Potential use of molecular markers for prediction of genotypic values in hybrid maize performance

M. Balestre[1], R.G. Von Pinho[2], J.C. Souza[1] and R.L. Oliveira[1]

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

Corresponding author: M. Balestre
E-mail: marciobalestre@hotmail.com

ABSTRACT. We evaluated the potential of genetic distances estimated by microsatellite markers for the prediction of the performance of single-cross maize hybrids. We also examined the potential of molecular markers for the prediction of genotypic values and the applicability of the Monte Carlo method for a correlation of genetic distances and grain yield. Ninety $S_{0:2}$ progenies derived from three single-cross hybrids were analyzed. All 90 progenies were genotyped with 25 microsatellite markers, including nine markers linked to quantitative trait loci for grain yield. The genetic similarity datasets were used for constructing additive genetic and dominance matrices that were subsequently used to obtain the best linear unbiased prediction of specific combining ability and general combining ability. The genetic similarities were also correlated with grain yield, specific combining ability and heterosis of the hybrids. Genetic distances had moderate predictive ability for grain yield (0.546), specific combining ability (0.567) and heterosis (0.661). The Monte Carlo simulation was found to be a viable alternative for a correlation of genetic distances and grain yield. The accuracy of genotypic values using molecular data information was slightly higher than if no such information was incorporated. The estimation of the

relationship using molecular markers proved to be a promising method for predicting genetic values using mixed linear models, especially when information about pedigree is unavailable.

Key words: BLUP; Relationship; Partial diallel; Genetic accuracy

INTRODUCTION

The evaluation and selection of inbred lines to obtain single-cross hybrids is a common process and one of the major objectives in maize breeding programs. However, the testing stage can be considered the most expensive, since it requires a greater amount of resources, even when few inbred lines are crossed. This is because performance *per se* of the inbred lines does not always assure superior hybrids, and it is then necessary to evaluate these inbred lines in hybrid combinations (Hallauer et al., 1988; Troyer, 1994).

Diallel crosses can be used to evaluate the inbred lines' general combining ability (GCA) and specific combining ability (SCA). However, due to the great number of crosses that can be obtained with few genotypes, this cross design is little used in the initial stages of breeding programs.

One alternative for this large number of genotypes in the early evaluation stages is the use of "top crosses". However, according to Araújo (2001), this method is less accurate for GCA estimates in relation to complete and partial diallels; moreover, it is not possible to obtain SCA estimates between the inbred lines, which help the breeder in the selection of superior genotypes.

Another alternative is a prior selection, or screening, based on genetic distances of partially endogamic lines to cross only the promising genotypes in diallel crosses. The advantages of performing a diallel in partial inbred lines are due to the exploration of hybrid heterosis, which can be used directly for the production of hybrids prior to these lines, reaching a high endogamy level, such as with lines S_6 or S_7 (Carlone and Russel, 1988; Amorim et al., 2006). Moreover, there is a high correlation between progeny performances in early and advanced generations, making it possible to use diallel crosses with partial inbred lines for prior selections of the elite single-cross hybrids (Bernardo, 1991).

According to Falconer and Mackay (1996), based on the quantitative genetics theory, there is a linear relationship between genetic distance, in the presence of loci with dominance, and heterosis. Thus, some studies have been conducted to predict heterosis, combining ability and grain yield on the basis of parental genetic distance, as pointed out by Dias et al. (2004). Moreover, several types of germplasm have been used from inbred and partial inbred lines to single-cross hybrids (Amorim et al., 2006; Guimarães et al., 2007; Balestre et al., 2008a). However, the results have demonstrated little efficacy and repeatability (Dias et al., 2004).

Several factors can contribute to low efficacy in prediction performance from genetic distances using molecular markers. Among these factors, Bernardo (1992) highlights the presence of a small dominance effect on quantitative trait loci, low heritability, markers not linked to quantitative trait loci, or more than 70% of these markers not linked to quantitative loci, besides the great variation in allele frequency of the progenitors. Moreover, the choice of an adequate genetic distance coefficient for the datasets and study objectives can contribute to the magnitude of the correlation between genetic distance and

heterosis (Reif et al., 2005; Balestre et al., 2008b).

Correlation studies between genetic distance and grain yield commonly use hypothesis tests, such as the Student *t*-test, which are not robust for data with strong deviation from normality or sample skewness. Data asymmetry can occur with genetic distance estimates (due to dependence on the data), leading to violations of the test assumptions and increasing the rate of type I and type II errors, i.e., rejecting a true hypothesis or accepting a false one, respectively. Thus, alternative tests for the *t*-test have been proposed for correlation studies by non-parametric statistics, such as the Mantel Z-test (Mantel, 1967). However, several studies have suggested the empirical correlation test between variables using Monte Carlo simulation (May et al., 2001; Fan et al., 2002).

This study evaluated the predictive potential of genetic distances estimated by microsatellite markers in the performance of hybrids from $S_{0:2}$ lines, the potential of molecular markers in predicting the accuracy of genotypic values, and the applicability of the Monte Carlo method in the correlation study of genetic distance and grain yield.

MATERIAL AND METHODS

Material description

Ninety $S_{0:2}$ progenies from three populations originating from three single-cross hybrids were used. There were 30 progenies from population A (hybrid P30F45), 31 from population B (hybrid Dow 657) and 29 from population C (hybrid DKB333B).

All 90 progenies were genotyped with 25 microsatellite markers (simple sequence repeats), with 9 markers linked to quantitative trait loci for grain yield. The others were selected according to the results from "bootstrap" analysis for those same populations (Amorim et al., 2006). The extraction of DNA, reaction preparation and estimates of genetic distances were described earlier by Balestre et al. (2008b). Based on the results obtained by these authors, Roger's modified distance (RMD) was used for the selection of 17 partial inbred lines with the objective of recovering the five most and least divergent combinations. Roger's modified distance is the best suited expression for this type of analysis because it provides a coefficient recommended for the make up of heterotic groups (Reif et al., 2005), besides showing acceptable values of distortion and stress for these populations (Balestre et al., 2008b).

A partial diallel (11 x 6) was performed with the 17 selected partial inbred lines where, within each group, the lines were used to obtain the crosses between the most and least divergent ones.

Experimental design and correction of averages by mixed models

Due to the small number of seeds, 60 were evaluated together with four checks and were evaluated in two locations in the county of Lavras (Campus of the Universidade Federal de Lavras and Vitorinha Farm), in the agricultural year of 2007/2008. One of the checks was the hybrid DKB199, and the other three checks were single-cross hybrids that originated populations A (P30F45), B (Dow 657) and C (DKB333B). The experimental design used was an 8 x 8 lattice with three repetitions. Each plot consisted of two 3-m long rows, fertilized with 450 kg/ha NPK (8-28-16) at sowing and 100 kg/ha nitrogen as side dressing.

The linear mixed model in the matrix form used for data analysis, considering recovery in inter-block information, was expressed as:

$$y = X\beta + Z_1 a + Z_2 g + \delta$$ **(Equation 1)**

where y is the vector of observed data, β is the vector of fixed effects (averages of repetitions), a is the random vector of genotypic values, g is the random vector of block effects within the repetitions, δ is the random error and X, Z_1 and Z_2 are the incidence matrices related to β, a and g, respectively.

Diallel analysis by mixed models

The joint diallel analysis, using method IV proposed by Griffing (1956), was performed with the data corrected for each location. The analyses were done with the SAS® PROC IML system (SAS Institute, 2000). The estimates of fixed effects and best lenear unbiased predictions (BLUPs) of combining abilities (SCA and GCA) were obtained as presented by Bernardo's (1995) proposition for partial diallel. However, in this study the complete diallel model was used, considering all genotypes within the same group.

The linear model considered is given by:

$$y = X\beta + Z_1 g + Z_2 s + e$$ **(Equation 2)**

where y is the vector of the observations, β is the vector of fixed effects (overall mean and local ones), g is the vector of effects of GCA, s is the vector of SCA, e is the vector of errors, and X, Z_1 and Z_2 are the incidence matrices of effects β, g and s, respectively.

The joint solution of the fixed and random effects was obtained by the equations system, according to Henderson (1984):

$$\begin{bmatrix} \beta^0 \\ \hat{g} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} X'X & X'Z_1 & X'Z_2 \\ Z_1'X & Z_1 Z_1 + A_1^{-1}\gamma_1 & Z_1'Z_2 \\ Z_2'X & Z_2'Z_1 & Z_2'Z_2 + A_2^{-1}\gamma_2 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z_1'y \\ Z_2'y \end{bmatrix}$$ **(Equation 3)**

with $\gamma_1 = \frac{\sigma_e^2}{\sigma_{GCA}^2}$ $\gamma_2 = \frac{\sigma_e^2}{\sigma_{SCA}^2}$, where: A_l is the additive genetic similarity matrix among the partial inbred lines. The additive LR relationship matrix was obtained according to Lynch and Ritland (1999) by the following expression:

$$\hat{r}_{xy(k)} = \frac{p_a(S_{bc} + S_{bd}) + p_b(S_{ac} + S_{ad}) - 4p_a p_b}{(1 + S_{ab})(p_a + p_b) + 2p_a p_b}$$ **(Equation 4)**

where:

$\hat{r}_{xy(k)}$ is the additive relationship estimator between individual X (taken as reference) with alleles a and b and individual Y with alleles c and d in locus k;

S_{ab} : equals to 1 if a and b are identical, and equals 0 if otherwise;

S_{ac}: equals to 1 if a and c are identical, and equals 0 if otherwise;

S_{ad}: equals to 1 if a and d are identical, and equals 0 if otherwise;

S_{bc}: equals to 1 if b and c are identical, and equals 0 if otherwise;

S_{bd}: equals to 1 if b and d are identical, and equals 0 if otherwise;

p_a and p_b are the frequencies of alleles a and b through the lines for a given locus k.

Considering the relationship estimates are obtained along many loci, the use of weights for the estimates is suggested, besides reciprocal estimation of the $\hat{r}_{xy(k)}$, i.e., initially, the individual X is taken as reference and later the individual Y (Lynch and Ritland, 1999). The final expression, considering all loci, the given weights for each locus, and the estimated relationship, is given by:

$$\hat{r}_{xy} = \frac{\sum_{1}^{L} w_{r,x(k)} \hat{r}_{x,y(k)} + \sum_{1}^{L} w_{r,y(k)} \hat{r}_{y,x(k)}}{W_{r,x} + W_{r,y}} \qquad \textbf{(Equation 5)}$$

where $w_{r,x(k)}$ and $w_{r,y(k)}$ are the weights for the kth locus of estimates of $\hat{r}_{xy(k)}$; $W_{r,x}$ and $W_{r,y}$ are the sum of weights attributed to all loci involved.

The dominance relationship matrix (A_2) was constructed according to Henderson (1984) and Van Vleck (1993).

Also, Roger's distances were used to make the matrices A_1 and A_2 using the expression $s_{ij} = 1 - d_{ij}$, where d_{ij} is the RMD value. Thus, two similarity estimates were compared to determine the effect of the use of similarity-in-state (Roger's distances) and similarity-by-descent (LR), in the prediction accuracy of GCA and SCA values. Accuracy estimates of GCA and SCA were obtained by:

$$r_{GCA} = \sqrt{\left(1 - d_i \gamma_1\right)} \quad r_{SCA} = \sqrt{\left(1 - d_i \gamma_2\right)} \qquad \textbf{(Equation 6)}$$

where d_i is the ith element of the diagonal of $Z_1 Z_1 + A_1^{-1} \gamma_1$ and $Z_2' Z_2 + A_2^{-1} \gamma_2$ resulting from the inverse of the joint matrix of fixed and random effects.

Correlation study between RMD with grain yield, SCA and heterosis

The study of Pearson's correlation was carried out using the variables unhusked ear yield (HEY), SCA, and heterosis with RMD. The analyses were performed with SAS®, using PROC CORR. The Mantel test was used to study the significance of the correlations (Manly, 1991). Heterosis values were estimated based on the single-cross hybrids that originated the populations A, B and C, as suggested by Amorim et al. (2006).

Besides the Mantel test, the correlation significance between RMD and HEY was tested by Monte Carlo simulation, which used the algorithms recommended by Fan et al. (2002). These authors suggested elaborating the data, fitted to the characteristics of a common sample, to observe how much the correlation coefficient can vary when several different-sized samples are removed from the group of data.

Basically, the algorithm consists in: a) obtaining the first, second, third, and fourth moments of the sample, i.e., the average, variance, skewness and kurtosis; b) observing if there is any degree of dependence between the samples from which inferences will be made (correlation), and c) generating data fitted to the sample characteristics (Table 1).

Table 1. Moments referring to genetic distance and yield of 60 maize hybrids, and group of fitted and simulated data using the Monte Carlo method based on sample data.

Sample data	N	Average	Standard deviation	Skewness	Kurtosis
RMD	60	0.169	0.101	-0.332	-1.098
Yield	60	8.962	1.529	*	*
Correlation	0.546				
Simulated data	N	Average	Standard deviation	Skewness	Kurtosis
RMD	2,000,000	0.169	0.102	-0.345	-0.975
Yield	2,000,000	8.958	1.528	*	*
Correlation	0.547				

*Data with normal distribution by the Shapiro-Wilk test. RMD = Roger's modified distance.

The algorithm proposed by Fleishman (1978) was used to obtain random data with the desired skewness and kurtosis, and that proposed by Kaiser and Dickman (1962) for the data with the desired dependence (correlation).

Four different sample sizes were used (10, 30, 60, and 100). Each sample was repeated 10,000 times, for a total of 40,000 samples with 2 million pieces of data. The probability of $r > 0$ was tested by the expression proposed by May et al. (2001) and Fan et al. (2002). Moreover, a confidence interval was calculated for the correlation value between the variable HEY and RMD for a sample size of 60, which is equal to that obtained in this experiment.

RESULTS

BLUPs of HEY for each hybrid and the four checks are found in Table 2. Three crosses were superior to the check with the greatest yield (DKB199). The interpopulation hybrid 1BX28C was 1.66 t/ha more productive in terms of HEY than the hybrid DKB199. Moreover, six crosses had better performance in comparison to the hybrid DKB333B, which was one of the hybrids used to form one of the three base populations (population C) and which showed the best performance in relation to the other hybrids (DOW657 and P30F45).

Intrapopulation hybrids as productive as interpopulation ones were found (Table 2). However, the greatest yields were found in interpopulation hybrids, and together, their average performance was 9.84 t/ha versus the intrapopulation average yield, which was 8.24 t/ha.

Only 30% of the interpopulation hybrids showed negative heterosis, while all intrapopulation ones had performance below that of the genitor single-cross hybrid, reflecting negative heterosis estimates. This result demonstrates that obtaining interpopulation hybrids from $S_{0:2}$ progenies is extremely viable, since 70% of the interpopulation hybrids surpassed the average performance of the single-cross hybrids used in the formation of the base population (Table 2). It is important to note that the heterosis applied in this study does not reflect the classic concept of heterosis, since single-cross hybrids that originated the populations were used as reference instead of the $S_{0:2}$ lines.

Table 2. BLUPs of unhusked ear yield (t/ha) and heterosis of intra- and interpopulation hybrids evaluated in two locations in the agricultural year 2007-2008.

Genotype	Yield	Heterosis	Genotype	Yield	Heterosis
1BX28C	12.375	2.998	19CX25C	8.713	-1.710
14CX29A	11.222	1.300	15CX7C	8.712	-1.711
30BX28C	10.813	1.436	12CX25C	8.707	-1.716
*DKB199	10.715	-	19CX23C	8.695	-1.728
30BX23C	10.690	1.313	22CX23C	8.583	-1.840
1BX7C	10.649	1.272	30BX29A	8.579	-0.296
29CX28B	10.628	1.251	12CX28C	8.560	-1.863
*DKB333B	10.423	-	15CX29A	8.545	-1.377
30BX7C	10.412	1.035	*DOW 657	8.331	-
12CX28B	10.264	0.887	19CX7C	8.511	-1.912
15CX28B	10.194	0.817	12CX23C	8.467	-1.956
5CX29A	10.029	0.107	14CX23C	8.464	-1.959
27BX28C	9.942	0.565	1CX29A	8.434	-1.488
27BX23C	9.928	0.551	5CX25C	8.301	-2.122
22CX29A	9.894	-0.027	5CX7C	8.238	-2.185
5CX28B	9.894	0.517	12CX7C	8.167	-2.256
14CX28B	9.879	0.502	14CX28C	8.152	-2.271
19CX28B	9.745	0.368	19CX28C	8.142	-2.281
27BX7C	9.653	0.276	5CX23C	8.134	-2.289
1BX23C	9.631	0.254	1BX25C	8.091	-1.286
22CX28C	9.566	-0.857	29CX25C	8.087	-2.336
1CX28C	9.558	-0.865	22CX25C	8.084	-2.339
1BX29A	9.467	0.591	29CX23C	7.866	-2.557
5CX28C	9.424	-0.999	29CX7C	7.835	-2.588
*P30F45	9.420	-	1CX7C	7.797	-2.626
19CX29A	9.417	-0.504	14CX7C	7.741	-2.682
1CX28B	9.273	-0.104	15CX25C	7.480	-2.943
15CX23C	9.268	-1.155	14CX25C	7.407	-3.016
27BX25C	9.115	-0.262	1CX23C	7.398	-3.025
15CX28C	9.113	-1.310	1BX28B	7.310	-1.021
27BX29A	9.064	0.189	22CX7C	7.050	-3.373
1CX25C	8.761	-1.662	27BX28B	5.627	-2.704
Average	9.010				

Lavras, UFLA, 2009. *Controls; A = population originating from the single-cross hybrid P30F45; B = population originating from the single-cross hybrid Dow 657; C = population originating from the single-cross hybrid DKB333B.

The estimates of SCA and GCA are presented in Table 3. The estimates of GCA were extremely low, i.e., varying from -0.009 to 0.005 and reflecting a small variance of GCA (Table 2). In contrast, the estimates of SCA varied from 3.339 to -3.399 converging to a high variance value of the SCA (Table 3).

Table 3. BLUPs of general combining ability (GCA) and specific combining ability (SCA) of 17 lines $S_{0:2}$, considering a complete unbalanced diallel model.

	23C	25C	28B	7C	28C	29A	GCA
30B	1.661	-	-	1.386	1.776	-0.457	0.005
27B	0.905	0.098	-3.399	0.633	0.911	0.034	-0.001
19C	-0.328	-0.304	0.718	-0.508	-0.889	0.387	-0.001
22C	-0.438	-0.931	-	-1.968	0.537	0.866	-0.002
1B	0.604	-0.930	-1.721	1.625	3.339	0.432	0.004
1C	-1.623	-0.253	0.248	-1.219	0.530	-0.594	-0.003
14C	-0.557	-1.608	0.855	-1.277	-0.877	2.194	-0.003
29C	-1.156	-0.928	1.602	-1.183	-	-	-0.002
5C	-0.889	-0.715	0.867	-0.781	0.393	0.999	-0.001
12C	-0.553	-0.306	1.240	-0.849	-0.468	-	-0.004
15C	0.247	-1.535	1.169	-0.306	0.084	-0.484	-0.002
GCA	-0.003	-0.009	0.001	-0.006	0.005	0.004	
σ^2_{GCA}	0.002						
σ^2_{SCA}	1.622						
σ^2_{e}	1.137						
h^2_a	0.741						

Lavras, UFLA, 2009.

The model considering the genotypes in a single group (unbalanced complete diallel) was more accurate than the partial diallel model with and without the relationship information, i.e., a model that considers the genitors in distinct groups (Figure 1).
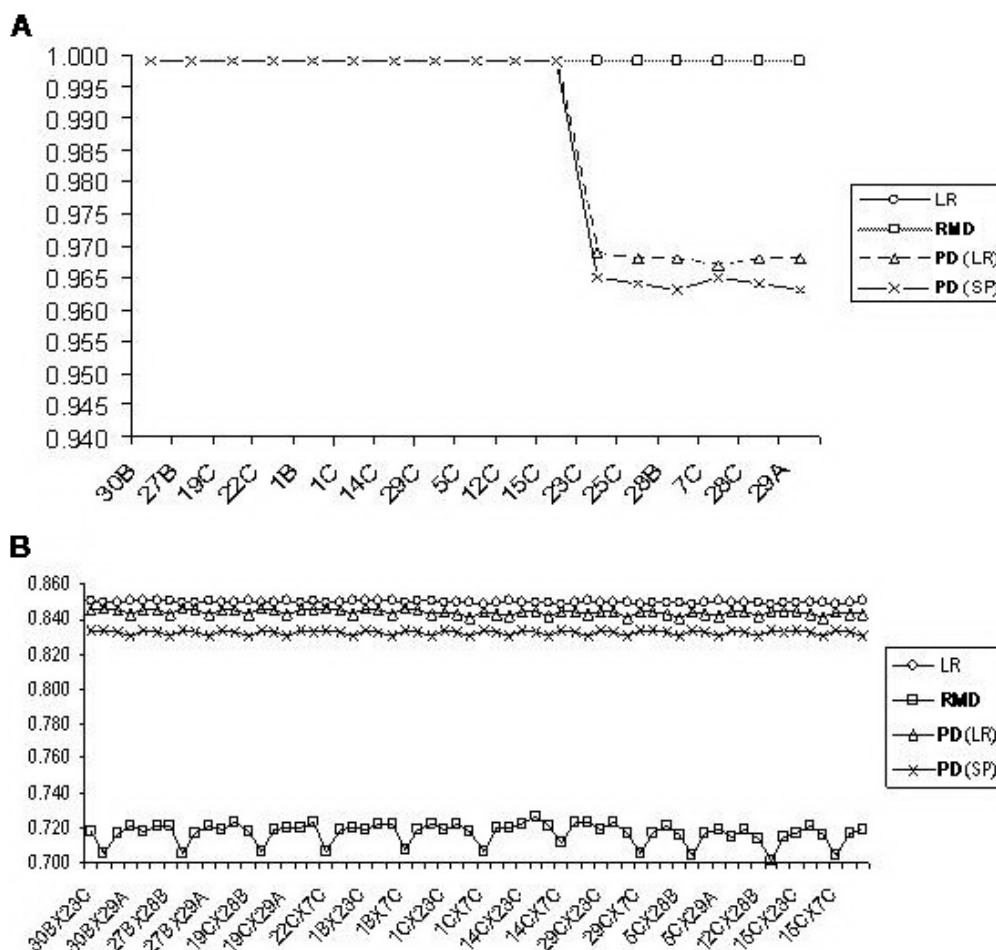

Figure 1. Accuracy of general combining ability (A) and specific combining ability (B) BLUPs considering the unbalanced complete diallel with Lynch and Ritland's (LR) relationship, Roger's modified distance (RMD), partial diallel with Lynch and Ritland's [PD (LR)] relationship, and partial diallel with no marker information [PD (SP)].

Accuracy using the RMD coefficients, representing similarity-by-state, can lead to an increase in the prediction error variance and, consequently, loss of accuracy of genetic values in comparison with estimators that eliminate this bias (Figure 1). Moreover, it can be noted that the model without relationship information had accuracy loss in GCA and SCA estimates in relation to the coefficient LR.

Nevertheless, accuracy values using RMD were advantageous only for GCA BLUPs in relation to the two models of partial diallel. However, for SCA accuracy values, RMD gave lower values in comparison with the partial diallel model with and without relationship information (Figure 1).

The twelve most divergent hybrids, on average, were very productive and showed, on average, positive estimates of SCA and heterosis (Table 4). In contrast, the twelve least divergent hybrids displayed poor performance in yield and negative SCA and heterosis values.

Table 4. Selection of the 12 most divergent crosses and comparison with the 12 least divergent ones, considering the values of yield, heterosis and specific combining ability (SCA).

Order	Cross	Distance	Yield	SCA	Heterosis
		Twelve most divergent crosses			
1	1BX25C	0.337	8.091	-0.926	-1.286
2	27BX25C	0.313	9.115	0.110	-0.262
3	14CX28B	0.306	9.879	0.854	0.502
4	14CX29A	0.306	11.222	2.190	1.300
5	22CX29A	0.283	9.894	0.860	-0.027
6	19CX28B	0.274	9.745	0.716	0.368
7	1BX23C	0.272	9.631	0.598	0.254
8	30BX23C	0.271	10.690	1.656	1.313
9	30BX28C	0.271	10.813	1.759	1.436
10	1CX28B	0.265	9.273	0.248	-0.104
11	5CX28B	0.265	9.894	0.865	0.517
12	15CX28B	0.265	10.194	1.169	0.817
Average		0.285	9.087	0.842	0.402
		Twelve least divergent crosses			
Order	Cross	Distance	Yield	SCA	Heterosis
1	27BX28B	0.074	5.627	-3.406	-2.704
2	1CX28C	0.061	9.558	0.524	-0.865
3	12CX28C	0.058	8.560	-0.473	-1.863
4	19CX7C	0.030	8.511	-0.499	-1.912
5	22CX23C	0.030	8.583	-0.433	-1.840
6	22CX7C	0.030	7.050	-1.958	-3.373
7	1CX7C	0.030	7.797	-1.207	-2.626
8	14CX7C	0.030	7.741	-1.265	-2.682
9	29CX7C	0.030	7.835	-1.172	-2.588
10	5CX7C	0.030	8.238	-0.771	-2.185
11	12CX7C	0.030	8.167	-0.836	-2.256
12	15CX7C	0.030	8.712	-0.294	-1.711
Average		0.039	8.032	-0.982	-2.217

Lavras, UFLA, 2009. A = population originating from the single-cross hybrid P30F45; B = population originating from the single-cross hybrid Dow 657; C = population originating from the single-cross hybrid DKB333B.

However, the most divergent hybrid was not the most productive; on the contrary, its performance was below the overall mean. Similarly, the least divergent hybrid was not the one with the worst performance of the group of hybrids evaluated (Table 4).

Still, on average, the most divergent hybrids were the most productive ones, concurring with medium-range correlation values between RMD and yield, SCA and heterosis (Table 5). All these values differed from zero by the Mantel test. However, when considered sepa-

rately, intra- and interpopulation hybrids, there was no significant correlation between RMD and the variables yield, SCA and heterosis.

Table 5. Correlation between Roger's modified distance (RMD) and yield, specific combining ability (SCA) and heterosis referring to 60 hybrids evaluated in two locations in the agricultural year 2007/2008.

Distance	Yield	SCA	Heterosis
RMD	0.546[+]	0.567[+]	0.661[+]
RMD intra	0.107	0.141	0.153
RMD inter	0.031	0.015	0.047

Lavras, UFLA, 2009. [+]Significant at 1% probability by the Mantel test based on 10,000 simulations.

The coefficient of correlation (r) between RMD and yield (0.546) varied from -0.64 to 0.97, depending on the sample size (Figure 2). The probability of r being greater than zero for sample size 10 was 95.7%. This probability $P(r \leq 0|r)$ was practically zero for all other sample sizes, demonstrating that, independent of the sample size, there was a positive association between RMD and ear yield.
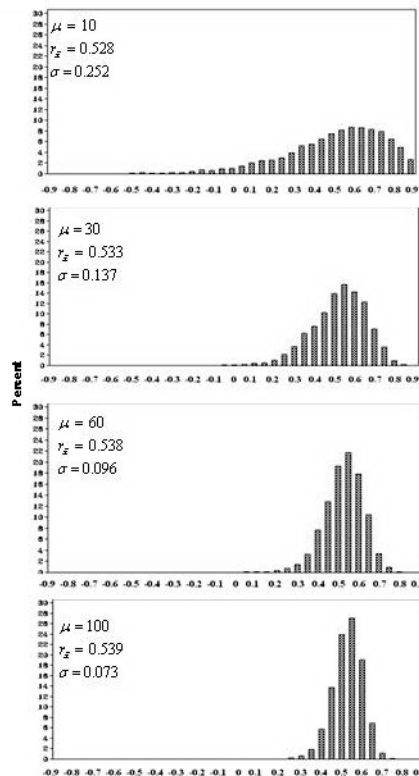

Figure 2. Sample size (μ), standard deviation (σ) and average correlation ($r_{\bar{x}}$) of data obtained by Monte Carlo simulation, fitted to genetic distance and yield values of 60 hybrids from partially endogamic lines.

Taking sample size 60 as reference, one can note that the correlation value between RMD and yield, in 98% of the samples, varied from 0.35 to 0.70, with an average of 0.538 using 600,000 data randomly sampled within the populations simulated and fitted according to phenotypic values and genetic distance. This result supports the correlation values found for each environment (0.402 and 0.603).

On average, all samples showed correlation values near those of sample values (Figure 2), which was expected from the result of the test population (Table 1).

DISCUSSION

The use of hybrids from partially endogamic lines, obtained from single-cross commercial hybrids, has been suggested by several authors (Souza Jr., 1992; Carvalho et al., 2003; Amorim et al., 2006). All these studies have reported genotypes with superior performance compared to the checks (parental hybrids). In the present study, the results were not different, i.e., most interpopulation hybrids were better than the average of the parental hybrids (71%).

This result was greater than that found by Amorim et al. (2006) with S_0 hybrids for the same populations (11.65%). Moreover, six crosses were greater than DKB333B, while Amorim et al. (2006) evaluated 163 hybrids and found nothing better than this commercial hybrid. In contrast, Carvalho et al. (2003) found that 17% of the S_2 x S_2 hybrids were superior to the hybrid DKB333B, demonstrating the great potential of using this type of germplasm for hybrid synthesis.

Although some intrapopulation crosses gave hybrids with performance above the overall mean, these hybrids were no better than their parents. One possibility is that only 33 intrapopulation hybrids were evaluated, accounting for the low performance of these genotypes, due to the greater probability of finding endogamic combinations. However, considering that only 27 interpopulation hybrids were evaluated and that most of them showed high performance, it is reasonable to state that this type of cross would be preferred for these populations.

Table 2 presents yield genotypic values and not the phenotypic averages. Thus, multiple comparison tests are not recommended for genotypic values since BLUP is considered to be an estimator of the "shrinkage" type, which penalized the genotypes that were most affected by the genetic evaluation, bringing the values near the overall average.

Some aspects of SCA and GCA and their respective variances deserve being highlighted. According to Sprague and Tatum (1942), GCA reflects the average performance of a progenitor in a group of crosses. Moreover, according to Boer and Hoeschele (1993), endogamic crosses affect the estimates of variance components, a fact that can be observed in a complete diallel, where related and non-related individuals are crossed. In the present study, the results were not different, i.e., the estimate of GCA variance component reflecting the variability of the parent's performance in the crosses was very low and near zero.

Considering the concept of GCA and endogamic and non-endogamic crosses evaluated jointly in complete diallels, one can note in Table 2 that the lines showing good performance in interpopulation crosses (non-endogamic) usually displayed low yield in intrapopulation crosses (endogamic). This result could have contributed to the underestimation of GCA values, which were near zero, and consequently, GCA variance was near zero (Table 3). Thus, it can be inferred that the group of hybrids evaluated in this study was not strongly affected by the parents' GCA in the crosses, converging to very distinct values

of additive and non-additive variances in relation to GCA and SCA from what would be expected for these variance components in maize crop (Hallauer and Miranda Filho, 1981).

Another important aspect to be highlighted is the reason for using an unbalanced complete diallel instead of a partial diallel model for the analysis. As stated in Material and Methods, the crosses were performed in an 11 x 6 scheme, which would lead to a partial diallel to evaluate GCA and SCA. However, these crosses were obtained with the objective of recovering the most and least divergent combinations, which led to the allocation of related individuals out the groups, i.e., not necessarily in distinct groups.

In this study, the use of a partial diallel loses part of the relationship in matrix A, while the complete diallel, even when highly unbalanced, can recover this information, leading to greater accuracy in predicting GCA and SCA values. This fact is evident in Figure 1, where the six last progenitors, when allocated in a group apart from all others (11 x 6), had smaller GCA BLUP accuracy in relation to the unbalanced complete diallel and, consequently, smaller accuracy values of SCA, which is also affected by the loss of relationship information (Figure 1).

Therefore, the preference for the use of an unbalanced complete diallel instead of a partial one can be justified. It should be noted that the traditional approach of diallel crosses by ordinary least squares, or even the use of mixed models, without the use of relationship information, would not allow the recovery of information, since it does not assume covariance among the genotypes. Thus, it is possible to infer that, when relationship information on the genotypes evaluated is unavailable, the information provided by markers can contribute to prediction accuracy of genetic values.

Nevertheless, some authors (Bauer et al., 2006; Balestre et al., 2008c) have suggested the use of similarity-in-state to predict genetic values using the methodology of BLUP. However, observing accuracy values for the predictions of SCA and GCA (Figure 1), it can be noted that the LR coefficient yielded a greater magnitude of accuracy values in comparison with RMD.

This greater prediction error caused by RMD on SCA and GCA values was expected, due to the recovery of genetic information that takes into consideration only the evolution aspect (similarity-in-state) of the individuals and not their relationship. Thus, in greater or lesser weight, all genetic covariance values are recovered by the linear predictor (BLUP) for each SCA and GCA values, including those individuals that do not share alleles inherited from common progenitors. However, this bias is eliminated by the use of Lynch and Ritland's (1999) coefficient since, despite using molecular data (similarity-by-state), it takes into consideration the conditional probability that two individuals have alleles inherited from common progenitors (similarity-by-decent). Thus, it can be inferred that the use of clean similarity-in-state for the construction of matrix A can lead to error in the prediction of genetic values.

The genetic distances obtained by microsatellite markers showed moderate efficacy in the prior selection of the hybrids evaluated in this study. The twelve most divergent hybrids, on average, were greater in yield, SCA and heterosis in relation to the least divergent, resulting in a selection efficiency of 20% in relation to the greatest observed yield values. However, the top performance hybrid was not within the group of selected hybrids, demonstrating that a promising combination can be lost by the use of genetic distance prior to selection (Table 4).

The correlation results between RMD and yield (0.546) were practically identical to those obtained by Amorim et al. (2006) for S_0 progenies from these same populations (0.55). This allows the inference that there is repeatability in the association between RMD and yield for the populations involved in this study. This correlation value is considered of medium magnitude, but does allow prediction reliability. This occurs because there was no association

between RMD and yield in the separation of intra- and interpopulation hybrids, i.e., although there is average joint prediction efficacy for the inter- and intrapopulation hybrids, RMD had small prediction efficacy for intra- or interpopulation crosses.

This result disagrees with those obtained by Reif et al. (2003), Amorim et al. (2006) and Balestre et al. (2008c), who found large differences in correlation values in separate inter- and intrapopulation groups. However, those authors disagree on the magnitude of the association found between RMD with yield in intra- and intergroup hybrids.

Thus, it can be inferred that, although the correlation results are apparently acceptable, RMD showed only average magnitude selection efficacy, when high variability in the values for genetic distance and yield was observed. When this variability decreased, in the case of separating inter- and intrapopulation hybrids, this moderate efficacy vanished, which was expected by Pearson's correlation. These same inferences can be made for SCA and heterosis, although the correlation between RMD and heterosis (0.661) showed a greater value in comparison with yield and SCA. Also, for the correlation between RMD and yield, Monte Carlo simulation results can be used, for illustration purposes, to demonstrate its application in correlation studies.

Focusing on the study of Amorim et al. (2006), it can be noted that these authors used only 24 hybrids for the correlation study of a group of 163 hybrids evaluated, where only one-half of them (12) were intra- and the other half interpopulation. These authors found that the high correlation value between RMD and yield was obtained only when inter- and intrapopulation hybrids were separated, and then, only the interpopulation hybrids showed high correlation (0.84), while low correlation values were obtained (0.18) with intrapopulation ones.

Considering that the same populations were used in the present study, that the progenies used in the crosses are derived from the same ones used by Amorim et al. (2006), and that the correlation coefficient obtained in this study was practically the same as that obtained by those authors, it is possible to infer that these authors obtained a high correlation value between RMD and yield for interpopulation hybrids (0.84) by chance, due to the small sample size (12) removed from the data group with a correlation value of 0.55 and size of 24. The Monte Carlo simulation method supports this inference (Figure 1).

It can be noted that drawing several samples of size 10 from populations with a fixed correlation of 0.55, the density of samples that presented similar values (0.55) was basically the same density of samples with correlation of 0.80 (Figure 2), i.e., incurs in sampling error on the correlation estimate.

However, this parity disappears when using larger samples (60 and 100), with the population coefficient prevailing (Figure 2). Similar results were observed by Millsap (1988), in correlation studies with Monte Carlo simulation. That author concluded that samples smaller than 60 can be biased due to sampling error, which was also observed in this study, where a sample size of 60 gave practically the same result as that with sample size of 100. Moreover, the correlation values obtained on the individual analyses were within the confidence interval for sample size of 60 (0.402 and 0.603).

Thus, it is possible to test the significance of the correlation coefficient for asymmetric data (RMD) with a correlation study using Monte Carlo simulation, as well as to obtain the required information to determine sample size in order to avoid the risk of obtaining random correlation values, as emphasized by Fan et al. (2002).

Based on the results of the present study, it can be concluded that the use of partially endogamic lines is a viable alternative for the production of high performance hybrids and that ge-

netic distances provide moderate yield predictive ability, combination ability and hybrid heterosis.

Monte Carlo simulation is a viable alternative to test the correlation between genetic distance and yield, and can be used to determine the adequate sample size in association studies between these two variables.

Relationship estimates by molecular markers through Lynch and Ritland's coefficient have been shown to be a promising methodology for predicting genetic values in mixed linear models (BLUP), especially when no pedigree information is available.

ACKNOWLEDGMENTS

REFERENCES

Amorim EP, Amorim UBO and Santos JB (2006). Genetic distance based on SSR and grain yield of inter- and intrapopulational maize single cross hybrids. *Maydica* 51: 507-513.

Araújo PM (2001). Dialelo Parcial Circulante Interpopulacional e Cruzamentos "Top Cross" na Avaliação de Linhagens Parcialmente Endogâmicas de Milho. PhD thesis, Universidade de São Paulo, Piracicaba.

Balestre M, Machado JC, Lima JL, Souza JC, et al. (2008a). 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.

Balestre M, Von Pinho RG, Souza JC and Lima JL (2008b). Comparison of maize similarity and dissimilarity genetic coefficients based on microsatellite markers. *Genet. Mol. Res.* 7: 695-705.

Balestre M, Von Pinho RG, Souza JC and Machado JC (2008c). Potential of maize single-cross hybrids for extraction of inbred lines using the mean components and mixed models with microsatellite marker information. *Genet. Mol Res.* 7: 1106-1118.

Bauer AM, Reetz TC and Leon J (2006). Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. *Crop Sci.* 46: 2685-2691.

Bernardo R (1991). Correlation between testcross performance of lines at early and late selfing generations. *Theor. Appl. Genet.* 82: 17-21.

Bernardo R (1992). Relationship between single-cross performance and molecular marker heterozygosity. *Theor. Appl. Genet.* 83: 628-634.

Bernardo R (1995). Genetic models for predicting maize single-cross performance in unbalanced yield trial data. *Crop Sci.* 35: 141-147.

Boer IJM and Hoeschele I (1993). Genetic evaluation methods for populations with dominance and inbreeding. *Theor. Appl. Genet.* 86: 245-258.

Carlone MR Jr and Russel WA (1989). Evaluation of S_2 maize lines reproduced for several generations by random mating within lines. II: Comparisons for testcross performance of original and advanced S_2 and S_8 lines. *Crop Sci.* 29: 899-904.

Carvalho ADF, Souza JC and Ribeiro PH (2003). Performance of hybrids of partially endogamic lines S_2 of maize in regions of the States of Roraima and Minas Gerais [Desempenho de híbridos de linhagens parcialmente endogâmicas de milho em regiões dos Estados de Roraima e Minas Gerais]. *Cienc. Agrotec.* 27: 985-990.

Dias LA, Picoli EA, Rocha RB and Alfenas AC (2004). *A priori* choice of hybrid parents in plants. *Genet. Mol. Res.* 3: 356-368.

Falconer DS and Mackay TFC (1996). Introduction to Quantitative Genetics. 4th edn. Longman, Essex.

Fan XO, Felsóvály IA, Sivo AS and Keenan SC (2002). SAS® for Monte Carlo Studies: A Guide for Quantitative Researchers. SAS Institute Inc., Cary.

Fleishman AI (1978). A method for simulating non-normal distributions. *Psychometrika* 43: 521-532.

Griffing B (1956). Concept of general and specific combining ability in relation to diallel crossing system. *Aust. J. Biol. Sci.* 9: 463-493.

Guimarães OS, Paterniani MEAGZ, Luders RR, Souza AP, et al. (2007). Correlation between the heterosis of maize hybrids and genetic divergence among lines [Correlação da heterose de híbridos de milho com divergência genética entre linhagens]. *Pesq. Agropec. Bras.* 42: 811-816.

Hallauer AR and Miranda Filho JB (1981). Quantitative Genetics in Maize Breeding. Iowa State University Press, Ames.

Hallauer AR, Russel WA and Lamkey KR (1988). Corn Breeding. In: Corn and Corn Improvement. 3rd edn (Sprage GF and Dudley JW, eds.). Iowa State University Press, Madison, 453-464.

Henderson CR (1984). Applications of Liner Models in Animal Breeding. University of Guelph Press, Guelph.

Kaiser HF and Dickman K (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika* 27: 179-182.

Lynch M and Ritland K (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* 152: 1753-1766.

Manly BFJ (1991). Randomization and Monte Carlo Methods in Biology. Chapman & Hall, London.

Mantel N (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209-220.

May EC, Spottiswoode SJP and Faith LV (2001). A Methodological Issue in the Study of Correlation Between Psychophysiological Variables. In: Proceedings of Presented Papers. The Parapsychological Association 44th Annual Convention, August 2-4, 2001, New York, 166-178.

Millsap RE (1988). Sampling variance in attenuated correlation coefficients. *J. Appl. Psychol.* 73: 316-319.

Reif JC, Melchinger AE, Xia XC, Warburton ML, et al. (2003). Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. *Crop Sci.* 43: 1275-1282.

Reif JC, Melchinger AE and Frisch M (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Sci.* 45: 1-7.

SAS Institute (2000). Version 8. SAS Institute Inc., Cary.

Souza Junior CL (1992). Interpopulation genetic variances and hybrid breeding programs. *Rev. Bras. Genet.* 15: 643-652.

Sprague GF and Tatum LA (1942). General *vs* specific combining ability in single crosses of corn. *J. Am. Soc. Agron.* 34: 923-932.

Troyer AF (1994). Breeding Early Corn. In: Specialty Corns (Hallauer AR, ed.). CRC Press, Ames, 342-396.

Van Vleck LD (1993). Selection index and introduction to mixed model methods for genetic improvement of animals: the green book. CCR Press, Boca Raton.