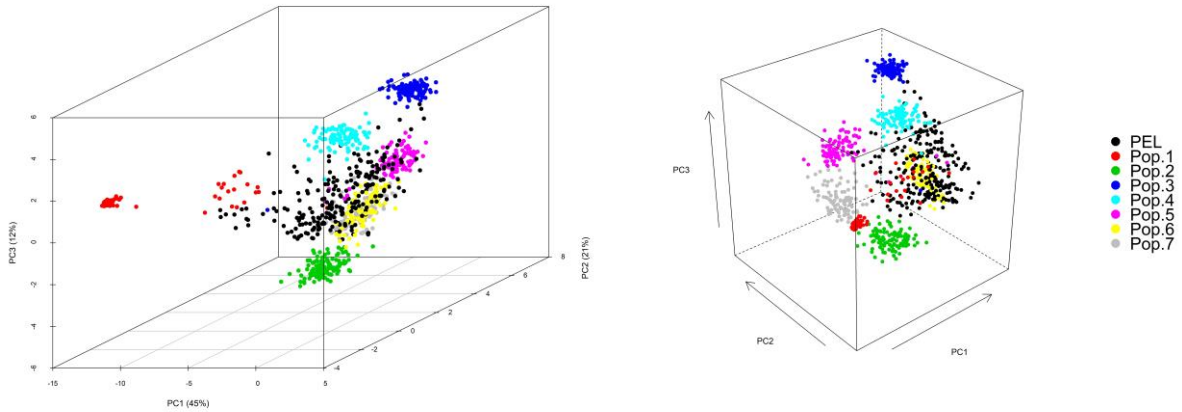

Figure 3. PCA analysis of M4 populations and their PEL.

3.3.2. Prediction abilities and optimization of training sets

Among all the scenarios we tested, we observed a range in PA from 0.09 to 0.51 for GY and 0.01 to 0.53 for MG. For GY, the PA was very similar between M2 and M4. In both datasets, scenario E obtained the highest performance, followed by A and D. Scenarios B and C resulted in a weak PA in both datasets. When we only considered the better scenarios, M4 was better than M2 on A and D; and similar in E. For MG, the results were very different compared to GY. For this trait, scenario A was the best and no other scenario had a high value for PA. Scenario E was second after scenario A and achieved almost half of the PA shown in scenario A. Also, scenario D, which was a promising scenario for GY, was too weak in MG. Furthermore, B and C were now better than D, but only in M2. Likewise, GY, M2, and M4 show similar results, with little superior performance in M2 (Figure 4 A).

Considering that a normal soybean breeding timeline is five years, the strategies A and E demand two additional years, hence a previous evaluation of full sibs is necessary. This way, the response to selection in these scenarios are penalized, what increase the advantage of scenario D for GY in both datasets. For MG, nothing chance, once low performance of all scenarios, except for A (Figure 4 B).

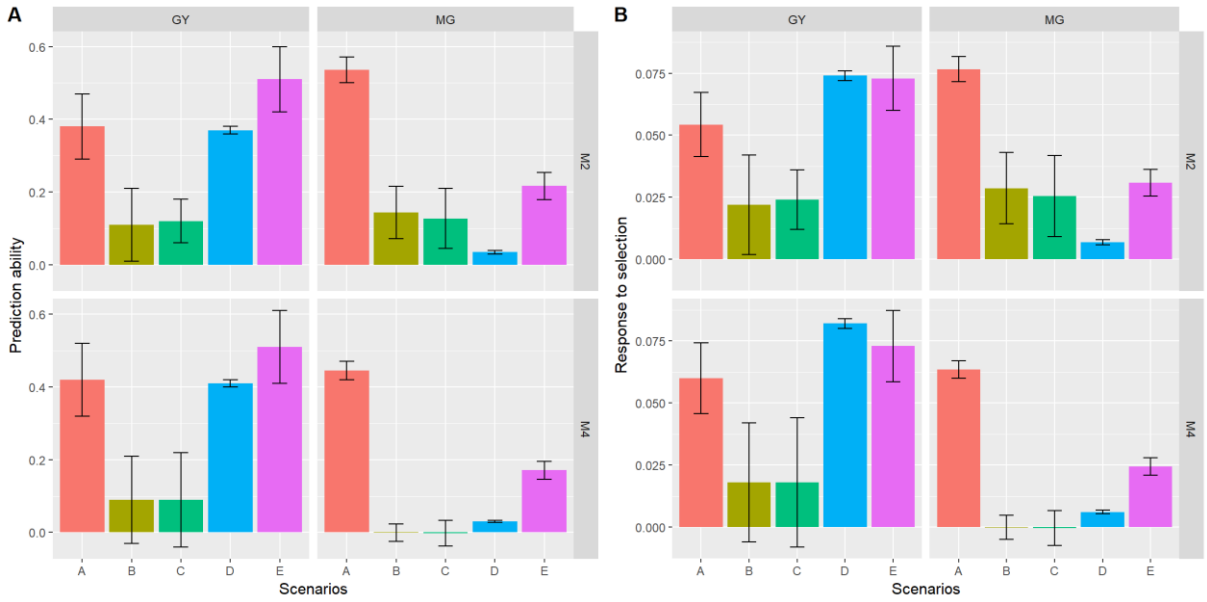

Figure 4. Prediction ability and response to selection obtained for each scenario in GY and MG for datasets M2 and M4.

As in PA, results of the optimization of the TS were different for GY and MG. For GY, the OTS does not improve the PA, but M2 and M4 showed different results. While in M2, the OTS and random sample performed similar from 100% to 20%; in M4, the PA started to decay in 60%, but only in the random scenario. As in M2, the PA remained stable using OTS until 20%. Despite some small variations, there were no differences in the use of all lines from PEL or any percentage of OTS in both datasets for GY (Figure 5). For MG, the results were weak for both (optimized and random). The highest PA obtained was 0.08 for M2 (40%) and 0.25 for M4 (60%). Both are better than the use of all PELs, showing that is it possible to improve PA for MG by selecting some specific lines from the dataset; however, those results were too far removed from scenario A, which was the scenario that produced the best results for MG (Figures 4 and 5).

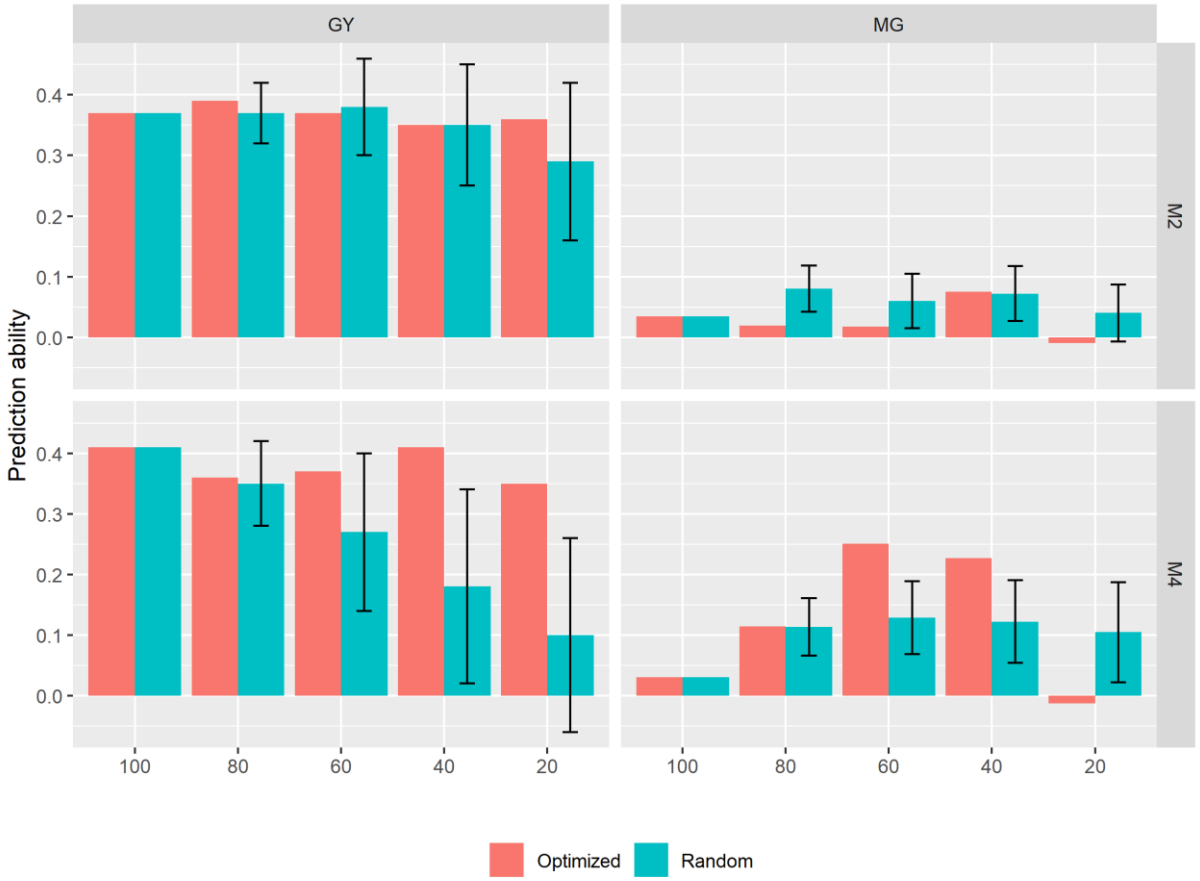

Figure 5. Comparison of PA using the optimized and random PEL reduction as TS for GY and MG in datasets M2 and M4.

3.3.3. Selection coincidences

The results for SC were highly correlated with the PA obtained for both traits, in which a higher PA value correlated with a better SC. A more successful selection than a random choice implies that the SC must be higher than the percentage of selection used in the population. This phenomenon indicates an increase in the percentage of desirable individuals in the selected population compared to the original population. This condition was observed in scenarios A, D, and E for GY in both datasets. Scenarios B and C reached values near the percentage of selection for SC, showing little advantage for GS in these scenarios. For MG, scenarios A, B, C, and E were satisfactory for M2; however, scenarios A and E were the best for M4. For both datasets, scenario A was the most successful and had the greatest PA. The high performance of GS compared to random choice was continuously verified in most of the percentiles of selection testing. Hence, considering the successful scenarios for GS, the three

percentiles reached values above SC. This did not occur only for MG and in M4 where the 10% does not achieve good coincidences of selection (Figure 6).

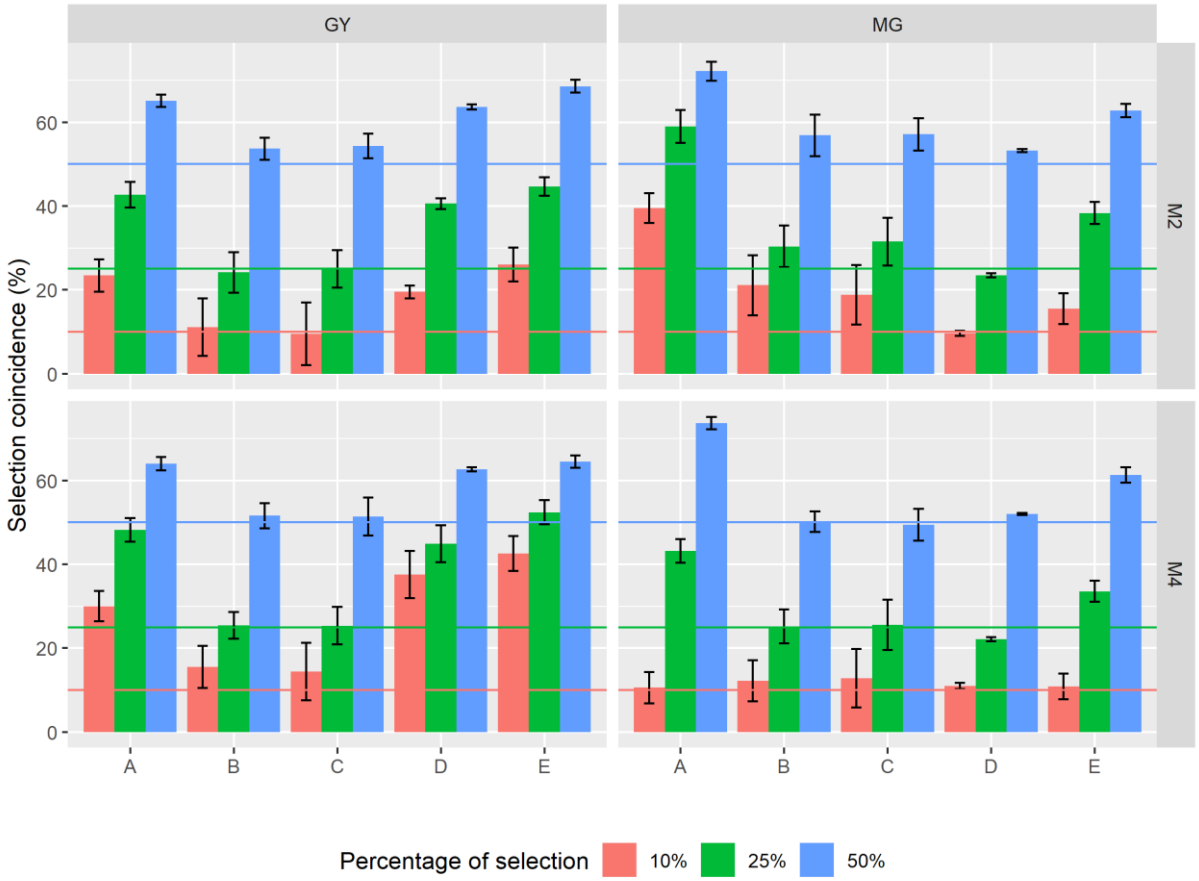

Figure 6. Selection coincidences obtained for GY and MG in datasets M2 and M4, considering 10%, (coral); 25% (green); and 50% (blue) as the percentage of selection. Horizontal lines of the same colors indicate the respective percentiles; these give an idea of which scenarios select better than random choice.

3.4. Discussion

Soybean breeders attempt to predict new recombinant lines. Therefore, every year they perform new crosses and test the progenies in the field. However, in the beginning (e.g., F_2), there are many progenies to be tested. This problem makes it economically unviable to explore all the variability generated by crosses. Nevertheless, because genotype prices have fallen and phenotype costs have risen (Cobb et al. 2013), the GS approach could increase the amount of variability that can be explored for the same economic investment.

Yield and maturity are two of the most important traits in soybean, hence, the simultaneous selection for those is essential. In some macroregions of Brazil (e.g. M2)

precocity allows farmers to perform more than one crop in the same year. This combined with high yield, characterize the most successful varieties in this Zone. Therefore, a multi-trait approach could be an option, but the correlation of those traits was close to zero in both datasets. In this situation, nonadditional gains are verified using multi-trait genomic prediction (Lyra et al. 2017), hence, two single-trait models can be used.

Because TS development is the main cost for GS (Rajsic et al. 2016), the strategy used to build it needs to take genetic and economic factors into account. In this context, strategy A—which obtained the first and second highest values of PA for MG and GY, respectively—adheres well to the contributing genetic factors once the TS comprises progenies of the same crosses. Scenario A is the perfect scenario for performing predictions because the TS and breeding populations are closely related. On the other hand, breeders are interested in new recombinants. Thus, if they also generate those for TS, there is less interest in predicting progenies that comes from these crosses. Moreover, addition time is need, what increase the time of breeding and reduces the gains per time unit (Figure 4 A and B).

Scenarios B and C were performed to verify the possibility of predicting progenies from crosses that were not included in the TS. Nevertheless, the PA results were low for both traits, indicating a limited use for a segregated population at this point. Similar results were verified by Schopp et al. (2017) using *in silico* bi-parental crosses in maize. These authors showed that the use of unrelated crosses in the TS reduces the PA by an average of 40–60%. The same trend was observed by Riedelsheimer et al. (2013) and Würschum et al. (2017) using real populations of maize and barley, respectively.

The low PA performance in scenarios B and C can be justified in two ways: effective size and genetic variability. Although these populations have a high number of progenies (~800), there are few founder lines; so, the external population representation is limited, and this leads to a low PA. Additionally, even though there is a high amount of phenotypic variance within bi-parental crosses, the genetic variability is low (data not shown). This creates a problematic prediction scenario when only one family is considered in the VS. Thus, using just one or two families as the VS will naturally lead to a low PA, and this scenario might not adequately represent the performance of a group of unrelated families predicting another group. To minimize this problem, we tested the ability of the seven bi-parental populations of M2 to predict the other seven from M4, and vice-versa. The PA was close to zero in both cases, and this is explained by the genetic distance between them (Figure 7) since they came from different breeding zones. These findings indicate the inability to predict the results of bi-parental crosses using other families as the TS.

Consequently, we performed scenarios D and E, which used information from the PEL to increase the genetic variability of the TS. For GY, the PA was satisfactory in both cases, matching and surpassing the PA obtained by scenario A. Although E showed the greatest performance in perdition ability, it was equivalent with strategy D in response to selection, hence D only uses PEL information, therefore, no addition time is needed just to generate the TS (Figure 4 B). The scenario E has the same problem as A, on what part of the lines at the same crosses was represented in the TS. Then, it requires an evaluation from part of the family, increasing the costs and time.

Scenarios D and E were not satisfactory for MG, indicating that the inclusion of all PELs without any selection process was harmful for prediction. In this case, the simple inclusion of variability in the TS does not increase accuracy. This finding suggests that predictions demand less diversity but more trait segregation resolution, similar to the findings in scenario A. In fact, the primary genetic selection strategy for MG is marker-assisted selection (MAS) because of the oligolectic genetic control of this trait (Zhang et al. 2015; Contreras-Soto et al. 2017b; Langewisch et al. 2017); for the genomic selection approach, MG is usually used as a division criterion in TS building (Smallwood et al. 2019). However, there is a lack of information in the literature about primary genes controlling MG up to group V (Langewisch et al., 2014; Jiang et al., 2014; Wolfgang and An, 2017; Li et al., 2017). Therefore, for the tropical conditions, the GS approach could be an alternative that could be used to assist breeding programs.

As expected, the performance of SC coincides with the results of the PA. Despite a sizeable initial population size, a non-restricted selection is expected in the early steps of the program, e.g., 50%. In this context, the results show superior performance in this selection percentage in scenarios A, D, and E for GY, and in A and E for MG. A more intensive selection percentage can be explored by accounting for the variability in advanced steps and initial population size. Nowadays, the cost of genotyping restrict the application of GS in the early phase, but strategies that reduce the set of markers for genotyping (Pembleton et al. 2016; Gorjanc et al. 2017; e Sousa et al. 2019) or use the same criteria to preselect lines for the TS (Fristche-Neto et al. 2018; Akdemir and Isidro-Sánchez 2019) can make GS application possible.

Therefore, with an optimal strategy to select lines for a TS, given a specific prediction set, it would be possible to keep the accuracies at the same level of using the whole dataset while reducing the costs in genotyping, phenotyping, and computing resources. As we have shown, the breeding populations used have a considerable population structure (Figures

2 and 3). In this situation, the use of optimization algorithms for TS is suggested. Rincent et al. (2012) reported improvement using principal component analysis on two diversity panels for predictions in maize. Furthermore, the incorrect inclusion of some groups of lines or progenies in the TS can reduce the PA (Bernardo and Yu 2007; Isidro et al. 2015; Lado et al. 2016). Therefore, in the context of a large diverse panel and external segregated progenies, the algorithm that optimizes the TS can be used to take out genotypes that will not contribute to predictions (Akdemir et al. 2015).

The optimization showed two exciting results for GY: in the case of our datasets, it is not possible to increase—but it is possible to keep—the PA by reducing the number of lines on TS; also, M2 and M4 demonstrated different performance. The first result indicates there are not distance groups of lines distorting the predictions; this was confirmed for PCA analysis (Figures 2 and 3), in which all PELs are taken from segregated populations. The second shows that M2 was capable of keeping its PA in a more drastic reduction, even using the random scenario, showing that M2 has a less complex genomic population structure. This result can be correlated with the MG of those datasets and the relationship between MG and population structure. M2 comes from a small breeding zone, in which the MG used is mainly between V and VI; the M4 zone is large, and groups range from VI to VIII (Zdziarski et al. 2018). In other words, due to latitude conditions, M4 is a more diverse zone compared to M2. It can be seen in the joint PCA graphic, where M2 progenies have a narrow distribution (Figure 7). Hence, it is expected that fewer lines are capable of representing M2 well.

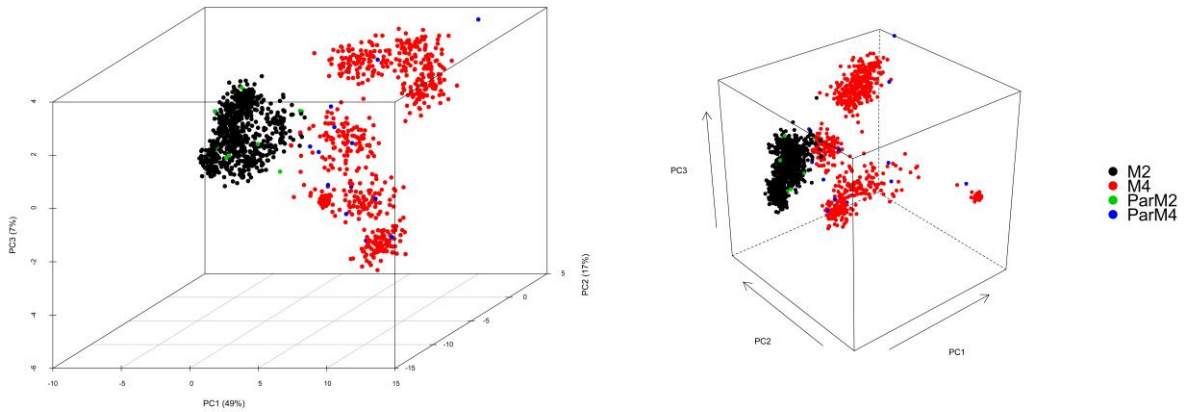

Figure 7. PCA graphic of M2 and M4 progenies and their parents

The performance of OTS in MG was not as high as in GY. In would be possible to see an improvement in M4, but not significant enough to achieve the same PA obtained for

the best scenario (A). However, a tendency toward improvement could be observed with the reduction of the panel followed by a long decay caused by the small number of lines in the more reduced scenario. This indicates that even MG can benefit from the OTS, but the most important factor is to keep the TS relationship with the predictions set.

The use of a panel of lines as TS for predictions is extensively discussed in the literature, possibly because the PA potential of these could be stable for years. He et al. (2016), for example, used more than 2000 European winter wheat lines and obtained a PA superior to 0.50 for GY, and the authors assessed a selection of subsets. Meanwhile, Jarquin, Specht, and Lorenz (2016) used robust historical data from the USDA Soybean Germplasm Collection and obtained almost 0.8 of PA. In general, our PA was lower than those reported; this is explained by the low heterozygosity levels of genotypes tested in previous articles. Hence, the adjustment of an additive model for prediction is more parsimonious in these cases. Because we are combining lines and segregating the population to perform the early selection, a lower PA is expected. Moreover, in the initial steps, with a high number of progenies, a slightly reduced PA is not reckless because the phenotypic selection in these steps is usually unsuccessful (Resende et al. 2017).

Even the scenario D showed the best performance in response to selection, the idea behind scenario E cannot be ignored. Sometimes, for many reasons, breeding programs repeat some crosses. Thus, those progenies can be incused on TS to increase prediction accuracy. Moreover, the same parent can be used in crosses for years, which creates the possibility of half-sibs in the PEL after multiple breeding cycles. In this context, Xavier, Muir, and Rainey (2016) reported that a significant increment of PA increases the size of the TS with related progenies—up to 2000 individuals. However, results can vary at this point. For instance, Neyhart et al. (2017) reported that the best performance occurred when they used a smaller but more recent TS. In this context, breeding programs, which generally accumulate lots of phenotypic data every year, have an excellent opportunity to accumulate genotype data as well, which can be used to improve their TS.

In conclusion, the use of PELs as TS (scenario D) is recommended for GY when applying GS in earlier steps of soybean breeding programs. Regarding genetic factors, the strategy mentioned previously uses a group of lines that is continuously renewed and genetically close to the breeding populations. Moreover, no additional phenotyping—e.g., building an experiment only for the TS—is needed. Furthermore, no further generations are needed to build the TS, a fact that saves time and resources. For MG, the predictions are successful in scenarios with great variability control and when the TS is highly correlated. All

SCs attained satisfactory results. Hence, the choice depends on breeding parameters, such as the size of the populations, scales of the experiments, and budget for phenotypes and genotypes. Finally, the optimization of the PEL did not improve the PA for either the datasets or traits. However, for the GY, it kept the PA at almost the same level for the whole scenario, indicating a potential use for saving resources.

REFERENCES

Akdemir D, Isidro-Sánchez J (2019) Design of training populations for selective phenotyping in genomic prediction. Sci Rep 9:1446. doi: 10.1038/s41598-018-38081-6

Akdemir D, Sanchez JI, Jannink J-L (2015) Optimization of genomic selection training populations with a genetic algorithm. Genet Sel Evol 47:38. doi: 10.1186/s12711-015-0116-6

Alliprandini LF, Abatti C, Bertagnolli PF, et al (2009) Understanding Soybean Maturity Groups in Brazil: Environment, Cultivar Classification, and Stability. Crop Sci 49:801. doi: 10.2135/cropsci2008.07.0390

Bernardo R, Yu J (2007) Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci 47:1082. doi: 10.2135/cropsci2006.11.0690

Bradbury PJ, Zhang Z, Kroon DE, et al (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. doi: 10.1093/bioinformatics/btm308

Browning SR, Browning BL (2007) Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet 81:1084–1097. doi: 10.1086/521987

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml-R reference manual mixed models for S language environments. Train. Ser. QE02001 149

Cobb JN, Declerck G, Greenberg A, et al (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theor Appl Genet 126:867–87. doi: 10.1007/s00122-013-2066-0

Contreras-Soto RI, Mora F, Lazzari F, et al (2017) Genome-wide association mapping for flowering and maturity in tropical soybean: implications for breeding strategies. Breed Sci 67:435–449. doi: 10.1270/jsbbs.17024

Costa MM, Di Mauro AO, Unêda-Trevisoli SH, et al (2008) Heritability estimation in early generations of two-way crosses in soybean. Bragantia 67:101–108. doi: 10.1590/S0006-87052008000100012

e Sousa MB, Galli G, Lyra DH, et al (2019) Increasing accuracy and reducing costs of genomic prediction by marker selection. Euphytica 215:18. doi: 10.1007/s10681-019-2339-z

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

Gorjanc G, Dumasy J-F, Gonen S, et al (2017) Potential of Low-Coverage Genotyping-by-Sequencing and Imputation for Cost-Effective Genomic Selection in Biparental Segregating Populations. Crop Sci 57:1404. doi: 10.2135/cropsci2016.08.0675

Hartwig EE (1973) Varietal development. In: Caldwel BE (ed) Soybeans: Improvement, production, and uses, 1st edn. ASA, Madison, WI, pp 187–207

Hayes BJ, Pryce J, Chamberlain AJ, et al (2010) Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. PLoS Genet 6:e1001139. doi: 10.1371/journal.pgen.1001139

He S, Schulthess AW, Mirdita V, et al (2016) Genomic selection in a commercial winter wheat population. Theor Appl Genet 129:641–651. doi: 10.1007/s00122-015-2655-1

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic Selection for Crop Improvement. Crop Sci 49:1. doi: 10.2135/cropsci2008.08.0512

Isidro J, Jannink J-L, Akdemir D, et al (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128:145–158. doi: 10.1007/s00122-014-2418-4

Jarquin D, Specht J, Lorenz A (2016) Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions. G3: Genes|Genomes|Genetics 6:2329–2341. doi: 10.1534/g3.116.031443

Jiang B, Nan H, Gao Y, et al (2014) Allelic Combinations of Soybean Maturity Loci E1, E2, E3 and E4 Result in Diversity of Maturity and Adaptation to Different Latitudes. PLoS One 9:e106042. doi: 10.1371/journal.pone.0106042

Lado B, Barrios PG, Quincke M, et al (2016) Modeling Genotype × Environment Interaction for Genomic Selection with Unbalanced Data from a Wheat Breeding Program. Crop Sci 56:2165. doi: 10.2135/cropsci2015.04.0207

Lam H-M, Xu X, Liu X, et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059. doi: 10.1038/ng.715

Langewisch T, Lenis J, Jiang G-L, et al (2017) The development and use of a molecular model for soybean maturity groups. BMC Plant Biol 17:91. doi: 10.1186/s12870-017-1040-4

Langewisch T, Zhang H, Vincent R, et al (2014) Major Soybean Maturity Gene Haplotypes Revealed by SNPViz Analysis of 72 Sequenced Soybean Genomes. PLoS One 9:e94150. doi: 10.1371/journal.pone.0094150

Li J, Wang X, Song W, et al (2017) Genetic variation of maturity groups and four E genes in the Chinese soybean mini core collection. PLoS One 12:e0172106. doi: 10.1371/journal.pone.0172106

Lyra DH, de Freitas Mendonça L, Galli G, et al (2017) Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. Mol Breed 37:80. doi: 10.1007/s11032-017-0681-1

Matei G, Woyann LG, Milioli AS, et al (2018) Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. Mol Breed 38:117. doi: 10.1007/s11032-018-0872-4

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29

Mourtzinis S, Conley SP (2017) Delineating Soybean Maturity Groups across the United States. Agron J 109:1397. doi: 10.2134/agronj2016.10.0581

Neyhart JL, Tiede T, Lorenz AJ, Smith KP (2017) Evaluating Methods of Updating Training Data in Long-Term Genomewide Selection. G3: Genes|Genomes|Genetics 7:1499–1510. doi: 10.1534/g3.117.040550

Pembleton LW, Drayton MC, Bain M, et al (2016) Targeted genotyping-by-sequencing permits cost-effective identification and discrimination of pasture grass species and cultivars. Theor Appl Genet 129:991–1005. doi: 10.1007/s00122-016-2678-2

Rajsic P, Weersink A, Navabi A, Peter Pauls K (2016) Economics of genomic selection: the role of prediction accuracy and relative genotyping costs. Euphytica 210:259–276. doi: 10.1007/s10681-016-1716-0

Resende RT, Resende MD V, Silva FF, et al (2017) Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. Heredity (Edinb). 119:245–255

Riedelsheimer C, Endelman JB, Stange M, et al (2013) Genomic Predictability of Interconnected Biparental Maize Populations. Genetics 194:493–503. doi: 10.1534/genetics.113.150227

Rincent R, Laloë D, Nicolas S, et al (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics 192:715–728. doi: 10.1534/genetics.112.141473

Schopp P, Müller D, Wientjes YCJ, Melchinger AE (2017) Genomic Prediction Within and Across Biparental Families: Means and Variances of Prediction Accuracy and Usefulness of Deterministic Equations. G3 (Bethesda) 7:3571–3586. doi: 10.1534/g3.117.300076

Smallwood CJ, Saxton AM, Gillman JD, et al (2019) Context-Specific Genomic Selection Strategies Outperform Phenotypic Selection for Soybean Quantitative Traits in the Progeny Row Stage. Crop Sci 59:54. doi: 10.2135/cropsci2018.03.0197

VanRaden PM (2007) Genomic Measures of Relationship and Inbreeding. Interbull Bull 111–114. doi: 10.1007/s13398-014-0173-7.2

Wimmer V, Albrecht T, Auinger H-J, Schön C-C (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086–7. doi: 10.1093/bioinformatics/bts335

Wolfgang G, An YC (2017) Genetic separation of southern and northern soybean breeding programs in North America and their associated allelic variation at four maturity loci. Mol Breed 37:8. doi: 10.1007/s11032-016-0611-7

Würschum T, Maurer HP, Weissmann S, et al (2017) Accuracy of within- and among-family genomic prediction in triticale. Plant Breed 136:230–236. doi: 10.1111/pbr.12465

Xavier A, Muir WM, Rainey KM (2016) Assessing Predictive Properties of Genome-Wide Selection in Soybeans. G3 (Bethesda) 6:2611–6. doi: 10.1534/g3.116.032268

Zdziarski AD, Todeschini MH, Milioli AS, et al (2018) Key Soybean Maturity Groups to Increase Grain Yield in Brazil. Crop Sci 58:1155. doi: 10.2135/cropsci2017.09.0581

Zhang J, Song Q, Cregan PB, et al (2015) Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. BMC Genomics 16:217. doi: 10.1186/s12864-015-1441-4