Received: 27 October 2018 | Revised: 29 March 2020 | Accepted: 4 April 2020

DOI: 10.1111/pbr.12827

REVIEW ARTICLE

Plant Breeding WILEY

Using public databases for genomic prediction of tropical maize lines

Pedro Patric Pinho Morais[1] | Deniz Akdemir[2] | Luciano Rogério Braatz de Andrade[1] | Jean-Luc Jannink[3] | Roberto Fritsche-Neto[4] | Aluízio Borém[1] | Filipe Couto Alves[4] | Danilo Hottis Lyra[4] | Ítalo Stefanine Correia Granato[4]

[1]Department of Crop Sciences, Federal University of Viçosa, Viçosa, Brazil

[2]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan

[3]Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York

[4]Department of Genetics, Luiz de Queiroz College Agriculture, University of São Paulo, Piracicaba, Brazil

Correspondence
Pedro Patric Pinho Morais, Nações Unidas Avenue, nº 12.901, 7th and 8th Floors. Monsanto do Brasil LTDA, São Paulo, 04578-910 São Paulo, Brazil.
Email: patric_pinho@hotmail.com

Funding information
Fundação de Amparo à Pesquisa do Estado de São Paulo; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

Communicated by: T. Lübberstedt

Abstract

In this paper, the aims were (a) to test the usefulness of using genomic and phenotypic information from public databases (open access) to predict genetic values for tropical maize inbred lines regarding plant and ear height; (b) to identify how the population structure, the use of optimized training sets (OTSs) and the amount of information originating from public databases affect the predictive ability. Thus, 29 training sets (TSs) were defined considering three diversity panels: the University of São Paulo (USP—validation set (VS)) and the ASSO and USDA North Central Regional Plant Introduction Station (NCRPIS) (external public panels—predictors), which were divided into four scenarios with different TS configurations. We showed that it is possible to use public datasets as a primary TS and that population structure can modify the predictive abilities of GS. In the four scenarios proposed, very large or very small sets did not provide predictive abilities over 0.53 for GS. However, OTSs composed of 250 individuals were sufficient to achieve predictive abilities over this limit.

KEYWORDS

diversity panels, GBLUP, optimized training set, population structure, predictive ability

1 | INTRODUCTION

Proposed at the beginning of the twenty-first century by Meuwissen, Hayes, and Goddard (2001), genomic selection (GS) arose as a broad methodology for using information from markers spread over the entire genome to predict the performance of genotypes. With the advent of genomic selection, it is possible to increase selection gains per unit of time and reduce costs in breeding programmes (Meuwissen et al., 2001;Wong & Bernardo, 2008). Two main phases are required to apply GS: training and selection (Jonas & de Koning, 2013). In the former, two population sets are used, the training set (TS) and the validation set (VS), in which the individuals are phenotyped and genotyped. In the TS, the effects of the markers are estimated through prediction models, whereas in the VS, these effects are tested to verify the accuracies of the prediction models. In the former phase, the genomic estimated breeding values (GEBVs) of the genotypes are predicted from the marker effects estimated based on the TS (Bernardo, 2014; Newell & Jannink, 2014).

Therefore, the composition and size of the training sets are critical to obtaining high accuracies of prediction (Isidro et al., 2015).

Abbreviations: ASSO, Maize 282 Association Panel; BLUPs, best linear unbiased predictions; EH, ear height; GBS, genotyping by sequencing; GEBVs, estimated breeding values; GS, genomic selection; h^2, heritability; LRT, likelihood ratio test; NAM, nested association mapping population (inbred lines); NCRPIS, USDA North Central Regional Plant Introduction Station maize collection (inbred lines); OTS, optimized training set; PC, principal components; PEV, prediction error variance; PH, plant height; RTS, randomized training set; SNP, single nucleotide polymorphism; TS, training set; TSG1, training set group 1; TSG2, training set group 2; TSG3, training set group 3; TSG4, training set group 4; USP panel, São Paulo University panel (inbred lines); VS, validation set.

© 2020 Blackwell Verlag GmbH

Thus, it is necessary to include an adequate number of individuals in the TS, making the GS feasible compared to phenotypic selection. Consequently, in breeding programmes with small TS size, the incorporation of genomic and phenotypic data from public datasets seems to be a straightforward solution. Hence, public and private small-scale breeding programmes with limited budgets to establish a base population for GS would be profitable. Several panels of maize diversity, representing a broad sample of inbred lines and heterotic groups, are available, for instance, the nested association mapping (NAM) population (Hung et al., 2012;McMullen et al., 2009), the Maize 282 Association Panel (Flint-Garcia et al., 2005), a collection of the USDA North Central Regional Plant Introduction Station (NCRPIS) (Romay et al., 2013) and the MAGIC population (Dell'Acqua et al., 2015).

Another critical factor affecting the accuracy, or predictive ability, in this case, is the genetic relationship between TS and VS (Albrecht et al., 2011; Clark, Hickey, & Werf, 2011; Habier, Fernando, & Dekkers, 2007; Hayes, Visscher, & Goddard, 2009; Pszczola, Strabel, Mulder, & Calus, 2012). Recently, the development of algorithms to assist in choosing individuals for the establishment of optimized training sets (OTSs) has drawn the attention of animal and plant breeders (Rincent et al., 2012). The reason for this is that an increase in the predictive ability of the models tested is a crucial point for efficiently allocating the resources of a breeding programme and having the best balance between genetic variability, cost and accuracy. This procedure has had significant outcomes in the form of increased accuracy for GS in recent studies (Akdemir, Sanchez, & Jannink, 2015; Isidro et al., 2015). Therefore, the aims of this study were (a) to test the usefulness of using genomic and phenotypic information from public databases to predict tropical maize inbred lines and (b) to identify how the population structure, the use of optimized training sets (OTSs) and the amount of information originating from public databases affect the predictive ability of tropical maize inbred lines.

2 | MATERIALS AND METHODS

2.1 | Genotype group and field experiments

We used three panels composed of maize inbred lines: (a) São Paulo University (USP) validation set (VS), (b) the nested association mapping (NAM) population (Hung et al., 2012; McMullen et al., 2009) and the Maize 282 Association Panel (Flint-Garcia et al., 2005) (the combination of both is referred to as the associative panel, ASSO) and (c) Department of Agriculture—Agricultural Research Service (USDA–ARS), North Central Regional Plant Introduction Station (NCRPIS) (Romay et al., 2013).

The USP dataset was composed of 64 tropical lines. The experimental trial used a simplex lattice design (8 × 8) with two replications. Trials were carried out in Anhembi (22°50′51″S, 48°01′06″W, 466 m) and Piracicaba (22°42′23″S, 47°38′14″W, 535 m), São Paulo State, Brazil, during the second growing season of 2014 and 2015. The plots consisted of a 4-m-long row (2014) and a 5-m-long row (2015), with a spacing of 0.85 m between rows and 0.20 m between plants, under conventional fertilization, weed and pest control.

The NAM dataset and Maize 282 Association Panel were evaluated in ten environments in the United States (US) under a conventional tillage system. The environments were Aurora, NY; Clayton, NC; Columbia, MO; and Urbana, IL in the 2006 and 2007 crop years; Aurora, NY, in the 2008 crop year; and Columbia, MO, in the 2009 crop year.

The NCRPIS dataset was evaluated in 2010 in plots of a single row in three environments in the US: Aurora, NY; Clayton, NC; and Columbia, MO. In Aurora, the rows were planted in plots of 12 plants in the Muskgrave Research Station. In Clayton, the experiment was set up at the Central Crops Research Station, and in Columbia, experiments were conducted at the South Farm, with 15 plants per plot. In these experiments, the NCRPIS set was stratified into nine maturity groups, and the lines were randomly designed in incomplete blocks of 19 lines, with the lines B73, IL14H, KI11, P39, SA24 and TX303 as controls.

2.2 | Phenotypic traits

In the three panels, the traits evaluated were plant height (PH, cm) and ear height (EH, cm). PH was measured from soil surface to the flag leaf collar and EH from soil to the primary node of the ear.

2.3 | Predicted genetic values by REML/ BLUP modelling

The best linear unbiased predictions (BLUPs) of NAM, Maize 282 and NCRPIS datasets were obtained by Peiffer et al. (2014). It means the BLUPs were accessed freely. However, for this study, we had to use a linear mixed model to calculate the BLUPs of USP inbred lines following the model below:

$$y = Jb + Zg + Vs + Tf + \varepsilon \tag{1}$$

where y is the line's adjusted mean of the traits evaluated; b is the block-effect vector within the replication, considered fixed; g is the genotype-effect vector (lines), considered random, in which $g \sim N(0, G)$ and G = $I\sigma_g^2$ and σ_g^2 is genetic variance; s is the environment-effect vector, considered fixed; f is the genotype × environment interaction-effect vector, considered random, in which $f \sim N$ (0, F) and $F = I\sigma_f^2$; and ε is the error vector, in which $\varepsilon \sim N(0, \mathcal{E})$ and $\mathcal{E} = I\sigma_\varepsilon^2$. Note that I is the identity matrix, and J, Z, V and T are incidence matrices that relate the effects of the independent vectors of each matrix to the dependent vector y.

Variance components and entry-mean-based heritability (h^2) were obtained for PH and EH for the USP dataset. The significance of the random effects of genotypes was assessed by the likelihood ratio test (LRT) at 5% probability using ASReml-R (Gilmour, Gogel, & Cullis, 2015).

2.4 | Genotypic data

The genotyping of the 64 tropical inbred lines was performed by an Affymetrix® Axiom® Maize Genotyping Array (Unterseer et al., 2014) containing approximately 614,000 SNPs. The genotyping information for the panels ASSO and NCRPIS were obtained using genotyping by sequencing (GBS) in ZeaGBS v2.7 (Glaubitz et al., 2014). This dataset, including the two panels, contained 17,280 thousand public lines, genotyped with 955,690 SNPs (available at www.panzea.org in a version partially imputed in file HDF5).

In the USP panel, markers with a low call rate (<90%) and minor allele frequency (<0.05) were removed, as were individuals with more than 10% heterozygous loci. A total of 409,000 high-quality polymorphic SNPs were obtained for the USP panel, and 359,000 SNPs and 12,149 inbred lines were obtained for the public dataset.

After this quality control, two pairing processes were carried out: (a) pairing of information from the set of markers obtained via GBS and Affymetrix, considering for this purpose the reference number of the chromosome and the exact physical position of the marker (bp) and (b) pairing of the set of lines genotyped via GBS with the lines of the dataset of Peiffer et al. (2014) (NCRPIS = 2,815, NAM = 4,982, Maize 282 = 282). Consequently, 28,260 SNPs were found to be common between the two sets of genotyping and 2,685 paired lines. Among these genotypes, 2,237 lines belonged to the NCRPIS Panel, and 448 belonged to ASSO (NAM = 166, Maize 282 = 282). Thus, when added to the 63 lines of the USP panel, the final dataset included 2,748 individuals genotyped for the same 28,260 SNPs. There was no common line or duplicate between the USP panel and other panels. These panels only had common SNP markers.

The entire process was carried out using TASSEL for Quality Control (Bradbury et al., 2007) and R software (R Core Team, 2017) for data pairing.

2.5 | Kinship and population structure

The genomic relationship matrix (GRM) was estimated using the VanRaden (2008) calculation methodology. Principal component analysis (PCA) was used to detect population structure, and following Romay et al. (2013), who used NCRPIS inbred lines, the number of groups was 5 ($K = 5$). For both analyses, the arrangement of 28,260 SNPs and 2,748 lines was considered. Thus, the clusters of tropical, popcorn, nonstiff stalk, stiff stalk and sweet corn were defined.

2.6 | Training set (TS) and validation set (VS)

A total of 29 training set groups (TSG) were defined, divided into four scenarios (Table 1).

TABLE 1 Groups of training sets (TS) and validation sets (VS) and their respective population size, N_t and N_v, used for the scenarios of training set groups TSG1, TSG2, TSG3 and TSG4

Group	TS	N_t	VS	N_v
TSG1	USP	10	USP	53
		20		43
		30		33
TSG2	NCRPIS + ASSO + USP	2,465	USP	32
	NCRPIS + ASSO	2,434		32
	NCRPIS	2,046		32
	ASSO	388		32
	$\text{NCRPIS}_{\text{(USP cluster)}}$	512		32
	$\text{ASSO}_{\text{(USP cluster)}}$	136		32
TSG3	OTS1 − RTS1	50	USP	32
	OTS2 − RTS2	250		32
	OTS3 − RTS3	500		32
	OTS4 − RTS4	1,000		32
	OTS5 − RTS5	1,500		32
TSG4	OTS6 − RTS6	81	USP	32
	OTS7 − RTS7	281		32
	OTS8 − RTS8	531		32
	OTS9 − RTS9	1,031		32
	OTS10 − RTS10	1,531		32

Abbreviations: ASSO, Association panel; NCRPIS, US USDA–ARS panel; N_t, number of inbred lines in the training; N_v, number of inbred lines in the validation set; OTS, optimized training set; RTS, randomized training set; USP Cluster, genotypes associated with USP dataset in PCA; USP, São Paulo University panel.

2.6.1 | TSG1—training set group 1—built only by private lines

We used all of the USP panel to establish this TS. This scenario sought to infer the predictive ability of the panel by itself, without any addition of outside information. Therefore, the TS was defined by three training groups ($k = 3$), each one considering different proportions of TS and VS within the USP panel. Hence, in the first, second and third groups, 10, 20 and 30 individuals were placed in the TS, respectively, and the VS had 53, 43 and 33 individuals to be predicted. The designation of the lines included in the TS was defined at random, later incorporating the remaining lines in the VS.

2.6.2 | TSG2—training set group 2—constituted by public database or public database and some private lines

The USP, NCRPIS and ASSO panels were used to establish the TS. However, the population structure information was used to define it. In this scenario and the others described below, a situation often found in plant breeding was considered, in which the environment used to phenotype the TS differs from the environments used for the VS. Additionally, the lines to be predicted had a weak relationship with the training group (Windhausen et al., 2012), just like the existence of a population structure. Therefore, three different TSs (scenarios) were considered: (a) individual panel (NCRPIS or ASSO); (b) combination of panels (NCRPIS + ASSO + USP or NCRPIS + ASSO); and (c) selection of lines belonging to the panels NCRPIS or ASSO that were allocated to the same cluster of the USP lines via PCA, in such a way that these selected lines had a greater genetic relationship with the USP panel (VS).

Specifically, in the second TS (NCRPIS + ASSO + USP), the VS was composed of 32 lines taken at random from the USP panel; the other lines (31) were then incorporated into the TS, together with the lines from the NCRPIS and ASSO panels.

2.6.3 | TSG3—training set group 3—formed by candidates selected by the optimized training set method or randomized training set from public databases

For the determination of the TS in this scenario, we applied the optimized training set (OTS) method proposed by Akdemir et al. (2015), considering a predefined population size. In this method, the selection of lines requires only genotypic information on the individuals present in a group of candidates (NCRPIS + ASSO) and in the group to be predicted (USP). Subsequently, based on this information, a genetic algorithm approximates the prediction error variance (PEV) with the principal components, via the marker matrix, and selects determined lines that will establish the OTS. In this scenario, five groups with different sizes of OTS (50, 250, 500, 1,000 and 1,500) were determined. The algorithm for the establishment of OTS was implemented via the STPGA package (Akdemir, 2017) using R software (R Core Team, 2017).

To make inferences and comparisons regarding the efficacy of the proposed method, we defined a second group determined at random, consisting of one of five randomized training sets (RTS), with the same sizes as the groups in the OTS. Thus, RTS with 50, 250, 500, 1,000 and 1,500 individuals were established via random sampling within the group of candidates, without the criterion of proportionality. The VS in TSG3 was composed of 32 lines taken at random from the USP panel.

2.6.4 | TSG4—training set group 4—developed by candidates selected by optimized training set method or randomized training set from a public database and some private lines

In this scenario, additional information from the USP panel was combined with the TSG3 to establish the TS. The inclusion of some USP lines in the TS attempted to simulate a condition in which the breeding programme phenotypes part of the lines of the total group (previously genotyped) and includes them with the external data for prediction of the other lines. Therefore, ten groups of different sizes of TS were established here, five via OTS and five via RTS (with 50, 250, 500, 1,000 and 1,500 individuals). The remaining lines of the USP panel (N = 31) that were not used in the TS were considered the VS.

2.7 | Genomic prediction models

In all the scenarios described above, the performance of the USP panel lines for PH and EH was predicted via G-BLUP, following the model:

$$y = Xu + Zg + \varepsilon \tag{2}$$

where y is the vector of genotypic values obtained via REML/BLUP, u is the fixed-effect vector (mean of the population), g is the random-effect vector of the genomic values, and Ɛ is the residual vector. The variance of the random effects of g is $var\ (g) = K\sigma_g^{\ 2}$, in which K is the genomic relationship matrix (kinship) and $\sigma_g^{\ 2}$ represents genetic variation. The residual variance is given by $var\ (Ɛ) = I\sigma_e^{\ 2}$. Lastly, Z is an incidence matrix that relates to the effects of the independent vector to the dependent vector y, X is an $n\ x\ 1$ column matrix, and I is the identity matrix. We used rrBLUP-R (Endelman, 2011) to run the model.

The predictive ability ($r_{(\hat{y},g)}$) was obtained by the Pearson correlation coefficient (r) between the predicted genetic value and the GEBV in the validation set. For each of the different scenarios tested, 50 replications were performed (preceded by loops). The accuracies obtained from the various scenarios were transformed into variables and subjected to analysis of variance to compare the effect of training set composition:

$$y_{ij} = \mu + T_i + \varepsilon_{ij} \tag{3}$$

$$y_{ij} = \mu + T_i + N(T_i) + \varepsilon_{ij} \tag{4}$$

in which $j = 1, \ldots, r$; $i = 1, \ldots, k$; y_{ij} is the j-th observation of the level i of the treatment factor; μ is the overall mean; T_i is the treatment effect (training set); N is the number of individuals within the TS; and $Ɛ_{ij}$ is the residual value. The model in Equation 3 refers to scenarios TSG1 and TSG2. Equation 4 designates scenarios TSG3 and TSG4. From these data, standard deviations (σ) were also estimated and the mean values compared by the Tukey test with α = 0.05. Before performing the analysis of variance, predictive abilities were assessed for normality assumption. All the procedures were carried out via R software.

3 | RESULTS

3.1 | Variance components and heritabilities

With the likelihood ratio test, we found significant differences between USP inbred lines for PH (126.84^{***}, chi-square test at 1%) and EH (78.74^{***}, chi-square test at 1%). The heritability estimates were considered moderate to high: ASSO panel (0.93 for PH and 0.94 for EH); NCRIPS (0.87 for PH and 0.86 for EH)—Heritabilities of ASSO and NCRIPS were obtained in Peiffer et al. (2014); and USP panel (0.71 for PH and 0.60 for EH).

3.2 | Population structure

Principal component analysis revealed a structure compatible with the accessed and paired data published by Romay et al. (2013). Thus, cluster analysis acted in such a way that all the groups could be separated in the first two axes of principal components (PC), showing estimates of 4.9% and 3.3% of the genetic variance, respectively (Figure 1). These results are consistent with previous studies on maize (Guo et al., 2013).

Considering the USP, ASSO and NCRPIS panels (Figure 1), the number of lines allocated in the clusters obtained was distributed as follows:

1. Tropical = USP (63), ASSO (136) and NCRPIS (512);
2. Popcorn = ASSO (60) and NCRPIS (91);
3. Nonstiff stalk = ASSO (207) and NCRPIS (1,272);
4. Stiff stalk = ASSO (15) and NCRPIS (150);
5. Sweet corn = ASSO (30) and NCRPIS (112).

As described above, the 151 lines belonging to the popcorn cluster did not factor into the establishment of the training sets in any proposed scenario.

3.3 | Methods of establishing the training set and prediction accuracies

3.3.1 | TSG1

In this first scenario, the prediction accuracies ranged from 0.16 to 0.20 for PH and from 0.02 to 0.05 for EH (Table 2). Therefore, although this panel exhibited genetic variability among the lines for the traits considered, the number of individuals included in the TS was a limiting factor regarding the predictive ability of GS.

3.3.2 | TSG2

For this scenario, where different panels were used to build the TS and VS, the predictive abilities obtained indicated that the addition of USP lines into the TS increased the predictive abilities of traits, though this increase in plant height was not significant (Table 3). Moreover, the panels NCRPIS and ASSO predicted USP panel with predictive abilities between 0.08 and 0.29 for PH and 0.09 and 0.44 for EH, regardless of whether they were combined.

3.3.3 | TSG3

The results obtained for TSG3 indicate that higher values of predictive ability were reached for plant height when optimized training sets (OTSs) were used with an N_t of 250, 500 or 1,000 individuals.

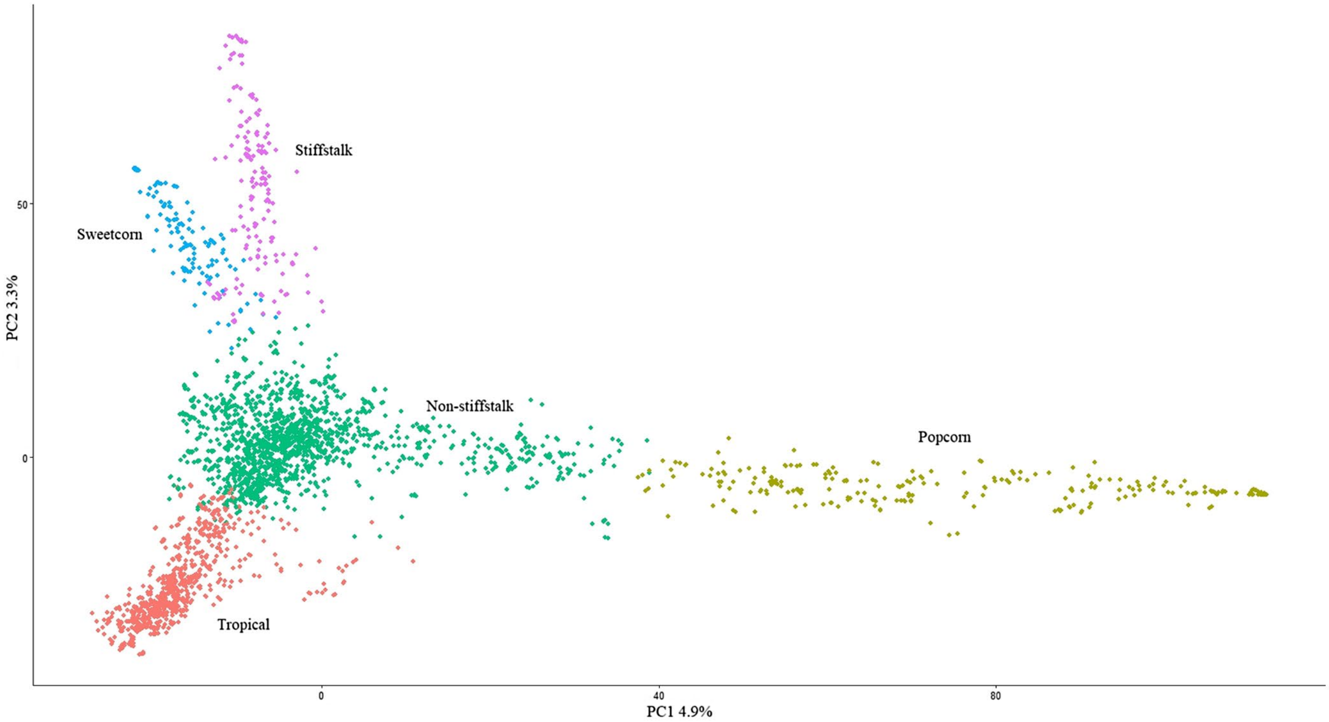

FIGURE 1 Population structure via principal component analysis (PCA), with k = 5 clusters, using 28,260 SNPs and 2,748 maize lines. Each solid circle represents a line

TABLE 2 Mean and standard deviation of predictive ability ($r_{(\hat{y},g)}$) of the genomic predictions obtained in TSG1 for different sizes of training sets (N_t) and validation sets (Nv) in the traits of plant height (PH) and ear height (EH)

Traits	N_t	N_v	$r_{(\hat{y},g)}$
PH	10	53	0.16 ± 0.20^{a}
	20	43	0.20 ± 0.13^{a}
	30	33	0.22 ± 0.15^{a}
EH	10	53	0.02 ± 0.11^{a}
	20	43	0.03 ± 0.09^{a}
	30	33	0.05 ± 0.10^{a}

[a]Mean values followed by the same letter in the column do not differ by the Tukey test at 5%; coefficient of variation: PH = 0.85, EH = 3.09.

TABLE 3 Mean and standard deviation of predictive ability ($r_{(\hat{y},g)}$) obtained in TSG2 for different sizes of the training set (N_t) and accordance with the panels USP, NCRPIS and ASSO for the traits of plant height (PH) and ear height (EH)

Trait	TS panel	N_t	$r_{(\hat{y},g)}$
PH	USP + NCRPIS + ASSO	2,465	0.29 ± 0.13^{a}
	NCRPIS + ASSO	2,434	0.20 ± 0.13^{a}
	NCRPIS	2,046	0.24 ± 0.11^{a}
	ASSO	388	0.08 ± 0.17^{b}
	NCRPIS (*tropical*)	512	0.21 ± 0.11^{a}
	ASSO (*tropical*)	136	0.20 ± 0.13^{a}
EH	USP + NCRPIS + ASSO	2,465	0.44 ± 0.11^{a}
	NCRPIS + ASSO	2,434	0.15 ± 0.11^{cd}
	NCRPIS	2,046	0.18 ± 0.10^{c}
	ASSO	388	0.11 ± 0.15^{cd}
	NCRPIS (*tropical*)	512	0.09 ± 0.09^{d}
	ASSO (*tropical*)	136	0.30 ± 0.11^{b}

Note: Validation set (VS): 32 lines (USP panel) determined at random; coefficient of variation: PH = 0.67, EH = 0.55. Mean values followed by the same letter in the column do not differ by the Tukey test at 5%.

TABLE 4 Mean and standard deviation of predictive ability ($r_{(\hat{y},g)}$) of the genomic predictions obtained in TSG3 for different sizes of the optimized training set (N_t) obtained via panel NCRPIS and ASSO for the traits of plant height (PH) and ear height (EH)

Trait	N_t	$r_{(\hat{y},g)}$	
-	-	OTS	RTS
PH	50	0.17 ± 0.14^{bB}	0.14 ± 0.16^{bB}
	250	0.32 ± 0.10^{aA}	0.17 ± 0.19^{bB}
	500	0.31 ± 0.19^{aA}	0.18 ± 0.19^{bB}
	1,000	0.31 ± 0.10^{aA}	0.22 ± 0.18^{bAB}
	1,500	0.18 ± 0.15^{bB}	0.20 ± 0.11^{bAB}
EH	50	0.15 ± 0.14^{aBC}	0.06 ± 0.19^{aAB}
	250	0.32 ± 0.11^{aA}	0.07 ± 0.16^{bAB}
	500	0.22 ± 0.11^{aB}	0.12 ± 0.18^{bAB}
	1,000	0.06 ± 0.12^{aCD}	0.16 ± 0.16^{aA}
	1,500	0.03 ± 0.12^{aD}	0.12 ± 0.11^{aAB}

Note: Validation set (VS): 32 lines (USP panel) determined at random; coefficient of variation: PH = 0.68, EH = 1.12. Mean values followed by the same lowercase letter in the same row and the same uppercase letter in the same column do not differ by the Tukey test at 5%.

Abbreviations: OTS, optimized training set; RTS, randomized training set.

The values observed ranged from 0.31 to 0.32 (Table 4), without significant differences among them. Furthermore, for EH, high estimates of predictive ability were obtained using OTSs with N_t = 250 individuals (Table 4).

Therefore, OTSs consisting of 250 individuals may be used, without the risk of penalizing the estimate of predictive ability, as long as they are chosen by the method tested here. As proof of the effectiveness of the method, the results indicated that in the groups of RTS, predictive ability increased following the addition of individuals in the training set. Nevertheless, this increase did not have a practical or statistical effect. That was because the variation observed was insufficient to generate differences between the predictive abilities (Table 4). Additionally, the values observed (RTS) were lower than those obtained when using an OTS composed of 250 individuals. Indeed, the best or clear comparison between RTS and OTS was when the differences in predictive abilities with N_t = 250 were analysed. In this group, the OTS had estimates 88.2% and 457% higher than RTS for PH and EH, respectively (Table 4).

3.3.4 | TSG4

The addition of lines belonging to the USP panel in the OTS and RTS groups led to an increase in the estimates of predictive ability for all the N_t tested (Table 5). The mean increase concerning the TSG3 was 0.55× greater for OTS and 1.08× greater for RTS in PH. For EH, these values were 2.21 and 4.03× greater in OTS and RTS, respectively. Both training set tested (OTS and RTS) reached estimates of predictive ability up to 0.59 (Table 5), in the model tested and for the population of interest. However, as shown above, very small groups were not effective in predictions. Thus, if the scenario had a VS greater than that tested here ($32 < N_v < 63$), the size of N_t would decrease ($50 < N_t < 81$; $250 < N_t < 281$). It seems that N_t near 50 individuals reduces the predictive ability. However, N_t near 250 individuals would tend to maintain high levels of the estimates of abilities because it was the most robust group size in other scenarios.

4 | DISCUSSION

The moderate to high heritability estimates that we found are relevant to the analysis. It is because heritability affects the predictive ability due to a positive tendency between them, with a significant response in the selection of traits with high heritability (Cavalcanti, Resende, Santos, & Pinheiro, 2012; Muranty et al., 2015).

TABLE 5 Mean and standard deviation f predictive ability ($r_{(\hat{y},g)}$) of the genomic predictions obtained in TSG4 for different sizes of the optimized training set (N_t) obtained via the USP, NCRPIS and ASSO panels for the traits of plant height (PH) and ear height (EH)

Traits	N_t	$r_{(\hat{y},g)}$	
-	-	OTS	RTS
PH	81	0.43 ± 0.10^{aA}	0.41 ± 0.13^{aA}
	281	0.44 ± 0.08^{aA}	0.40 ± 0.13^{aA}
	531	0.41 ± 0.09^{aA}	0.37 ± 0.14^{aA}
	1,031	0.40 ± 0.09^{aA}	0.39 ± 0.11^{aA}
	1,531	0.33 ± 0.10^{aB}	0.33 ± 0.11^{aB}
EH	81	0.57 ± 0.09^{aA}	0.58 ± 0.09^{aA}
	281	0.59 ± 0.10^{aA}	0.55 ± 0.13^{aA}
	531	0.51 ± 0.14^{aB}	0.53 ± 0.13^{aAB}
	1,031	0.42 ± 0.15^{bC}	0.53 ± 0.12^{aAB}
	1,531	0.42 ± 0.14^{bC}	0.48 ± 0.11^{aB}

Note: Validation set (VS): 32 lines (USP panel) determined at random; coefficient of variation: PH = 0.27, EH = 0.23. Mean values followed by the same lowercase letter in the same row and the same uppercase letter in the same column do not differ by the Tukey test at 5%.

Abbreviations: OTS, optimized training set; RTS, randomized training set.

Regarding the TSG1 process, predictions within very small groups limit the predictive ability of the model and do not significantly change its estimates, even if the size of the training set (N_t) is increased. The mean coefficient of the relationship (r_{xy}), via the genomic relationship matrix, was 0.19 (Figure S1) in this panel, which is high, because it is near the half-sib mean. Thus, the need for including more significant genetic variability for carrying out GS is evident.

Concerning TSG2, analysing its inference (Table 3) and taking into consideration the differences in the sizes of the training set between the panels (NCRPIS = 512 and ASSO = 136), it seems that even with a smaller proportion of the training set, the ASSO panel can predict better than NCRPIS regarding USP panel. This result indicates a significant relationship between ASSO lines and USP lines. However, the NCRPIS and ASSO panels have an r_{xy} of 0.02 and 0.017 with the USP, respectively. Another result of this study is the clear structure of the tropical cluster within the panels NCRPIS and ASSO (Figure 2A1,B1). Considering only this cluster and relating it to the USP panel (Figure 2A2,B2; Figure S2), the values of r_{xy} are greater than those described above, namely 0.071 for NCRPIS and 0.08 for ASSO, with maximum values of 0.263 and 0.184 for each panel, respectively. In this context, the tropical group inside ASSO (ASSO - tropical) panel provides higher predictive abilities when used as a training group than the whole ASSO panel (Table 3). The same tendency was not observed for the NCRPIS panel. In this case, the trend was a decline in predictive ability for EH. Thus, it might be expected that the strong genetic relationship between USP and ASSO was

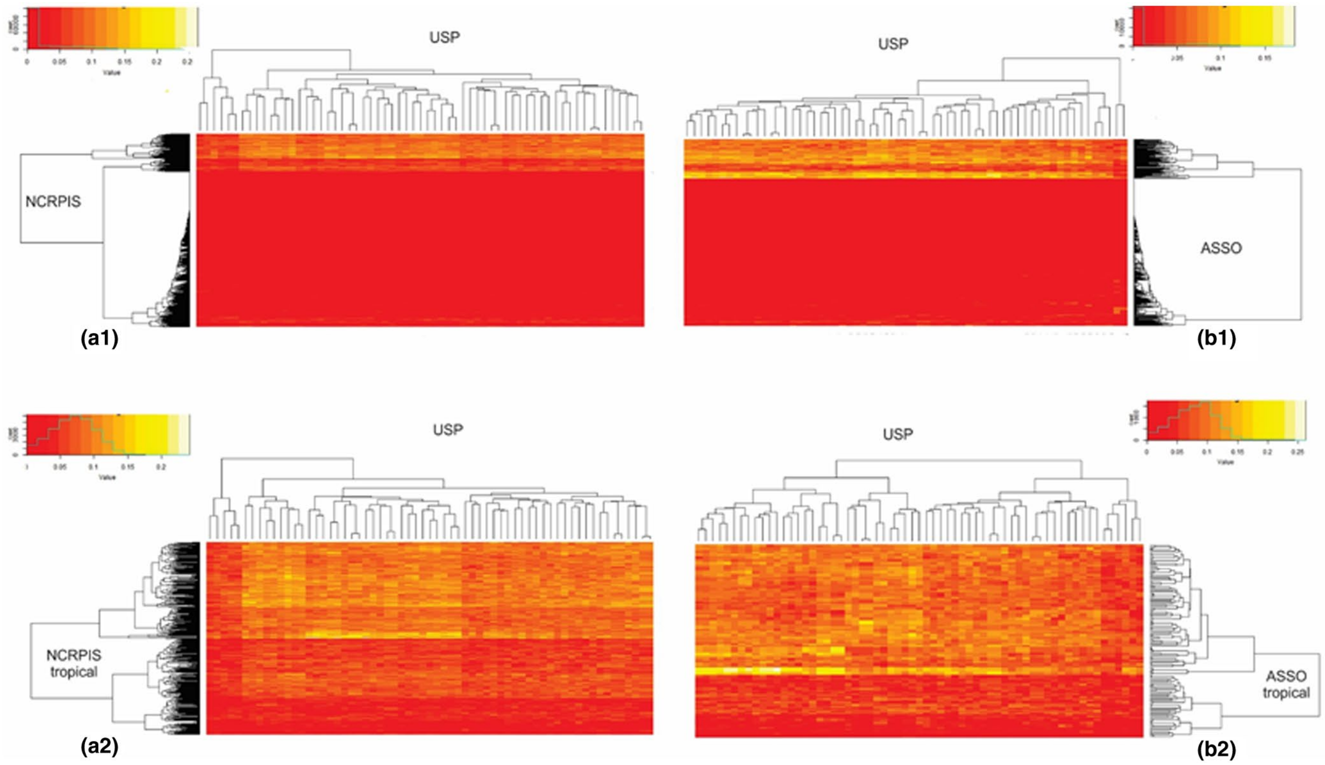

FIGURE 2 Heat map of the genomic relationship matrix (GRM), using 28,260 SNP markers, for the direct relationship between the panels USP and NCRPIS (A1), USP and ASSO (B1), USP and NCRPIS within the tropical cluster (A2) and USP and ASSO within the tropical cluster (B2)

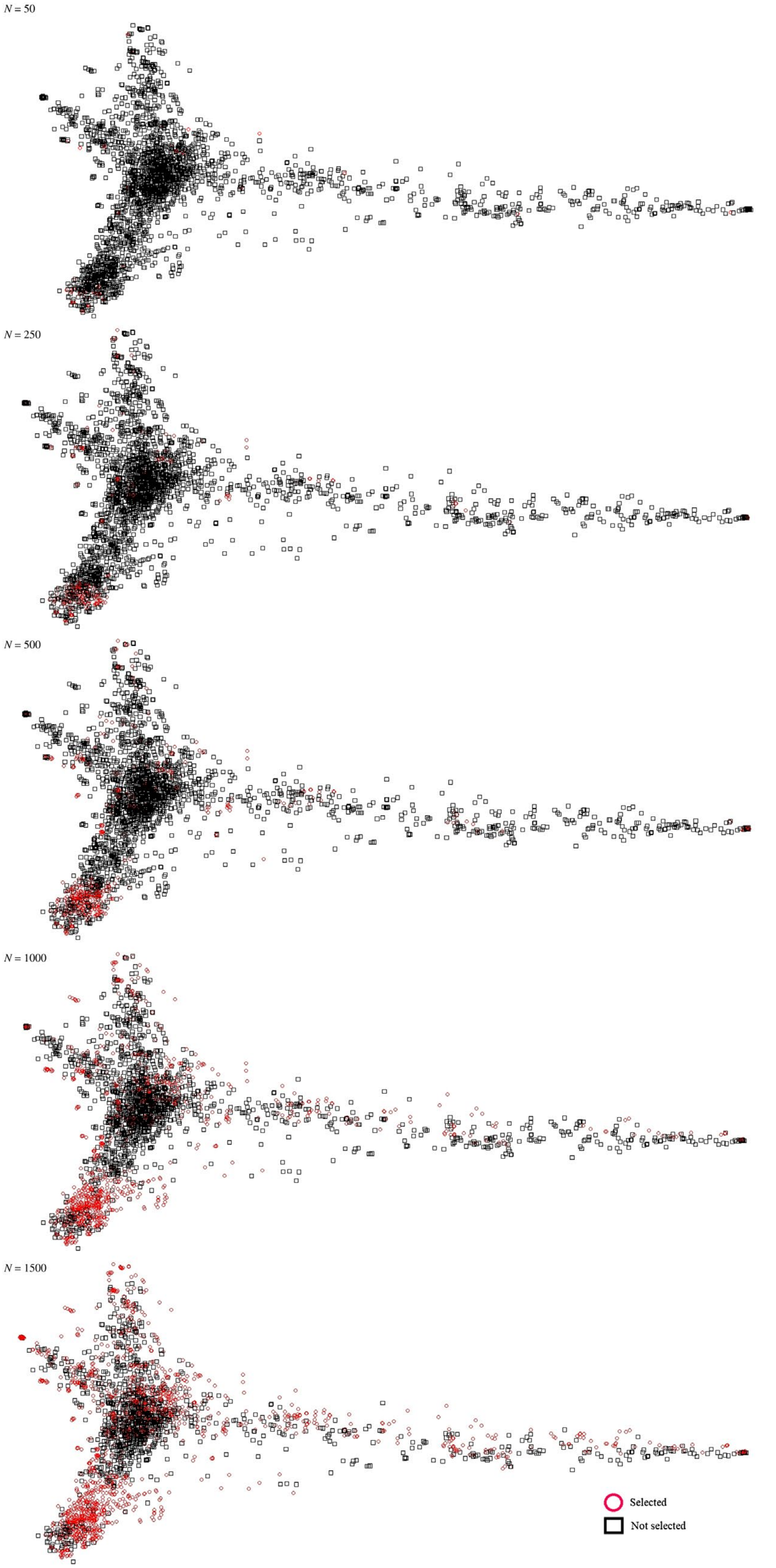

FIGURE 3 Population structure of the lines selected to compose the optimized training set (OTS) following their size (N_t) in the TSG3 scenario

determinant for the predictive abilities observed. Nevertheless, the high dispersion of NCRPIS lines seems to play a role in the absence of high prediction accuracies of USP lines. It has already been established that the population structure is a critical factor affecting the predictions in GS (Guo et al., 2013). Therefore, it is necessary to consider it in the establishment of the TS or definition of the models, to avoid unrealistic and spurious estimates of predictive abilities and accuracies (Riedelsheimer et al., 2013; Wray et al., 2013). Confirming the results we found, Albrecht et al. (2011) and Guo et al. (2013) observed that TS and VS groups established from the same cluster could improve predictive ability. On the other hand, when TS and VS diverge or when the genotypes come from different crosses or families, the predictive abilities and accuracies of the models of genomic selection tend to decrease.

In the TSG3 scenario, the superiority of OTSs is due to the optimization method used. It is based on a derivation of a computational method published by Akdemir et al. (2015). This process makes an efficient approximation of the prediction error variance (PEV), based on the principal components, via markers in the candidate and tested lines. The PEV was estimated in the test group (USP panel) instead of the candidate group (NCRPIS + ASSO). Consequently, the selection of lines for the optimized group occurs following those that minimize the PEV in the USP panel. This tendency can be observed for smaller groups, N_t = 50 and 250, in which the selection of lines to build the OTS is, in most cases, from the tropical group (Figure 3). In contrast, as the size of the selected group increases, lines that are divergent and lack any relationship with the USP panel tend to be included in the training set (Figure 3). Thus, there is a decrease in predictive ability, N_t = 1,500 and 500, for PH and EH, respectively (Table 4). These results show that the structure of the population plays a vital role in optimizing the training sets because when the population effect is smaller, the OTSs can more accurately predict the lines of the USP panel. Moreover, under strong population structuring, the OTS is less accurate. Nevertheless, small population sizes, such as N_t = 50, also limit predictions because they do not include sufficient genetic variability. Hence, not only should the effective size of the population be taken into consideration, but its degree of relationship with the population to be predicted should as well.

For TSG4, another important point is that the establishment of the groups via RTS depends on a random sampling of the lines. Thus, lines belonging to a single subpopulation and that are highly related should be selected, which results in a loss of diversity in the breeding programme, just as in the predictive ability of the GS models. On the other hand, the OTS follows a pattern to select individuals that will compose it, through which kinship and genetic variability of the chosen materials are controlled (Figure 3). Optimization within only one group with kinship does not improve the correlation of the GEBVs with the nonphenotyped lines. However, the association of this information with the genetic distance can lead to a selection of a stable and genetically representative sample of lines to be phenotyped (Schmidt et al., 2016), thus optimizing the process of establishing the training set.

Windhausen et al. (2012) indicate that when 50% of the genotypes in the validation group are included in the training group, the predictive ability increases for all the traits and the most significant improvements are for those with high heritability. Additionally, the results obtained via TSG3 and TSG4 indicate that predictive ability can be improved if the lines that compose the training group are correctly selected by efficient methods, such as that proposed by Akdemir et al. (2015), making satisfactory results feasible mainly through smaller groups of individuals. Simulations also suggest that small training groups may be as precise as large ones (Habier, Fernando, & Dekkers, 2009). This prediction has been confirmed in real breeding populations (Isidro et al., 2015; Wong & Bernardo, 2008) and this study. In general, large TSs are recommended for traits controlled by a large number of genes of small effect and with a large influence of the environment on phenotype (Goddard & Hayes, 2009). Nevertheless, the traits used in this study are not so complex as grain yield (Multani et al., 2003) and are highly heritable (Peiffer et al., 2014), which allowed us to use small training sets.

The standard deviation estimates (σ) were similar in most scenarios, except for TSG3 (Table 4), which showed the most significant deviation for RTS. Knowledge of standard deviations of the predictive ability is important for small training sets, mainly because some individuals with rare alleles might have a considerable effect and contribute to increasing the predictive ability (Schmidt et al., 2016).

4.1 | Application to plant breeding and perspectives

Taking into consideration the central theme of the problem presented, in which the resources used are limited, and there is the possibility of using the genotypic information from the whole programme population, the use of external panels to predict small and nonrelated populations is possible. A similar procedure was successfully used by Jarquin, Specht, and Lorenz (2016) in soybean, achieving predictive abilities of up to 0.92 for oil and protein and 0.79 for yield. Those results and the results presented here support the recommendation to breeders to access public genomic and phenotypic databases and use them in their GS programmes. A large number of genotypes in the TSs established are not necessary when optimization of the population is used. However, it is a large group forming the initial data with wide genetic variability. Based on that, the optimization process of the TS can be effective.

Furthermore, it is recommended to combine the use of external information and of some of the lines of the programme in establishing the TS. This approach will lead to the selection of the best individuals in future phases, reducing the amount of material to be screened in the field trials. Accordingly, costs will be reduced while increasing gains by making the right selection. However, this is valid only for traits with high heritability. Additional studies are necessary for traits with lower heritabilities. It should be noted that other factors were not studied in this work, such as the statistical models used (Heslot, Yang, Sorrells, & Jannink, 2012), the number and type of markers (Chen & Sullivan, 2003; Poland & Rife, 2012) and the imbalance of linkage (Habier et al., 2007). These factors can influence the predictive ability of GS and should be studied within the proposed panorama.

A prospect of how to apply this procedure in a breeding pipeline is described below. Let us consider a framework with a TS of 480 individuals to predict new double haploid (DH) lines. In this context, a VS may be defined with Nv = 48 (10% of the TS). The strategy in this procedure is to genotype and phenotype only a part of the internal population in each cycle or each year and to use the other data from public databases. Through the years, there should be a substitution of the external lines for private lines from the programme. In this way, the financial resources necessary to compose a TS will be reduced through the years. This strategy can be summarized as follows: first year: (a) genotyping and phenotyping 96 individuals. (b) Among the 96 individuals, 48 are established as the VS; the remaining ones will make up part of the TS, together with lines from public databases. Thus, 48 breeding lines and 432 lines from external panels are selected based on the OTS. (c) Validation and prediction of the GEBVs. Second year: (a) genotyping and phenotyping of 96 new individuals. (b) The VS is established based only on internal lines. Thus, among the 192 private lines (96 from the 1st year + 96 from the 2nd year), 48 will be part of the VS and 148 the TS. Consequently, 332 external lines selected via optimization procedures are required. (c) Validation and prediction of the GEBVs. This method is repeated up to the fifth year, in which practically the entire TS will be substituted by internal lines of the programme, considering the introduction of data from 96 lines annually. Additionally, as shown here, this procedure has the potential to maintain the adequate predictive ability of the predictive model.

Another advantage of our approach is the possibility of defining the best technical, operational and financial balance for each programme according to the resources available over time and for each crop. Furthermore, using this procedure, small investments per year are necessary, amortizing the total value over time. Even though the environment has a smaller effect on PH and EH than on yield, for example, in terms of predictive ability, there is evidence that the predictive ability decreases when the TS and VS are evaluated under contrasting environments (Windhausen et al., 2012). Moreover, the substitution of public lines by internal ones should respect the technical and economic limitations of each breeding programme. Thus, creating nets of collaboration, which develop public databases of tropical corn, may help small breeding programmes implement GS, especially in recently created programmes that have low genetic variability in their germplasm and small budgets.

5 | CONCLUSIONS

Public databases of genomic and phenotypic information are valid sources for creating training sets to implement genomic selection in breeding programmes with limited resources. However, the natural population structure of these datasets may affect the predictive ability. Thus, we recommend the use of optimization methods, as an example Akdemir et al. (2015), to build the training sets. Moreover, small groups of individuals (250) selected from public panels are sufficient to achieve satisfactory, over 0.53, for predictive abilities of genomic selection.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

AUTHORS CONTRIBUTIONS

Morais P.P.P., Akdemir D., Andrade L.R.B., Fritsche-Neto R. and Jannink J-L. contributed to data analysis and writing of the manuscript. Morais P.P.P, Andrade L.R.B., Alves F.C., Lyra D.H. and Granato I.S.C. contributed to data assessment (field) and paper review. Borem, A. contributed to paper review.

DATA AVAILABILITY STATEMENT

USP data are available with this https://doi.org/10.6084/m9.figshare.5525263 and at http://www.genetica.esalq.usp.br/alogamas/data2.html. Nested Association Mapping Population data of McMullen et al. (2009) are available at http://mirrors.iplantcollaborative.org/browse/iplant/home/shared/panzea/genotypes/SNPs/NAM_map_and_genos-120731.zip. North Central Regional Plant Introduction Station (NCRPIS) data of Romay et al. (2013) are available at http://mirrors.iplantcollaborative.org/browse/iplant/home/shared/panzea/dataFromPubs/Romay2013GenomeBiol. Maize 282 Association Panel data of Flint-Garcia et al. (2005) are available at http://mirrors.iplantcollaborative.org/browse/iplant/home/shared/panzea/genotypes/SNPs/FlintGarcia2005PlantJ.

ORCID

Pedro Patric Pinho Morais https://orcid.org/0000-0001-9704-4816
Roberto Fritsche-Neto https://orcid.org/0000-0003-4310-0047
Filipe Couto Alves https://orcid.org/0000-0002-2127-4276
Ítalo Stefanine Correia Granato https://orcid.org/0000-0003-2093-6810

REFERENCES

Akdemir, D. (2017). STPGA: Selection of training populations with a genetic algorithm. *arXiv Prepr*. https://doi.org/10.1101/111989

Akdemir, D., Sanchez, J. I., & Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, *47*, 38. https://doi.org/10.1186/s12711-015-0116-6

Albrecht, T., Wimmer, V., Auinger, H.-J., Erbe, M., Knaak, C., Ouzunova, M., … Schön, C.-C. (2011). Genome-based prediction of testcross values in maize. *Theoretical and Applied Genetics*, *123*, 339–350. https://doi.org/10.1007/s00122-011-1587-7

Bernardo, R. (2014). Genomewide selection when major genes are known. *Crop Science*, *54*, 68. https://doi.org/10.2135/cropsci2013.05.0315

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

Cavalcanti, J. J. V., de Resende, M. D. V., dos Santos, F. H. C., & Pinheiro, C. R. (2012). Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. *Revista Brasileira De Fruticultura*, *34*, 840–846. https://doi.org/10.1590/S0100-29452012000300025

Chen, X., & Sullivan, P. F. (2003). Single nucleotide polymorphism genotyping: Biochemistry, protocol, cost and throughput. *Pharmacogenomics Journal*, *3*, 77–96. https://doi.org/10.1038/sj.tpj.6500167

Clark, S. A., Hickey, J. M., & van der Werf, J. H. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genetics Selection Evolution*, *43*, 18. https://doi.org/10.1186/1297-9686-43-18

Dell'Acqua, M., Gatti, D. M., Pea, G., Cattonaro, F., Coppens, F., Magris, G., … Pè, M. E. (2015). Genetic properties of the MAGIC maize population:

A new platform for high definition QTL mapping in Zea mays. *Genome Biology*, *16*, 167. https://doi.org/10.1186/s13059-015-0716-z

Endelman, J. B. (2011). Ridge Regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome Journal*, *4*, 250. https://doi.org/10.3835/plantgenome2011.08.0024

Flint-Garcia, S. A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., … Buckler, E. S. (2005). Maize association population: A high-resolution platform for quantitative trait locus dissection. *Plant Journal*, *44*, 1054–1064. https://doi.org/10.1111/j.1365-313X.2005.02591.x

Gilmour, A. R., Gogel, B. J., Cullis, B. R.,Welham, S. J., & Thompson, R. (2015). *ASReml user guide release 4.1*. Hemel Hempstead, UK: Functional Specification, VSN International Ltd. Retrieved from www.vsni.co.uk

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q. I., & Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, *9*, e90346. https://doi.org/10.1371/journal.pone.0090346

Goddard, M. E., & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, *10*, 381–391. https://doi.org/10.1038/nrg2575

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., … Gay, G. (2013). The impact of population structure on genomic prediction in stratified populations. *Theoretical and Applied Genetics*, *127*, 749–762. https://doi.org/10.1007/s00122-013-2255-x

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*, 2389–2397. https://doi.org/10.1534/genetics.107.081190

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2009). Genomic selection using low-density marker panels. *Genetics*, *182*, 343–353. https://doi.org/10.1534/genetics.108.100289

Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Research*, *91*, 47. https://doi.org/10.1017/S0016672308009981

Heslot, N., Yang, H.-P., Sorrells, M. E., & Jannink, J.-L. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Science*, *52*, 146–160. https://doi.org/10.2135/cropsci2011.06.0297

Hung, H.-Y., Browne, C., Guill, K., Coles, N., Eller, M., Garcia, A., … Holland, J. B. (2012). The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity (Edinb)*, *108*, 490–499. https://doi.org/10.1038/hdy.2011.103

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., & Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics*, *128*, 145–158. https://doi.org/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. https://doi.org/10.1534/g3.116.031443

Jonas, E., & De Koning, D.-J. (2013). Does genomic selection have a future in plant breeding? *Trends in Biotechnology*, *31*, 497–504. https://doi.org/10.1016/j.tibtech.2013.06.003

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., … Buckler, E. S. (2009). Genetic properties of the maize nested association mapping population. *Science*, *325*, 737–740. https://doi.org/10.1126/science.1174320

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*, 1819–1829. 11290733

Multani, D. S., Briggs, S. P., Chamberlin, M. A., Blakeslee, J. J., Murphy, A. S., & Johal, G. S. (2003). Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. *Science*, *302*, 81–84.

Muranty, H., Troggio, M., Sadok, I. B., Rifaï, M. A., Auwerkerken, A., Banchi, E., … Bink, M. C. A. M. (2015). Accuracy and responses of genomic selection on key traits in apple breeding. *Horticulture Research*, *2*, 15060. https://doi.org/10.1038/hortres.2015.60

Newell, M. A., & Jannink, J.-L. (2014). Genomic selection in plant breeding. In D. Fleury, & R. Whitford (Eds.), *Crop breeding: Methods and protocols* (pp. 117–130). New York, NY: Springer New York.

Peiffer, J. A., Romay, M. C., Gore, M. A., Flint-Garcia, S. A., Zhang, Z., Millard, M. J., … Buckler, E. S. (2014). The genetic architecture of maize height. *Genetics*, *196*, 1337–1356. https://doi.org/10.1534/genetics.113.159152

Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome Journal*, *5*, 92. https://doi.org/10.3835/plantgenome2012.05.0005

Pszczola, M., Strabel, T., Mulder, H. A., & Calus, M. P. L. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of Dairy Science*, *95*, 389–400. https://doi.org/10.3168/jds.2011-4338

R Core Team (2017). *R: A language and environment for statistical computing*. R Found. Stat. Comput. Retrieved from http://www.R-project.org

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

Rincent, R., Laloe, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., … Moreau, L. (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. https://doi.org/10.1534/genetics.112.141473

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., … Gardner, C. A. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, *14*, R55. https://doi.org/10.1186/gb-2013-14-6-r55

Schmidt, M., Kollers, S., Maasberg-Prelle, A., Großer, J., Schinkel, B., Tomerius, A., … Korzun, V. (2016). Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. *Theoretical and Applied Genetics*, *129*, 203–213. https://doi.org/10.1007/s00122-015-2639-1

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

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

Windhausen, V. S., Atlin, G. N., Hickey, J. M., Crossa, J., Jannink, J.-L., Sorrells, M. E., … Melchinger, A. E. (2012). Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3: Genes|genomes|genetics*, *2*, 1427–1436. https://doi.org/10.1534/g3.112.003699

Wong, C. K., & Bernardo, R. (2008). Genomewide selection in oil palm: Increasing selection gain per unit time and cost with small populations. *Theoretical and Applied Genetics*, *116*, 815–824. https://doi.org/10.1007/s00122-008-0715-5

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Review Genetics*, *14*, 507–515. https://doi.org/10.1038/nrg3457

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Pinho Morais PP, Akdemir D, Braatz de Andrade LR, et al. Using public databases for genomic prediction of tropical maize lines. *Plant Breed*. 2020;00:1–11. https://doi.org/10.1111/pbr.12827