Genetics and Molecular Research
ISSN 16765680

Genomic breeding value prediction for simple maize hybrid yield using total effects of associated markers, under different imbalance levels and environments

N.F. Cantelmo[1], **R.G. Von Pinho**[2] **and M. Balestre**[3]

[1]Departamento de Biologia, Universidade Federal de Lavras, Lavras, MG, Brasil
[2]Departamento de Agricultura, Universidade Federal de Lavras, Lavras, MG, Brasil
[3]Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, MG, Brasil

Corresponding author: M. Balestre
E-mail: marciobalestre@dex.ufla.br

Genet. Mol. Res. 15 (1): gmr.15017232
Received July 16, 2015
Accepted November 19, 2015
Published March 11, 2016
DOI http://dx.doi.org/10.4238/gmr.15017232

ABSTRACT. The main objective of a maize breeding program is to generate hybrid combinations that are more productive than those pre-existing in the market. However, the number of parents, and consequently the number of crosses, increases so rapidly that the phenotypic evaluation of all the possible combinations becomes economically and technically infeasible. In this context, predicting the performance of the most promising genotypes may increase the genetic gains with increased selection intensity and reduced breeding cycles. Thus, the present study aimed to use the total effects of associated markers method to predict genomic breeding values (GBVs) via cross-validation and by using different imbalance levels (10, 30, 50, and 70%). A set of 51 genotyped strains was used with 79 microsatellite markers and 273 hybrids that were generated by a partial diallel. A total of 186 and 272 hybrids were analyzed in the experiments within the southern and central regions of Brazil, respectively. The GBVs were, thus, predicted

©FUNPEC-RP www.funpecrp.com.br

for each location in both the regions, and for training in one region and validation in another region. The correlation between the predicted and observed GBVs ranged from 0.48 to 0.91, depending on the imbalance level and the region analyzed. Overall, the results obtained in the present study were promising, particularly considering that a small number of markers were used and that the training and predictions occurred in the very distinct regions of southern and central Brazil.

Key words: Genome wide selection; Microsatellite markers; Genotype-by-environment interaction

INTRODUCTION

In maize breeding programs, to successfully obtain more productive hybrids adapted to different farming conditions, selection and evaluation of the best genotypes is extremely important (Ferreira et al., 2010) in identifying superior hybrids. As phenotypic evaluations of crosses between strains require many resources (Schrag et al., 2010), evaluating all the possible combinations is often economically and technically infeasible. With the growing number of strains, driven by the double-haploid technique, the number of crosses between the strains of different heterotic groups increases faster, thereby, further hindering the selection of the most promising combinations. Thus, in practice, only a small proportion of the crosses are evaluated in field experiments.

Identifying the best hybrids without phenotypically evaluating them has been called ‘prediction’; it involves assessment of field data from related trials and, has more recently, used molecular marker information (Meuwissen et al., 2001; Schrag et al., 2010; Guo et al., 2013a).

Prediction has a long history in plant genetic improvement. The development of quantitative genetic theories for both animal and plant genetic improvement was intensively motivated by the need for a genetic prediction structure to more clearly guide the breeding programs (Walsh, 2014). Therefore, the prediction of the performance of a hybrid based on the information of its parents is of great interest for breeders, and it may substantially increase the efficiency of breeding programs (Fu et al., 2012).

Molecular markers, discovered in the 1980s, began to be used in animal and plant breeding on the premise that the information at the DNA level could lead to faster genetic gain compared to the use of phenotypic data alone (Meuwissen et al., 2001). The possibility of using markers to establish heterotic groups arose from the theoretical relationship between genetic distance and heterosis (Charcosset and Essioux, 1994; Melchinger, 1999). Although a positive relationship between heterosis and genetic divergence was observed in such studies, conflicting results were reported by Balestre et al. (2008), Dhliwayo et al. (2009), and Devi and Singh (2011). Correlation between the predicted and observed values has been demonstrated to be low, which could be due to low or no linkage disequilibrium between the markers and genes involved in the trait (Charcosset and Essioux, 1994). Thus, these methods failed to make accurate predictions that could be integrated into the breeding programs (Melchinger, 1999).

In addition to molecular markers, several models have been proposed to predict the performance of maize hybrids. It was proposed that the performance of hybrids should be directly predicted based on the trait means of their parents. Although, this is considerably, a direct method, it is ineffective because it does not consider the high dominance level present in the yield trait (Guo et al., 2013a). Some authors have tried to predict the yield in maize hybrids using the general combining activity (GCA; Melchinger et al., 1987). However, this technique ignores the specific combination ability (SCA), which is related to heterosis and is an important component in hybrid performance (Gardner and Eberhart, 1966). Some models also proposed inclusion of the SCA in the prediction, which has demonstrated advantages when its variance is greater than or equal to the GCA variance (Schrag et al., 2006).

Best linear unbiased prediction (BLUP) is another approach for random effect prediction, used since 1990s (Henderson, 1984). The inclusion of marker information to calculate the genetic values obtained using BLUP was first demonstrated in animal breeding by Fernando and Grossman (1989). It has been estimated that the use of this method can increase genetic gains by 8-38% (Meuwissen et al., 2001). Thus, this method, which has previously been extensively used in animal breeding, also began to be adopted in plant breeding, particularly for maize (Bernardo, 1994, 1996).

BLUP uses any available marker to calculate the relationship between genotypes, thereby, estimating the performance of the untested hybrids based on the performance of the tested hybrids by comparing the relatedness between the two (Bernardo, 1994, 1996). Bernardo (1996) used this method in simple maize hybrids and obtained correlations between the predicted and observed values that ranged from 0.42 to 0.76 for kernel yield, 0.75 to 0.93 for kernel moisture content, 0.30 to 0.74 for breakage, and 0.16 to 0.53 for lodging. Since then, this method has been routinely used in maize breeding programs (Massman et al., 2013a).

More recently, Schrag et al. (2007, 2009) suggested replacing the relationship matrix used by Bernardo (1994) with the genotype matrix of the markers observed in the mixed model equation matrix in which they directly used the marker incidence matrix. This method was denominated ‘total effects of associated markers’ (TEAM). A training population is used to estimate genetic values for each marker and is then validated via untested hybrids (Schrag et al., 2009). This method demonstrated better results than those found with the method proposed by Bernardo (1994). Another important advantage of this method is the possibility of working with data sets that have missing observations (Schrag et al., 2007).

In this context, commercial hybrid breeding programs are optimal sources of data for studying prediction models because they generate a large number of hybrids and evaluate them under multiple environments (Massman et al., 2013a) and fulfill the needs of a real breeding program, such as phenotypic evaluation at several sites over the course of several years. The genotypes evaluated in the experiments may be used as a training population and are cross-validated using imbalance levels, thus identifying the most promising genotypes and testing the model’s efficiency under real conditions, which is not easily available in publications of this nature. These conditions give realism to the model analysis, leading to more certainty in choosing the best model.

The present study aimed to predict simple maize hybrid yield using TEAM, microsatellite marker data, and phenotypic information from Brazilian national maize breeding trials at different sites and in different cropping seasons.

MATERIAL AND METHODS

Phenotypic evaluation

A total of 51 strains belonging to different heterotic groups were used in the crosses to generate the population. From the crosses of these strains, 273 hybrids were obtained in a partial diallel system.

Of all the hybrids generated, 186 were evaluated for bean production at six sites, distributed throughout southern Brazil [Guarapuava, Paraná (PR); Vacaria, Rio Grande do Sul

(RS); Ipiranga, PR; Sananduva, RS; Faxinal dos Guedes, Santa Catarina (SC); and Itapeva, São Paulo (SP)], during the 2011/2012 crop season. Two hundred and seventy two hybrids (including the 186 hybrids evaluated from the southern region) were evaluated at nine sites within central Brazil [Presidente Olegário, Minas Gerais (MG); Uberaba, MG; Capinópolis, MG; Araguari, MG; Madre de Deus, MG; Nazareno, MG; Boa Esperança, MG; Lavras, MG; and Araguari, MG)] during the 2011/2012 crop season. The experiment was conducted using an incomplete block design with two replicates per site and plots comprising four 5-m rows spaced 0.7 m apart.

Hybrid production was evaluated, and the plot weight was corrected to a moisture content of 13%, which was converted into t/ha. The site was prepared, and the topdressing fertilizer was used according to that recommended for each experimental site; moreover, necessary crop practices were followed to control fall armyworm (*Spodoptera frugiperda*) and corn earworm (*Helicoverpa zea*), and to control weeds.

Genotypic analysis

A total of 79 microsatellite markers, distributed throughout the ten maize linkage groups, were used to genotype the 51 strains (Table 1).

Table 1. Distribution of the 79 microsatellite markers in the ten linkage groups (LGs) of maize.

Marker	LG	Bin	Marker	LG	Bin	Marker	LG	Bin	Marker	LG	Bin
bnlg1179	1	1.01	dupssr08	3	3.09	bnlg1200	7	7.01	umc1139	8	8.01
bnlg1014	1	1.01	bnlg1496	3	3.09	bnlg1808	7	7.02	bnlg1056	8	8.01
bnlg1007	1	1.02	umc1136	3	3.09	bnlg1305	7	7.03	bngl2082	8	8.03
bnlg1614	1	1.02	phi072	4	4.01	umc1342	7	7.04	bngl1067	8	8.03
bnlg1866	1	1.03	umc1101	4	4.09	bnlg2259	7	7.04	umc1858	8	8.04
umc1128	1	1.07	umc1109	4	4.1	umc1154	7	7.05	phi015	8	8.08
phi037	1	1.08	umc1197	4	4.11	umc1075	8	8.01	bnlg1131	8	8.09
bnlg1643	1	1.08	umc1058	4	4.11	umc1414	8	8.01	bnlg2122	9	9.01
umc1725	1	1.11	phi019	4	4.11	bnlg1194	8	8.01	umc1040	9	9.01
umc1797	1	1.12	umc1591	5	5.04	phi119	8	8.02	bnlg1724	9	9.01
umc1079	2	2.06	umc1482	5	5.04	umc1034	8	8.03	umc1078	9	9.05
bnlg1036	2	2.06	bnlg1237	5	5.05	phi115	8	8.03	umc1310	9	9.06
dupssr24	2	2.08	bnlg1118	5	5.07	mmc412	8	8.03	umc1319	10	10.01
bnlg1520	2	2.09	bnlg1371	6	6.01	umc2146	8	8.03	bnlg1079	10	10.03
umc1970	3	3.01	umc1006	6	6.02	phi121	8	8.03	umc2043	10	10.05
bnlg1601	3	3.05	umc1887	6	6.03	umc2147	8	8.03	bnlg1074	10	10.05
bnlg1160	3	3.06	umc1918	6	6.04	umc1157	8	8.03	umc1061	10	10.06
umc1148	3	3.07	bnlg1740	6	6.07	umc1202	8	8.05	bnlg1360	10	10.07
umc1167	3	3.08	phi089	6	6.08	bngl240	8	8.06	umc1084	10	10.07
bnlg1108	3	3.08	umc1066	7	7.01	umc1933	8	8.08			

A marker information matrix was constructed using a binary code: 1, for the presence of allele *t* in marker *m* in strain *i*, and 0 for the absence of the allele. This coding facilitates the construction of the additive and dominant matrices of the hybrids in contrast to the usual coding in which 2 and 0 code for homozygous and diploid strains, respectively.

Using this coding and considering that recombination was irrelevant in the homozygous strains, the additive matrix of the hybrids was constructed as follows:

$$\mathrm{A}_{lk} = \begin{cases} 2 & if \quad a_{1i} = a_{1j} = 1 \\ 1 & if \quad a_{1i} \neq a_{1j} \qquad \forall \; a_{1i} \vee a_{1j} = 1\,,0 \\ 0 & if \quad a_{1i} = a_{1j} = 0 \end{cases} \qquad \text{(Equation 1)}$$

where *a* was the phenotype of the t-ith allele of marker *m* in strains *i* and *j*, and $\forall$ indicated all the situations in which one or (∨) more strains exhibited different alleles.

Similarly, the matrix of the effects of dominance was constructed using the following relationship:

$$\Delta_{lk} = \begin{cases} a_{1i} \times a_{1j} \;\vee\; a_{2i} \times a_{2l} & \textit{if} \;\; t_i = t_j \\ a_{1i} \times a_{2j} + a_{1j} \times a_{2i} & \textit{otherwise} \end{cases} \qquad \text{(Equation 2)}$$

where 1 and 2 were the t-ith alleles of marker *m*. In the first case, we had the dominance deviation for the homozygotes, and in the second case, we had the deviation for the heterozygous complement.

Diallelic analysis and total effects of associated markers

Yield data in all the experiments for each region (southern and central) were subjected to analysis using mixed models. The mixed model used was given by:

$$\mathrm{y = X\beta + Zb + Wg + Ti + \varepsilon} \qquad \text{(Equation 3)}$$

where *X* corresponded to the matrix of the fixed effects: replicate within experiment within environment, experiment within environment and environment; β was the matrix of the incidence of the fixed effects; *Z* was the matrix of the random effects: block within replicate, within experiment, within site; *b* was the matrix of the incidence of the effects of blocks; *W* was the matrix of genotypes; *g* was the matrix of the incidence of the genotypes; *T* was the matrix of the genotype x environment interaction; *i* was the matrix of the incidence of the genotype x environment interaction, and ε was the matrix of the residues.

The estimated fixed effects and components of the phenotypic variance and predicted random effects were obtained via restricted maximum likelihood (REML) using the expectation-maximization algorithm. The means were adjusted for each site and the genomic breeding values were calculated. Heritability in the broad sense of the analyses between the environments was calculated using the following equation:

$$h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_{gxe}^2 / e + \sigma_r^2 / re) \qquad \text{(Equation 4)}$$

where σ^2_g, σ^2_{gxe}, and σ^2_r corresponded to genotypic variance, genotype x environment interaction variance, and residual variance, respectively, *r* corresponded to the number of replicates, and *a* corresponded to the number of environments.

In the marker model, the incidence matrices of the effects of the parental and specific combinations were replaced with matrices of the additive and dominant effects of the markers. Thus, the total effects of the associated markers model were given by:

$$\mathrm{y = X\beta + Aa + \Delta d + \varepsilon} \qquad \text{(Equation 5)}$$

where *y* was the vector of observations, *X* was the incidence matrix of the fixed effects (sites), β was the matrix of the fixed effects, *A* corresponded to the incidence matrix of the additive effects (*a*), and Δ corresponded to the incidence matrix of the dominance deviations (*d*), both aforementioned in the molecular data section. Unlike the RR-BLUP model, which is commonly used in genomic selection, TEAM is characterized as a mixed model because the environmental effects are considered in the fixed effects matrix. In addition, the marker environment interaction was confounded in the residual because of the high computational cost involved in estimating the effect of the interaction between each allele and their interactions with the environment.

The additive and dominant genetic values of each hybrid *i* were recovered by summing the additive effects of each allele that the individual possessed; this was described by the following equations:

$$\alpha_i = \sum_{j=1}^{n} \lambda a_j : \lambda = \{0,1,2\} \qquad \text{(Equation 6)}$$

$$\delta_i = \sum_{j=1}^{k} \phi d_j : \phi = \{0,1\} \qquad \text{(Equation 7)}$$

where *k* was the total number of allelic interactions *l* within each marker *m, n* was the number of alleles observed in the *m* markers, and λ and ϕ were variables indicating the status of the marker *m* in hybrid *i* for the additive and dominant effects, respectively. They were equivalent to the prediction point $\alpha_i = A_{(i\times n)}a$ and $\delta_i = \Delta_{(i\times k)}d$ in which *i* was the *i*-ith column of the matrices defined by Melchinger (1999).

The estimated components of phenotypic variance and fixed effects were obtained by predicting the additive and dominant effects contained in each marker via REML (Melo et al., 2014). The total genetic variance recovered was considered common for each marker and was calculated by:

$$\sigma_g^2 = n\sigma_a^2 + k\sigma_d^2 \qquad \text{(Equation 8)}$$

$$\sigma_a^2 = \left(\sum_{j=1}^{n} a_j^2 + trace[I_n W_n^{-1} \sigma^2]\right) \Big/ n \qquad \text{(Equation 9)}$$

$$\sigma_d^2 = \left(\sum_{j=1}^{k} d_j^2 + trace[I_n W_2^{-1} \sigma^2]\right) \Big/ k \qquad \text{(Equation 10)}$$

where W_n^{-1} and W_k^{-1} were sub-matrices of the inverse matrix of the mixed model equations.

Cross-validation

Cross-validation was performed in the set of hybrids that contained both the training and validation populations. For this analysis, different imbalance levels were applied to the data set.

The imbalance levels used were 10, 30, 50, and 70%. The process was repeated 100 times for each situation. The additive-dominant model was used to predict the genomic breeding value (GBV) of each hybrid. The correlation between the observed and predicted GBVs was used as a cross-validation parameter.

First, cross-validation was performed considering each location separately (southern and central regions); then, the data set was used in the cross-validation.

RESULTS

Table 2 shows the estimated variance components (σ_g, σ_{gxa}, and $\sigma_{residual}$) using the REML mixed model method (the genotypic variances were 0.34 within the southern region and 0.43 within the central region). The higher genotypic variance within the central region reflects the larger set of hybrids tested in this region that was responsible for higher variability. The genotype x environment interaction variance within the southern region was approximately 50% higher than that within the central region. The residual variance was essentially identical in both the regions. The difference in the genotype x environment interaction variance was reflected in the higher heritability within the central region (0.65) compared to the southern region (0.56).

Table 2. Components of genotypic variance (σ_g), genotype x environment interaction (σ_{gxa}), and residual variance ($\sigma_{residual}$) and their respective standard errors and heritability (h^2).

	Mega-environments	
	Southern Region	Central Region
σ_g^2	0.3451* (0.0426)	0.4336* (0.0485)
σ_{gxe}^2	0.1148* (0.0548)	0.0788* (0.0590)
σ_r^2	0.1564* (0.0428)	0.1511* (0.0500)
h^2	0.56	0.65

*Variance components were significant at 0.01% probability by the Z test.

The genotypic variance was higher than both the variances of the interactions and the residual variances at both the sites. Within the southern region, the genotypic variance was three times higher than the interaction variance, which was almost six-and-a-half times lower than the genotypic variance within the central region.

An analysis of the molecular markers in the 51 strains revealed 636 different alleles with a mean of 8.05 alleles per locus. The correlations shown in Table 3 reflect the accuracy when selection was based on the GBVs estimated from the markers and the phenotypic data. The correlation between the predicted and observed GBVs within the southern region ranged from 0.47 to 0.81, depending on the imbalance level used. At 10% imbalance, the mean correlation was 0.81, with a range of 0.47 to 0.94 and a variance of 0.009. At 30% imbalance, the mean correlation was 0.74, with a range of 0.50 to 0.90 and a variance of 0.007. At 50% imbalance, the mean correlation was 0.64, with a range of 0.32 to 0.79 and a variance of 0.009. Finally, at 70% imbalance, the mean correlation was 0.47, with a range of 0.07 to 0.68 (Table 3).

The central region exhibited the highest correlations, ranging from 0.76 to 0.90; these correlations were higher than those observed within the southern region. The highest correlation was observed when fewer hybrids were missing in the training population and the correlation was the lowest when only 30% of the hybrids were used to predict the remainder of the set (Table 3). This higher correlation could be explained by heritability, genotypic variance superiority compared to the interaction and residual variances, and using most hybrids within the central region.

Table 3. Correlations between predicted and observed genomic breeding values within the southern and central regions, the joint analysis, the southern region (training) for predicting the central region (southern → central) and the central region (training) for predicting the southern region (central → southern) using cross-validation.

Mega-environment	(%)	$r_{g\hat{g}}^{A}$	$r_{g\hat{g}}^{M}$	Variance	$r_{g\hat{g}}^{L}$	$r_{g\hat{g}}^{H}$
Southern	10	0.808	0.832	0.010	0.469	0.936
	30	0.738	0.750	0.007	0.499	0.895
	50	0.642	0.656	0.009	0.324	0.788
	70	0.465	0.482	0.015	0.073	0.684
Center	10	0.905	0.911	0.002	0.754	0.985
	30	0.838	0.842	0.002	0.72	0.923
	50	0.761	0.761	0.003	0.628	0.864
	70	0.63	0.647	0.007	0.328	0.767
Joint	10	0.659	0.675	0.013	0.303	0.902
	30	0.607	0.622	0.007	0.333	0.744
	50	0.557	0.566	0.007	0.194	0.712
	70	0.463	0.461	0.013	0.183	0.706
Center → Southern	10	0.853	0.863	0.004	0.629	0.956
	30	0.784	0.796	0.003	0.626	0.894
	50	0.732	0.738	0.002	0.618	0.841
	70	0.66	0.663	0.002	0.544	0.743
Southern → Center	10	0.781	0.786	0.009	0.468	0.943
	30	0.743	0.744	0.005	0.513	0.872
	50	0.717	0.721	0.002	0.61	0.808
	70	0.675	0.679	0.001	0.595	0.757

% Unbalance level, average correlation $r_{g\hat{g}}^{A}$, median correlation $r_{g\hat{g}}^{M}$, lowest correlation $r_{g\hat{g}}^{L}$, and largest correlation $r_{g\hat{g}}^{H}$ observed across 100 unbalanced simulation.

Data from both the southern and central regions were considered for the joint analysis. A total of 273 hybrid combinations were evaluated. The correlation between the estimated GBVs of the common hybrids within the southern and central regions was calculated (0.55) to test the effect of the genotype x environment interaction on the predictions. Figure 1A shows the regression between the calculated GBVs of the southern and central regions, with a fit of 0.3. This median correlation and consequent low-fit of the regression line reflected the interaction of the hybrids between the experiments, which were in the entirely contrasting regions. This interaction resulted in the genotypes having different responses to each site.

It is noteworthy that in addition to analyzing the hybrids at different locations, they were also analyzed in different crop seasons, which allowed for the obtained correlation to be considered satisfactory against the genotype x environment interactions. Interestingly, when ranking the best and worst 20% hybrids, it could be observed that by using the GBVs calculated for the central region, it was possible to correctly select the best 58% GBVs calculated within the southern region, where the opposite was true (considering the 186 common hybrids). There would be an error of 5% among the worst GBVs calculated, i.e., only two of the hybrids that had higher GBVs would be mistakenly discarded when they were present in another region. Thus, it is possible to affirm that the method is better for discarding the less promising hybrids than for selecting the most promising ones.

In the joint analysis, there were medium- to high-correlation values between the predicted and observed GBVs. At the highest imbalance level, the correlation between them decreased. At 10% imbalance, higher correlation estimates were obtained, with a mean correlation of 0.66 and a range of 0.30 to 0.90. At 30% imbalance, there was a mean correlation of approximately 0.60 and a range of 0.33 to 0.74. At 50% imbalance, there was a mean correlation of approximately 0.56 and a range of 0.19 to 0.71. As already expected, at 70% imbalance, there was a lower mean correlation (0.56) and a range of 0.19 to 0.71 (Table 3).

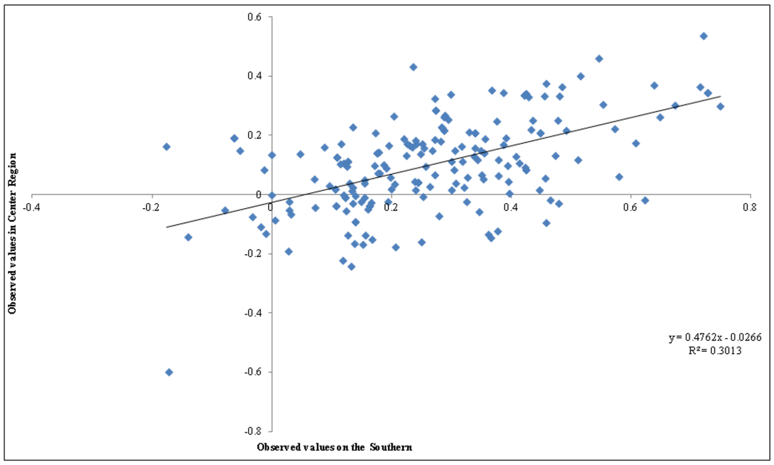

Figure 1. Regression of the genomic breeding values (GBVs) calculated using the total effects of associated markers (TEAM) method for the experiments in the southern and central regions.

The second portion of the study focused on using the GBVs obtained within one of the regions to predict the GBVs for the other region (Table 3). In this scenario, it was also observed that with more data to be predicted, i.e., a reduced training population, the lowest correlations values were observed. Interestingly, the correlation observed in different macroregions was equivalent to that obtained when predictions were made within the southern region or even when the joint analysis was performed with the data. This indicates that in the present study, the training population (using the data from the central region) was able to predict the southern data better, compared to the phenotype of the southern region itself. This conformed to the expectations, given the specific correlations of each macroregion.

When data of the central region were used to predict the southern region data, the mean correlation ranged from 0.85 to 0.66. Furthermore, the highest correlation was observed at 10% imbalance and the lowest was observed at 70% imbalance of the predicted data. However, when the data of the southern region were used to predict the GBVs of the central region at 10% imbalance, the observed mean was 0.78, with values ranging from 0.47 to 0.94. At 30 and 50% imbalance, the means were 0.74 and 0.72, with minimum values of 0.51 and 0.61 and maximum values of 0.87 and 0.81, respectively, whereas at 70% imbalance, the mean was 0.67, with the correlations ranging between 0.60 and 0.76. There was a decline in the correlations accompanied by a decline in the quantity of data used for prediction.

The frequency distributions of the correlations are shown in the histograms in Figure 2.

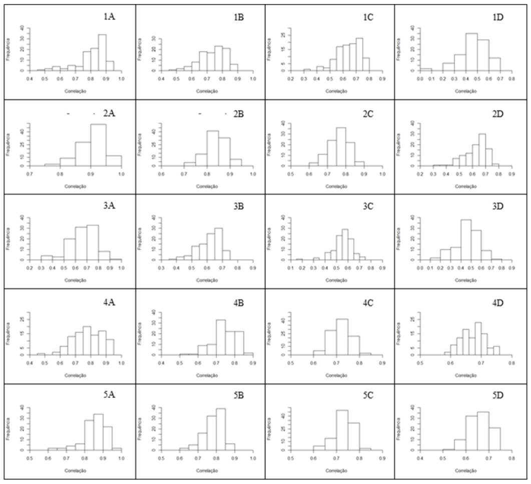

Figure 2. Frequency histograms of the correlations between the predicted and observed genomic breeding values; the numbers refer to the sites and the letters refer to the imbalance levels (1: southern region; 2: central region; 3: joint analysis; 4: southern region for predicting the central region; and 5: central region for predicting the southern region; **A.** 10% imbalance; **B.** 30% imbalance; **C.** 50% imbalance; and **D.** 70% imbalance).

The correlations tend to remain more asymmetrical as the imbalance increased. This was expected, considering that the model had greater difficulty in accurate predictions as more hybrids were removed from the training population. However, the ideal day-to-day model of a breeding program can accurately predict a large number of crosses that have not yet been evaluated in field trials, even with a small training population.

Figure 3 summarizes the results obtained in the present study. Among all the analyses, the joint analysis exhibited the lowest correlations at all the imbalance levels. This is most likely because of the interaction between the genotypes, sites, and crop seasons. It is also apparent that the decline in the correlation was more evident after 50% imbalance, in which the southern region exhibited the highest losses with high imbalance levels, and the cross predictions (central/southern and southern/central) were less affected.

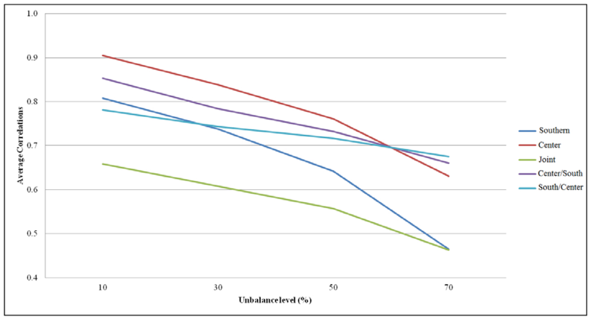

Figure 3. Relationship between all the performed analyses: southern region, central region, joint analysis, central region for predicting the southern region (central/southern), and southern region for predicting the central region (southern/central).

DISCUSSION

As reported in literature, heritability was observed in maize yield experiments over multiple years and sites, in both the southern and central regions (Lorenzana and Bernardo, 2009; Combs and Bernardo, 2013). It is noteworthy that the experiments were conducted at very distinct locations that exhibit very characteristic microclimates despite being located in the same macroregion. Thus, there was an effect of the genotype x environment interaction on the heritability.

Within the southern region, the effect of this interaction was further noticeable; heritability was lower in the southern region than in the central region, together with a greater effect of the interaction. There was also a higher proportion of the residual variance compared to the genotypic variance in the southern region.

The genotype x environment interaction is the differential response of the genotypes to environments, and this interaction represents a major challenge in genomic prediction (Heslot et al., 2015). The interaction can modify the genotype ranking according to each environment. Thus, the same genotype cannot have the best performance at all the sites. Genomic prediction can help improve the interaction because even if an individual was not tested in a particular environment, prediction is possible using their relative information (Heslot et al., 2015).

The correlation values observed in the present study, even under such contrasting conditions, were higher than several studies reported in the scientific literature. By using BLUP to predict the performance of simple maize hybrids based on information of the parental strains and RFLP markers in a population of 54 simple hybrids at different imbalance levels ranging from 10 to 30%, Bernardo (1994) achieved medium to high correlations (0.65 to 0.80). Thus, the results obtained in the present study are more encouraging than those reported by Bernardo (1994), even using higher imbalance levels and fewer markers.

Gains in the precision of the genomic selection may be affected by three factors: i) the proportion of the training population, ii) marker density, and iii) heritability (Guo et al., 2012). Daetwyler et al. (2008) expressed prediction accuracy as a function of the training population size (N), heritability (h^2), and the number of chromosomal segments affecting the trait (Me), as shown in the equation below.

$$r_{g\hat{g}} = \sqrt{Nh^2 / (Nh^2 + Me)}$$

Thus, the superior results found in the central region data set can be explained as the combined effect of both higher heritability and more individuals in the training population, as observed by the higher number of hybrids in the data set.

Because of the higher heritability observed within the central region, its correlation was higher than the correlations found in all the other analyses, considering all the imbalance levels. Several studies report heritability as an important point to increase in the correlations (Guo et al., 2012). For example, Guo et al. (2012), who studied the effect of heritability on the prediction accuracy with different statistical models and training population sizes, found that the prediction accuracy for the RR-BLUP model tend to increase with the increasing heritability in all proportions of the training population.

Population size is another important point when observing the higher estimates in the central region compared to those in the southern region. The population size was approximately 30% lower within the southern region. Cross-validation studies indicate an increased correlation with an increasing population size or the number of instances (Guo et al., 2012; Combs and Bernardo, 2013). Lorenzana and Bernardo (2009), who investigated the prediction of genomic breeding values for maize yield using the BLUP method and a constant number of markers, found that the correlation tends to decrease with decreasing numbers of progeny. The authors reported an approximate 40% reduction in the prediction accuracy when the number of progeny decreased from 178 to 95, with a constant number of markers.

It is also noteworthy that nine sites were used for the analyses within the central region whereas six sites were used within the southern region. Guo et al. (2013b) and Burgueño et al. (2012) also found increased correlation in multi-environment analyses. This may be explained by using information among related genotypes throughout the different environments and the same genotype among environments using the genetic and environmental covariance (Burgueño et al., 2012). It can also be explained by more precise estimates of the specific effects of the markers in each environment using genetic correlation (Guo et al., 2013b).

An important component of the genetic improvement programs is the evaluation of genotypes at locations with different environmental conditions. Thus, the studied genotype x environment interaction (GEI) can be compared more accurately and the best genotypes within and between the environments can be selected with higher reliability (Crossa et al., 2010). In this sense, prediction can be useful because it can use the hybrid data at one location to predict data at another entirely different location. However, a large portion of the studies on genomic prediction focused on the same site/crop season (Guo et al., 2013b), and in these cases, the interaction was capitalized for the prediction - which is not always possible to obtain in practice. In the present study, the data of one macroregion were used to predict the behavior of the same genotype in another macroregion with a very distinct characteristic climate. This was possible because southern and central Brazil have predominantly subtropical and tropical climates, respectively. This information is extremely important in a breeder's decision-making process, because it creates the potential for resource efficiency by avoiding planting unpromising genotypes at distinct locations.

In the joint analysis, data from both the southern and central regions were used to perform the prediction. It is noteworthy that in this case, imbalance was performed in the data set as a whole. The values were lower than those obtained in separate analyses of the regions. This was because the hybrids were planted at entirely discrepant regions and in different crop seasons and the training population comprised both the regions. Moreover, it was previously highlighted that the GEI among these regions was high and had a low correlation between the GBVs. Although multi-environment analyses usually tend to increase the correlation, it is noteworthy that this increase depends on the genetic correlation between the environments (Guo et al., 2013b).

In the second portion of the study, the southern region data set was used to predict the common genotypes tested within the central region and vice versa. In other words, training in the southern region and the performance of the predictions in the central region (southern/central) were evaluated and vice versa (central/southern). The results observed were encouraging, considering the large contrast between the sites and the crop seasons. When the southern region data were used to predict the central region data, the correlations were always greater than 0.68, even at the higher imbalance levels. Higher correlations were also observed when the central region data were used to predict those of the southern region, with means ranging from 0.85 to 0.66. The values were higher in the central region than in the southern region because of the higher quantity of information used and the larger data set, as previously mentioned. In this case, it is evident that the predictions for the central region were more informative than the data used in the training of the southern region.

Compared to the results obtained in other studies using TEAM, the correlations observed in the present study are encouraging, considering the training population size, the number of markers used, and the environmental differences between the sites and crop seasons. Schrag et al. (2009, 2010) obtained correlation values ranging from 0.16 to 0.65 using 50% imbalance in a set of 400 tested hybrids and more than 1000 molecular markers. Using less than 10% of the markers and approximately half of the hybrids in the training population, the results of the present study achieved mean correlations ranging between 0.56 and 0.76 for the same imbalance level.

Additionally, compared to the study by Massman et al. (2013b), who used cross-validation and obtained an accuracy ranging from 0.75 to 0.87 for 10% imbalance, our results exhibited similar mean correlations (0.66 to 0.91) for the same level of loss.

With more modern prediction methods, such as GBLUP and Bayes B, Technow et al. (2014) used a set of 1254 hybrids resulting from crosses between 'dent' and 'flint' heterotic groups over several years and sites, with a set of approximately 35,000 SNP markers. They observed correlations in the cross-validations ranging from 0.75 to 0.92 for kernel productivity, which were similar to the results obtained in the present study. In the study by Technow et al. (2014), the correlation mostly varied between the number of common parents in the training population and in the validation population.

The present study obtained high correlations even when using fewer markers compared to the current studies on genomic prediction. One explanation for this is that the current studies use SNP markers, which are bi-allelic, whereas SSR markers are multi-allelic, which allows them to provide the maximum number of alleles per marker (Lu et al., 2009) and, thus, deliver a higher level of information. According to Laval et al. (2002), (*k-1*) times bi-allelic markers are necessary to achieve the same genetic information obtained through a set of SSR markers with *k* alleles. If this proposition is true, approximately 550 SNPs would be necessary to obtain the same information rendered by the 79 SSRs used in the present study. In the studies on genetic distance, it has been demonstrated that on an average, the number of alleles detected by one SSR and ten SNP markers is same (Yan et al., 2010).

It has been reported that a marker density of 10-20 cM, which corresponds to approximately 150 markers, is sufficient to obtain good prediction estimates in bi-parental maize populations (Lian et al., 2014). Marker density marginally affects the prediction accuracy for the performance of hybrids with complex traits. However, the accuracy reaches a plateau with the use of a few hundred markers for bi-parental populations (Zhao et al., 2013). For example, Zhang et al. (2015) observed good predictive abilities in bi-parental maize populations for moderately- to highly-heritable traits, by using only 200 SNP markers.

The method described in the present study, which obtained high correlations despite using a restricted number of markers, can be particularly useful for small-scale breeding companies, public universities, and research centers, which often have limited resources for implementing selection in genetic improvement programs (Zhao et al., 2015).

It is noteworthy that only the additive-dominant model was considered in the present study and that epistasis was not used to simplify the model. Theoretically, including epistasis in genomic prediction may increase the prediction accuracy because it is important for the trait and can be modeled accurately (Lorenzana and Bernardo, 2009). However, in practice, both simulation and field data studies have not shown advantages or even exhibited reduced prediction accuracy when the epistatic effects were added to the model (Lee et al., 2008; Lorenzana and Bernardo, 2009).

Hybrid prediction, particularly of maize, must be further explored. The increase in the number of markers and individuals in the training population can be decisive for validating the technique. However, it was evident that validation, using trials over several years, in distinct sites and under different crop conditions adopted in the training populations may provide breeders with a broader view about the efficacy of this technique.

The continued advances in genotyping techniques and their reduced costs, statistical models, and breeding methods based on genomic selection could contribute to genetic improvement programs becoming more efficient. The increasing adoption of the double-haploid technique and increased phenotyping costs tend to further restrict evaluation of hybrid combinations in field experiments. It is believed that genomic prediction will be increasingly adopted in the genetic improvement programs in the coming years.

The TEAM method was efficient in predicting the GBVs related to yield in simple maize hybrids at different sites and during different crop seasons. Additionally, there were satisfactory correlations of the GBVs calculated under different environments.

The results of the present study suggest that validation under different farming conditions is possible, and the cross-validation results strongly demonstrate the real performance in the field.

The annual genetic gains can be increased by identifying the promising genotypes by genomic prediction. The results obtained in the present study highlight the necessity of further research for implementing genomic prediction in breeding programs.

Conflicts of interest

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

Research supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

Balestre M, Machado JC, Lima JL, Souza JC, et al. (2008). Genetic distance estimates among single cross hybrids and correlation with specific combining ability and yield in corn double cross hybrids. *Genet. Mol. Res.* 7: 65-73. http://dx.doi.org/10.4238/vol7-1gmr403

Bernardo R (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci.* 34: 20-25. http://dx.doi.org/10.2135/cropsci1994.0011183X003400010003x

Bernardo R (1996). Testcross additive and dominance effects in best linear unbiased prediction of maize single-cross performance. *Theor. Appl. Genet.* 93: 1098-1102. http://dx.doi.org/10.1007/BF00230131

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

Charcosset A and Essioux L (1994). The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. *Theor. Appl. Genet.* 89: 336-343.

Combs E and Bernardo R (2013). Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *Plant Genome* 6: 1-7. http://dx.doi.org/10.3835/plantgenome2012.11.0030

Crossa J, Campos GdeL, Pérez P, Gianola D, et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186: 713-724. http://dx.doi.org/10.1534/genetics.110.118521

Daetwyler HD, Villanueva B and Woolliams JA (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3: e3395. http://dx.doi.org/10.1371/journal.pone.0003395

Devi P and Singh NK (2011). Heterosis, molecular diversity, combining ability and their interrelationships in short duration maize (Zea mays L.) across the environments. *Euphytica* 178: 71-81. http://dx.doi.org/10.1007/s10681-010-0271-3

Dhliwayo T, Pixley K, Menkir A and Warburton M (2009). Combining ability, genetic distances, and heterosis among elite cimmyt and iita tropical maize inbred lines. *Crop Sci.* 49: 1201-1210. http://dx.doi.org/10.2135/cropsci2008.06.0354

Fernando RL and Grossman M (1989). Marker-assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.* 21: 467-477. http://dx.doi.org/10.1186/1297-9686-21-4-467

Ferreira DV, Von Pinho RG, Balestre M and Oliveira RL (2010). Prediction of maize hybrid performance using similarity in state and similarity by descent information. *Genet. Mol. Res.* 9: 2381-2394. http://dx.doi.org/10.4238/vol9-4gmr955

Fu J, Falke KC, Thiemann A, Schrag TA, et al. (2012). Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. *Theor. Appl. Genet.* 124: 825-833. http://dx.doi.org/10.1007/s00122-011-1747-9

Gardner CO and Eberhart AS (1966). Analysis and interpretation of the variety cross diallel and related populations. *Biometrics* 22: 439-452. http://dx.doi.org/10.2307/2528181

Guo T, Li H, Yan J, Tang J, et al. (2013a). Performance prediction of F1 hybrids between recombinant inbred lines derived from two elite maize inbred lines. *Theor. Appl. Genet.* 126: 189-201. http://dx.doi.org/10.1007/s00122-012-1973-9

Guo Z, Tucker DM, Lu J, Kishore V, et al. (2012). Evaluation of genome-wide selection efficiency in maize nested association mapping populations. *Theor. Appl. Genet.* 124: 261-275. http://dx.doi.org/10.1007/s00122-011-1702-9

Guo Z, Tucker DM, Wang D, Basten CJ, et al. (2013b). Accuracy of across-environment genome-wide prediction in maize nested association mapping populations. *G3 (Bethesda)* 3: 263-272. http://dx.doi.org/10.1534/g3.112.005066

Henderson CR (1984). Applications of linear models in animal breeding. 3rd edn. (Schaeffer LR, eds.). CGIL publications, University of Guelph, Guelph.

Heslot N, Jannink JL and Sorrells ME (2015). Perspectives for genomic selection applications and research in plants. *Crop Sci.* 55: 1-12. http://dx.doi.org/10.2135/cropsci2014.03.0249

Laval G, SanCristobal M and Chevalet C (2002). Measuring genetic distances between breeds: use of some distances in various short term evolution models. *Genet. Sel. Evol.* 34: 481-507. http://dx.doi.org/10.1186/1297-9686-34-4-481

Lee SH, van der Werf JH, Hayes BJ, Goddard ME, et al. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* 4: e1000231. http://dx.doi.org/10.1371/journal.pgen.1000231

Lian L, Jacobson A, Zhong S and Bernardo R (2014). Genomewide prediction accuracy within 969 maize biparental populations. *Crop Sci.* 54: 1514-1522. http://dx.doi.org/10.2135/cropsci2013.12.0856

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

Lu Y, Yan J, Guimarães CT, Taba S, et al. (2009). Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. *Theor. Appl. Genet.* 120: 93-115. http://dx.doi.org/10.1007/s00122-009-1162-7

Massman JM, Jung HJG and Bernardo R (2013a). Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Sci.* 53: 58-66. http://dx.doi.org/10.2135/cropsci2012.02.0112

Massman JM, Gordillo A, Lorenzana RE and Bernardo R (2013b). Genomewide predictions from maize single-cross data. *Theor. Appl. Genet.* 126: 13-22. http://dx.doi.org/10.1007/s00122-012-1955-y

Melchinger AE (1999). Genetic diversity and heterosis. In: The genetics and exploitation of heterosis in crops (Coors JG and Pandey S, eds.). ASA, CSSA and SSSA, Madison, WI, 99-118.

Melchinger AE, Geiger HH, Seitz G and Schmidt GA (1987). Optimum prediction of three-way crosses from single crosses in forage maize (Zea mays L.). *Theor. Appl. Genet.* 74: 339-345. http://dx.doi.org/10.1007/BF00274716

Melo WMC, Balestre M, Von Pinho RG and Bueno-Filho JSS (2014). Genetic control of the performance of maize hybrids using complex pedigrees and microsatellite markers. *Euphytica* 195: 331-344. http://dx.doi.org/10.1007/s10681-013-0999-7

Meuwissen TH, Hayes BJ and Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819-1829.

Schrag TA, Melchinger AE, Sørensen AP and Frisch M (2006). Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. *Theor. Appl. Genet.* 113: 1037-1047. http://dx.doi.org/10.1007/s00122-006-0363-6

Schrag TA, Maurer HP, Melchinger AE, Piepho HP, et al. (2007). Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. *Theor. Appl. Genet.* 114: 1345-1355. http://dx.doi.org/10.1007/s00122-007-0521-5

Schrag TA, Möhring J, Maurer HP, Dhillon BS, et al. (2009). Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. *Theor. Appl. Genet.* 118: 741-751. http://dx.doi.org/10.1007/s00122-008-0934-9

Schrag TA, Möhring J, Melchinger AE, Kusterer B, et al. (2010). Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. *Theor. Appl. Genet.* 120: 451-461. http://dx.doi.org/10.1007/s00122-009-1208-x

Technow F, Schrag TA, Schipprack W, Bauer E, et al. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* 197: 1343-1355. http://dx.doi.org/10.1534/genetics.114.165860

Walsh B (2014). Special issues on advances in quantitative genetics: introduction. *Heredity (Edinb)* 112: 1-3. http://dx.doi.org/10.1038/hdy.2013.115

Yan J, Yang X, Shah T, Sánchez-Villeda H, et al. (2010). High-throughput SNP genotyping with the GoldenGate assay in maize. *Mol. Breed.* 25: 441-451. http://dx.doi.org/10.1007/s11032-009-9343-2

Zhang X, Pérez-Rodríguez P, Semagn K, Beyene Y, et al. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity (Edinb)* 114: 291-299. http://dx.doi.org/10.1038/hdy.2014.99

Zhao Y, Gowda M, Liu W, Weurschum T, et al. (2013). Choice of shrinkage parameter and prediction of genomic breeding values in elite maize breeding populations. *Plant Breed.* 132: 99-106. http://dx.doi.org/10.1111/pbr.12008

Zhao Y, Mette MF and Reif JC (2015). Genomic selection in hybrid breeding. *Plant Breed.* 134: 1-10. http://dx.doi.org/10.1111/pbr.12231