INDIRECT RESPONSE IN GENOME WIDE SELECTION USING SELECTED MARKERS

RESPOSTA INDIRETA NA SELEÇÃO GENÔMICA AMPLA USANDO SELEÇÃO DE MARCADORES

Leonardo Lopes BHERING[1]; Shizhong XU[2]
1. Universidade Federal de Viçosa, Departamento de Biologia Geral, Viçosa, MG, Brasil. leonardo.bhering@ufv.br ; 2. University of California, Riverside, Department of Botany and Plant Sciences, Riverside, CA, United States

ABSTRACT: The goal of this work was to compare the effect of the accuracy and residual variance in genome wide selection using marker selection as well as using the effect of the indirect selection, using simulated and real data. In simulated data was used one sample with 200 individuals with 1,000 molecular markers in F2 population. The real data was obtained in maize with F2 population with 441 individuals and genotyping with 261 SSR markers. There was 11 traits evaluated (ear length, ear width, row number, kernels per row, 100-kernel weight, ear weight, grain yield, length of branch, number of branch, plant height and ear height). All data was analyzed using rrBLUP method and 10-fold cross-validation. In simulated and maize data the results were similar: the residual variance with few markers is lower than with the 1000 markers and the accuracy with few markers is bigger than with 1000 markers. For maize data multi trait selection, the accuracy increased when the correlation between traits is greater than 0.50 and residual variance decreased when the correlation is greater than 0.70. In this sense, these results showed that marker selection could be used as a first step in genome wide selection, improving the prediction and compute demand.

KEYWORDS: Genome Selection. Maize. Simulation. Data analysis. Plant breeding

INTRODUCTION

The principle of genomic selection is simultaneously estimate the effect of all markers in a training population comprised of phenotyped and genotyped individuals (MEUWISSEN et al., 2001). Thus, genomic estimated breeding values (GEBVs) could be calculated as the sum of estimated marker effects for genotyped individuals in a predicted population. Fitting simultaneously all markers ensures that marker-effect estimates are unbiased, small effects are captured (BROMAN; SPEED, 2002). Moreover, this can potentially capture all the quantitative trait loci (QTL) that contribute to the variation of a trait. The QTL effects, inferred from either haplotypes or individual single nucleotide polymorphism markers, are first estimated in a large reference population with phenotypic information. In subsequent generations, only the marker information is required to calculate the GEBVs (HEFFNER et al., 2009).

This approach is based upon the estimation of breeding values (EBV) available for genotyped individuals comprising a trained population using linear or non-linear models applied to phenotypes (DE LOS CAMPOS et al., 2013). After the estimate to EBV is determinate the accuracy of the derived prediction equations in an independent validation population and your application of the prediction equations to generate genomic estimated breeding values (GEBV) in selection candidates within an implementation population. These estimated values (GEBVs) are outputted from a model estimating the relationship between genome-wide markers and phenotypes of the individuals undergoing selection.

Genome Selection has been most successfully implemented in animal breeding (DAETWYLER et al., 2013; GARRICK, 2011; HAYES et al., 2009) and plant breeding (BERNARDO, 2010; JENA et al., 2008; SPINDEL et al., 2015). Genome-wide prediction is also being recognized as an important tool to predict phenotypes (LEE et al., 2008) and the genetic risk for diseases (WRAY et al., 2007) in other fields than animal or plant breeding. The key principle for all these applications is the simultaneous estimation of all genome-wide marker effects based on a reference population with known phenotypes.

Since the number of markers is typically much larger than the number of phenotyped animals in the reference population, most of the proposed models in genome selection attempt to reduce the effective dimensionality of the marker data, although the prediction models usually use only a single phenotypic trait (VAZQUEZ et al., 2010). However, new varieties of crops and animals are evaluated for their performance on multiple traits (DAETWYLER et al., 2010). Crop breeders record

Received: 13/05/18
Accepted: 05/12/18

http://dx.doi.org/10.14393/BJ-v35n5a2019-42245

phenotypic data for multiple traits in different categories such as yield components (e.g., grain weight or biomass), grain quality (e.g., taste, shape, color, nutrient content), and resistance to biotic or abiotic stress (JIA; JANNINK, 2012).

The goal of this work was to compare the effect of the accuracy and residual variance in genome wide selection using marker selection for each trait as well as using the effect of the indirect selection.

MATERIAL AND METHODS

Data simulation

Data were generated using the simulation module implemented in the GENES software (CRUZ, 2013). One sample with 200 individuals were generated with 10 linkage groups (LG) each. A genome of 10 LG was simulated, similar to a diploid species 2n=10, with 100 cM size, considering the existence of 100 molecular markers for linkage group, equally spaced; thus, a total of 1,000 molecular markers were evaluated. Contrasting homozygous parents were simulated to produce F1 generation; thus, parent 1 (AA) was coded with 1 for all markers, and parent 2 (aa) was coded with 0 for all markers.

For the population size (200 individuals) was generated an F2 population, and for that, each F1 individual produced 5000 gametes. There was random fecundation, generating F2 individuals. This process was repeated until all individuals were formed. This population was coded as 0, 1, and 2, for homozygous recessive individuals (aa), heterozygous individuals (Aa), and dominant homozygous individuals (AA) to the considered locus, respectively.

The phenotypic value for 4 different traits was simulated considering 1000 markers previously simulated, 200 of them controlled the characteristics, and the 20 first molecular markers in each LG were taken into account. Once there are 10 LG, there is a total of 200 loci.

It was also used the binomial distribution of effects for each characteristic in each LG. It was adopted the additive gene action of all loci, i.e., the dominance effect was considered null. To establish the phenotypic value it was added up a constant equal to 100, thus preventing that any of the individuals for each of these variables presented negative values.

The simulated data was analyzed using the rrBLUP method. This method was chosen owing to its accuracy and speed to perform the analysis. Data analyses were performed using R software (TEAM, 2012) with the package rrBLUP (ENDELMAN, 2011). A server DELL 12º generation, Intel Xeon E5-26 processor 3,30 GHz, RAM with 64 GB and Hard drive with 1024 GB was used to run the analyses.

After generating the data, mapping process was carried out, starting by the segregation of individual loci analysis. Chi-square tests were used (χ^2), at 5% probability, to confirm the result of segregation in each marker for all the generated populations. In addition, it was verified if all Linkage Groups had been restored, with size, distance and markers order, for concluding whether the simulated populations were F2 with the desired simulation properties.

A 10-fold cross-validation analysis was used as the validation methods. In the 10-fold cross-validation, the original sample was randomly partitioned into 10 equal sized subsamples. From the 10 subsamples, a single subsample was retained as the validation data for testing the model, and the remaining nine were used as training data. The cross-validation process was repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds was averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Initial analysis had 1000 molecular markers. The rrBLUP was used to obtain the effect of each molecular marker. The marker with the lowest value was removed from the matrix, so that the new molecular marker matrix has now 999 markers, which will be used as the input for the subsequent analysis with rrBLUP. This process continues, removing 1 marker every round (so 999 times in this work, or number de markers -1 times) until the molecular marker matrix has no more markers to be excluded. Also, in each round the cross-validation methods were used to estimate the accuracy in validation subsample and the residual variance (Figure 1).

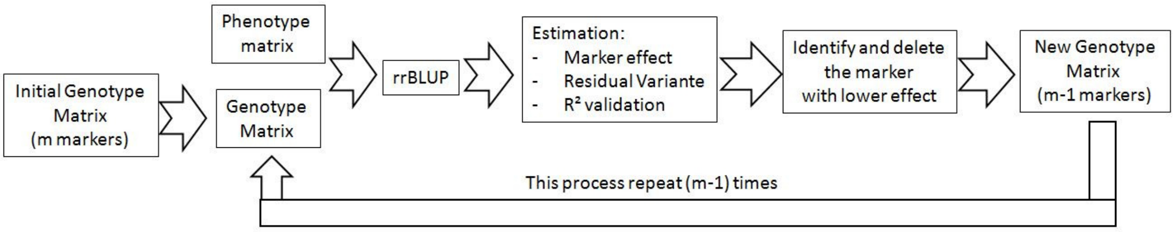
Initial Genotype Matrix (m markers)
Phenotype matrix
Genotype Matrix
rrBLUP
Estimation:
- Marker effect
- Residual Variante
- R² validation
Identify and delete the marker with lower effect
New Genotype Matrix (m-1 markers)
This process repeat (m-1) times

Figure 1. Representation of the process for selecting the number of markers.

After the marker selection, a matrix with the minimum number of markers possible was created. Such matrix only included the number of markers providing the maximum value of accuracy (given by the square of Pearson Correlation between EGBV and phenotypic value – R^2validation). This process was carried out for each phenotypic trait.

Having a matrix with the minimum number of markers for each trait, an analysis using rrBLUP was performed for another additional traits, so that is possible to compare which effect of the selection in one trait also occur in another trait.

Real data

The real data was obtained from maize F2 population with 441 individuals and genotyping with 261 SSR (simple sequence repeat) markers. There were 11 traits evaluated (ear length, ear width, row number, kernels per row, 100-kernel weight, ear weight, grain yield, length of branch, number of branch, plant height and ear height).

The F2 population was coded with 0, 1, and 2 for homozygous recessive individuals (aa), heterozygous individuals (Aa), dominant homozygous individuals (AA) to the considered locus, respectively. The same procedures of evaluation used for the simulated data were used for the real data. A 10 folds cross-validation was used with rrBLUP to produce the selection of markers and effect of indirect selection compared with uni trait selection.

RESULTS AND DISCUSSION

Simulated data

Genetic maps were constructed for the population in order to evaluate the quality of simulated data. Beginning with the analysis segregation of individual loci, chi-square tests (χ^2) were applied to verify the segregation ratio of all generated populations after the simulation process. There was no segregation distortion, that is, all markers typically segregate as a codominant F2 (1: 2: 1). Moreover, it was found that all linkage groups were restored according to the parameters used in simulation including the total size (100cM), in the main distance between markers, and the order of the markers that constitute the linkage group. Thus, it is concluded that the simulated population have characteristics of an F2 population, and therefore it would be appropriate for this study. This also shows the importance of performing such analyses to ensure F2 population usage, so that the genetic parameters set by Falconer and Mackay (1996) including genetic variance, phenotypic variance, and environmental variance, can be properly inferred for this type of population with rrBLUP.

The phenotypic correlation between all the simulated traits is presented in Table 1. All correlations were significant using the t-test with 1% of probability. The correlation coefficient values ranged from 0.50 to 0.844.

Table 1. Correlation between all four traits simulated in F2 population.

	1	**2**	**3**	**4**
1	1	0.844**	0.804**	0.795**
2	0.844**	1	0.500**	0.553**
3	0.804**	0.500**	1	0.549**
4	0.795**	0.553**	0.549**	1

**: Significant with 1% of probability by t test

The results were similar for all the four simulated traits (Table 2). The accuracy (R^2) of the process increased with the marker selection. At the same time, the residual variance with few markers was lower than with the original marker matrix (1000 markers). The accuracy of the process (R^2 validation) is measured by the square of the correlation between the phenotype values and the predict values after 10-fold cross-validation. Eliminating marker is possible to increase de accuracy. In this sense, it might be desirable to reduce the number of SNP to ease computation process when predicting individual SNP effects and summing effects to calculate genomic predictions (JARQUÍN et al., 2014).

Table 2. Effect in Accuracy (R^2) and residual variance (RV) with the 1000 markers and with the number of minimum markers (NM) for four simulated traits in F2 population.

	Trait	1000 markers	NM	R^2 or RV
Accuracy (R^2)	1	0.019	49	0.216
	2	0.023	45	0.236
	3	0.046	22	0.286
	4	0.027	34	0.285
Residual Varianc e (RV)	1	9.889	34	6.942
	2	9.960	35	6.689
	3	9.681	30	6.653
	4	9.706	34	6.018

For all scenarios, the value of accuracy increased when markers are added into the model until the maximum value (value where is the NM). After that point, the accuracy of prediction drastically decreased as additional markers are feed into the model (Figure 2 A). The residual variance, on the other hand, showed an inverse pattern, that is, its values decreased as markers are added into the model reaching its lowest value at about the same point (i.e. the same number of markers) the accuracy of prediction reached its highest value. Therefore, this point indicates the minimum number of markers needed to obtain the lowest residual variance as well as the highest accuracy of prediction (Figure 2 B).

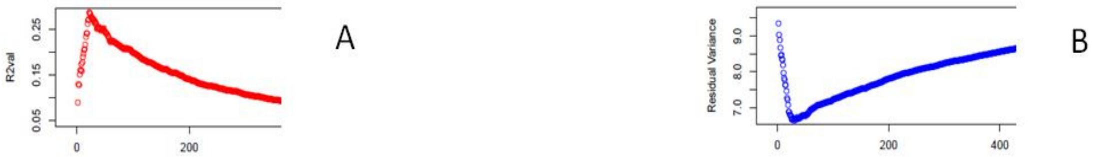

Figure 2. Effect of the number of markers in accuracy (A) and Residual Variance (B) for the trait 3.

In this sense, these results showed that marker selection could be used as a first step in genome wide selection. Another advantage with such approach is that with a low number of markers, analyses can be more rapidly performed, thus facilitating the use of methods demanding great computational capacity (e.g. Bayesian methods). Che and Xu (2010) obtained similar results on a simulated study using an optimal number of QTL determined by the cross-validation test. The authors, analyzing the predictor error, suggested that a reduction on QTLs does not lead to a significant loss in the precision of genome selection.

The response of each trait in relation to the minimum number of markers (NM) selected for one specific trait is presented on Table 3. Considering accuracy and trait 1, the minimum number of markers selected was 49 (Table 2). This new matrix (i.e. the 49 markers only) selected based on trait 1, was then used for analyzing the other traits. For example, for trait 2, accuracy was 0.148. This value is much higher (about six times more) than the accuracy value obtained with the original matrix based on 1000 (Table 2). Regarding accuracy, all possible comparisons showed similar results. In another words, the accuracy on the traits of interest is higher with the NM of markers than using all markers as it is usually done for genome selection prediction considering individual traits. Considering the residual variance (Table 3), the same pattern was observed. For example, the minimum number of markers selected for the trait 1 was 34 (Table 2). Using this matrix and considering trait 2, the residual variance was reduced from 9.889 (1000 markers) to 7.783 (34 markers). For animals in multitrait selection with only phenotypes on a correlated trait, the increase in accuracy was up to 0.04 and 0.18, (CALUS; VEERKAMP, 2011).

Table 3. Response in accuracy (R^2) and residual variance (RV) considering number of minimum matrix (NM) for each simulated trait.

		Response (R2 or RV)			
	NM	1	2	3	4
Accuracy (R^2)	1	-	0.148	0.12	0.115
	2	0.154	-	0.055	0.051
	3	0.137	0.027	-	0.067
	4	0.136	0.057	0.066	-
Residual Variance (RV)	1	-	7.783	8.270	7.954
	2	7.725	-	9.117	9.136
	3	8.052	9.869	-	9.382
	4	7.441	9.339	9.645	-

Currently, the genotyping costs per individual are considerably less expensive for low-density SNP (LD-SNP) panels than high-density (HD-SNP) panels. Thus, there is a significant interest on the development of methods that implement GS using LD-SNP panels. The most common strategy used to develop LD-SNP panels is to employ variable selection methods to identify a small set of markers that are predictive of trait phenotype or breeding value. A potential problem with variable selection for the development of a LD-SNP panel, however, is that selected HD-SNPs might be different for each quantitative trait and population, thereby increasing the number of SNPs that must be genotyped when GS is implemented for the multiple-trait breeding programs. In addition, the effectiveness of this approach may depend on the number of QTL affecting the trait; larger numbers of SNPs will be needed for traits with larger numbers of QTL (HABIER et al., 2009). However, the present work has shown that when traits are correlated, SNPs selected for one trait can also be used for another. In fact, Jia and Jannink (2012) showed that the prediction accuracy for a low-heritability trait could be significantly increased by multivariate genomic selection when a correlated high-heritability trait was available. Furthermore, multiple-trait genomic selection had higher prediction accuracy than single-trait genomic when phenotypes are not available for all individuals and traits.

Real data

The phenotypic correlation between all traits is presented in table 4. Most of all correlations were significant using a t test, except between 100-kernel weight (KW) and kernels per row (KR), length of branch (LB) and number of branch (NB), plant height (PH) and number of branch (NB), ear height (EH) and number of branch (NB). Correlation values ranged from -0.126 to 0.981.

Table 4. Correlation between the traits ear length (EL), ear width (ED), row number (RN), kernels per row (KR), 100-kernel weight (KW), ear weight (EW), grain yield (GY), length of branch (LB), number of branch (NB), plant height (PH) and ear height (EH) in F2 maize population.

	EL	**ED**	**RN**	**KR**	**KW**	**EW**	**GY**	**LB**	**NB**	**PH**	**EH**
EL	1	0.311**	0.168**	0.813**	0.180**	0.668**	0.667**	0.554**	0.126**	0.659**	0.447**
ED	0.311**	1	0.678**	0.336**	0.454**	0.599**	0.577**	0.328**	0.244**	0.647**	0.545**
RN	0.168**	0.678**	1	0.236**	- 0.126**	0.352**	0.369**	0.173**	0.248**	0.385**	0.255**
KR	0.813**	0.336**	0.236**	1	- 0.077^{ns}	0.650**	0.680**	0.416**	0.100*	0.615**	0.483**
KW	0.180**	0.454**	- 0.126**	-0.077 ^{ns}	1	0.411**	0.364**	0.200**	0.115*	0.344**	0.297**
EW	0.668**	0.599**	0.352**	0.650**	0.411**	1	0.981**	0.287**	0.144**	0.691**	0.618**
GY	0.667**	0.577**	0.369**	0.680**	0.364**	0.981**	1	0.266**	0.149**	0.668**	0.578**
LB	0.554**	0.328**	0.173**	0.416**	0.200**	0.287**	0.266**	1	0.084^{ns}	0.484**	0.238**
NB	0.126**	0.244**	0.248**	0.100*	0.115*	0.144**	0.149**	0.084^{ns}	1	0.061^{ns}	0.046^{ns}
PH	0.659**	0.647**	0.385**	0.615**	0.344**	0.691**	0.668**	0.484**	0.061^{ns}	1	0.807**
EH	0.447**	0.545**	0.255**	0.483**	0.297**	0.618**	0.578**	0.238**	0.046^{ns}	0.807**	1

ns,** e * :Non significant, significant with 1 and 5% of probability by t test, respectively

Regarding the accuracy (R^2), all traits showed a similar pattern, that is, the accuracy of the process increased with the marker selection (reduction of markers) and the residual variance decreased (Table 5). Nevertheless, it is important to confirm that marker selection is a better strategy for genome selection when using real data. In this sense, Spindel et al. (2015) tested a different number of markers on three traits in rice: grain yield, flowering time, and plant height. They concluded that for all traits, there was no significant difference in the best-performing genomic selection method for a given trait or validation season when 7,142 SNPs (approximately 1 SNP for every 0.2 cM) were used versus when 13,101 SNPs (1 SNP for every 0.1 cM) or the full 73,147 SNPs were used.

In this study, all traits showed a standard curve, that is, the value of accuracy increased when markers are added into the model until its maximum value was reached (61 markers). Beyond this point, the accuracy of prediction constantly decreased reaching the minimum value (Figure 3 A). The residual variance, on the other hand, displayed an inverse behavior for the same range of markers, that is, it initially decreased as markers are added into the model until its minimum value is reached (74 markers); after this point the residual variance constantly increased as more markers are included. Additionally, this point indicates the minimum number of markers to obtain the lowest residual variance (Figure 3 B). Six K SNP fixed arrays have been recently developed for using within specific rice breeding/research programs. Fixed arrays have established advantages in rice, including robust allele calling, cost-effectiveness per data point, and speed of genotyping turn-around (THOMSON, 2014). Similar results were obtained by Che and Xu (2010) obtained with *Arabidopsis* using an optimal number of QTL determined by the cross-validation tests that do not lead to any significant loss in the precision of genome selection when compared to the use of all QTLs, thus analyzing the predictor error. Using Barley, these authors found a similar pattern, that is, after a specific point (minimum number of QTLs) the square of prediction error increased, thus showing the need to select markers in genome selection.

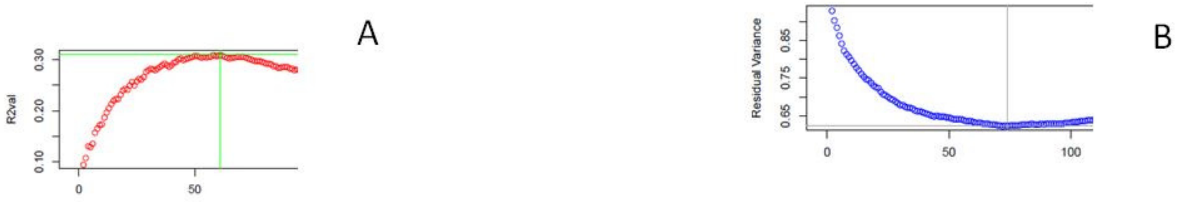

Figure 3. Effect of the number of markers in accuracy (R^2) (A) and Residual Variance (B) for the trait grain yield (GY).

The response in accuracy as well as residual variance for different traits after marker selection for one specific trait is shown in table 6. For the accuracy, the minimum number of markers for the trait EL was 63 (Table 5). This matrix was then used to estimate the accuracy for 10 additional traits. For example, for the trait ED the accuracy was 0.108 and this value is less (-) than that obtained with the original 1000 markers matrix (0.122, Table 5). This situation is not desirable, although it could be explained by the correlation between these two traits (0.31, Table 4). Thus, poor correlation between traits might result on a poor selection of marker, and therefore indirect selection is not advised as good strategy in genomic selection. Jia and Jannink (2012) showed that the prediction of accuracy for a low-heritability trait could be significantly increased by multivariate genomic selection when a highly correlated heritability trait is available. Furthermore, multiple-trait genomic selection showed higher prediction accuracy than compared to single-trait genomic selection when phenotypes are not available for all individuals and traits.

Still, considering the matrix selected for the trait EL, and now the response in KR, the accuracy was 0.179 (Table 6), which is once again greater (+) than the value using the 1000 markers matrix in the KR analysis (0.068) Table 5. This situation is desirable, and it shows the possibility of using the matrix selected for EL to predict KR. This effect could be explained by the correlation between these traits (0.813, Table 4). In this sense, high values of correlation between traits can result in good marker selection for these traits, and thus the indirect selection can be a good strategy in genomic selection. Fixed arrays of 6–12K could thus proving to be the most affordable and efficient way of genotyping for GS, especially for smaller breeding programs with less genotyping informatics expertise (SPINDEl et al., 2015).

Comparing all responses in Table 6, for the accuracy, and using the information of correlation between all traits (Table 4), it can be state that with correlations higher than 0.50 the accuracy is always improved and the multi-trait selection is useful. The only exceptions were between marker selection for EL and the response in EW where the value of accuracy decreased even though the correlation between traits was 0.66; and between marker selection for PH and response in ED where the correlation was 0.64.

For the residual variance, the minimum number of markers selected for EL was 79 (Table 5). This matrix with 79 molecular markers was then used to estimate the residual variance of additional 10 traits.

Table 5. Effect in Accuracy (R^2) and residual variance (RV) with the 1000 markers and with the number of minimum markers (NM) in traits ear length (EL); ear width (ED); row number (RN), kernels per row (KR), 100-kernel weight (KW), ear weight (EW), grain yield (GY), length of branch (LB), number of branch (NB), plant height (PH) and ear height(EH) in F2 maize population.

	Accuracy (R²)			Residual Variance (RV)		
Traits	1000 markers	NM	R2	1000 markers	NM	RV
EL	0.084	63	0.276	1.960	79	1.482
ED	0.122	60	0.339	0.04	76	0.031
RN	0.107	48	0.318	1.575	86	1.205
KR	0.068	69	0.286	8.089	80	5.999
KW	0.081	60	0.299	11.351	72	8.570
EW	0.118	65	0.302	1.011	88	0.828
GY	0.119	61	0.309	0.771	74	0.623
LB	0.08	40	0.28	6.323	75	4.952
NB	0.052	77	0.275	8.360	79	6.140
PH	0.09	61	0.297	196.446	76	151.770
EH	0.088	63	0.293	94.428	76	72.230

For the trait ED, the residual variance was 0.046 and that was greater (+) than the value based on the 1000 markers matrix (0.04, Table 5). This situation is not desirable, but once again it could be explained by the correlation between these traits (0.31, Table 4). Therefore, low values of correlation between traits can also negatively impact the residual variance (i.e. increase it) during the selection of marker for these traits, and thus multi-trait selection should be avoid in genomic selection.

Still, considering the matrix selected for EL, and now the response in KR, the residual variance was 7.264 (Table 6) and that is smaller (-) than the value using the 1000 markers matrix (8.089, Table 5). Therefore, the matrix selected for EL could be used to predict KR. This effect could be explained by the correlation between these traits (0.813, Table 4), so that if high values of correlation between traits are found, a good response is expected during marker selection for these traits, and consequently a multi-trait selection can be a good strategy in genomic selection.

Comparing all responses for the residual variance (Table 6) and using the information of correlation between all traits (Table 4), it is clear that for correlations higher than 0.70 the residual variance always decreases and the multi-trait selection is useful method. The best strategy, however, will likely have multiple genotyping platforms available and the flexibility of switching between them as needed. Genotyping turn-around time is ultimately key for GS because genotypes must be available in time for selections and the next generation crossing. It should be noted that depending on the platform, genotyping individuals with more markers than is necessary could be detrimental to breeding progress if it overloads the bioinformatics and computational capacities of a breeding program (SPINDEL et al., 2015).

Table 6. Response in accuracy (R^2) and residual variance (RV) considering number of minimum matrix (NM) for each trait: ear length (EL); ear width (ED); row number (RN), kernels per row (KR), 100-kernel weight (KW), ear weight (EW), grain yield (GY), length of branch (LB), number of branch (NB), plant height (PH) and ear height(EH) in F2 maize population.

		Response (R2 or RV)										
	NM	**EL**	**ED**	**RN**	**KR**	**KW**	**EW**	**GY**	**LB**	**NB**	**PH**	**EH**
Accuracy (R^2)	**EL**	-	0.108^{-}	0.106^{-}	0.179^{+}	0.033^{-}	0.112^{-}	0.121^{+}	0.083^{+}	0.026^{-}	0.125^{+}	0.074^{-}
	ED	0.06^{-}	-	0.148^{+}	0.048^{-}	0.057^{-}	0.156^{+}	0.142^{+}	0.04^{-}	0.037^{-}	0.095^{+}	0.114^{+}
	RN	0.059^{-}	0.172^{+}	-	0.042^{-}	0.095^{+}	0.111^{-}	0.108^{-}	0.033^{-}	0.046^{-}	0.083^{-}	0.069^{-}
	KR	0.189^{+}	0.056^{-}	0.051^{-}	-	0.074^{-}	0.132^{+}	0.149^{+}	0.085^{+}	0.026^{-}	0.138^{+}	0.085^{-}
	KW	0.048^{-}	0.091^{-}	0.094^{-}	0.057^{-}	-	0.109^{-}	0.1^{-}	0.036^{-}	0.012^{-}	0.055^{-}	0.068^{-}
	EW	0.12^{+}	0.182^{+}	0.065^{-}	0.148^{+}	0.096^{+}	-	0.297^{+}	0.036^{-}	0.029^{-}	0.146^{+}	0.086^{-}
	GY	0.116^{+}	0.182^{+}	0.102^{-}	0.137^{+}	0.092^{+}	0.29^{+}	-	0.029^{-}	0.044^{-}	0.148^{+}	0.106^{+}
	LB	0.057^{-}	0.066^{-}	0.051^{-}	0.04^{-}	0.026^{-}	0.063^{-}	0.057^{-}	-	0.024^{-}	0.078^{-}	0.045^{-}
	NB	0.051^{-}	0.062^{-}	0.093^{-}	0.041^{-}	0.079^{-}	0.061^{-}	0.059^{-}	0.062^{-}	-	0.043^{-}	0.049^{-}
	PH	0.085^{+}	0.086^{-}	0.043^{-}	0.088^{+}	0.046^{-}	0.128^{+}	0.123^{+}	0.041^{-}	0.02^{-}	-	0.194^{+}
	EH	0.065^{-}	0.1^{-}	0.028^{-}	0.069^{+}	0.03^{-}	0.128^{+}	$0.119^{=}$	0.024^{-}	0.047^{-}	0.181^{+}	-
Residual Variance (RV)	**EL**	-	0.046^{+}	1.738^{+}	7.264^{-}	13.029^{+}	1.082^{+}	0.811^{+}	6.676^{+}	9.227^{+}	201.045^{+}	104.621^{+}
	ED	2.135^{+}	-	1.641^{+}	8.860^{+}	12.740^{+}	1.044^{+}	0.805^{+}	7.058^{+}	9.164^{+}	213.349^{+}	100.506^{+}
	RN	2.217^{+}	0.042^{+}	-	8.852^{+}	12.566^{+}	1.143^{+}	0.865^{+}	7.411^{+}	9.043^{+}	220.228^{+}	106.968^{+}
	KR	1.766^{-}	0.049^{+}	1.823^{+}	-	12.054^{+}	1.054^{+}	0.784^{+}	6.626^{+}	8.940^{+}	195.511^{-}	100.150^{+}
	KW	2.196^{+}	0.046^{+}	1.743^{+}	8.754^{+}	-	1.112^{+}	0.852^{+}	7.100^{+}	9.294^{+}	220.514^{+}	106.682^{+}
	EW	1.990^{+}	0.042^{+}	1.834^{+}	7.628^{-}	11.929^{+}	-	0.635^{-}	7.124^{+}	9.019^{+}	191.210^{+}	98.378^{+}
	GY	2.021^{+}	0.041^{+}	1.740^{+}	7.835^{-}	12.115^{+}	0.853^{-}	-	7.306^{+}	9.111^{+}	192.189^{-}	97.229^{+}
	LB	2.167^{+}	0.048^{+}	1.845^{+}	8.958^{+}	13.324^{+}	1.164^{+}	0.883^{+}	-	9.120^{+}	227.591^{+}	110.993^{+}
	NB	2.167^{+}	0.047^{+}	1.727^{+}	8.899^{+}	12.487^{+}	1.164^{+}	0.882^{+}	6.880^{+}	-	223.342^{+}	106.813^{+}
	PH	2.070^{+}	0.046^{+}	1.929^{+}	8.196^{+}	12.753^{+}	1.039^{+}	0.793^{+}	7.149^{+}	9.273^{+}	-	84.903^{-}
	EH	2.159^{+}	0.045^{+}	1.921^{+}	8.481^{+}	13.264^{+}	1.070^{+}	0.821^{+}	7.382^{+}	8.994^{+}	184.492^{+}	-

CONCLUSIONS

In simulated and maize data the results were similar: the residual variance with few markers is lower than with the 1000 markers and the accuracy with few markers is bigger than with 1000 markers.

For maize data multi trait selection, the accuracy increased when the correlation between traits is greater than 0.50 and residual variance decreased when the correlation is greater than 0.70. In this sense, these results showed that marker selection could be used as a first step in genome wide selection, improving the prediction and compute demand.

ACKNOWLEDGEMENTS

Acknowledgements to Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq – Brasil that funded the scholarship, and Fapemig, Capes and CNPq - Brasil to financial support.

RESUMO: O objetivo deste trabalho foi comparar o efeito da precisão e da variância residual na seleção genômica ampla utilizando a seleção de marcadores, bem como utilizando o efeito da seleção indireta, utilizando dados simulados e reais. Foram usados simulados de uma amostra com 200 indivíduos com 1.000 marcadores moleculares na população F2. Os dados reais foram obtidos em milho com população F2 com 441 indivíduos e genotipagem com 261 marcadores SSR. Foram avaliados 11 caracteres (comprimento da espiga, largura da espiga, número da linha, grãos por linha, peso de 100 grãos, peso da espiga, produtividade de grãos, comprimento da espiga, número de espigas, altura da planta e altura da espiga). Todos os dados foram analisados usando o método rrBLUP, sendo realizada 10 vezes a validação cruzada. Em dados simulados e de milho, os resultados foram semelhantes: a variância residual com poucos marcadores é menor do que com os marcadores 1000 e a precisão com poucos marcadores é maior do que com os marcadores 1000. Para a seleção multi-característica dos dados do milho, a precisão aumentou quando a correlação entre as características é maior que 0,50 e a variância residual diminuiu quando a correlação é maior que 0,70. Nesse sentido, esses resultados mostraram que a seleção de marcadores poderia ser usada como um primeiro passo na seleção genômica ampla, melhorando a previsão e a demanda computacional.

PALAVRAS-CHAVE: Seleção de genoma. Milho. Simulação. Análise de dados. Melhoramento de plantas.

REFERENCES

BERNARDO, R. Genome wide Selection with Minimal Crossing in Self-Pollinated Crops. **Crop Science,** v.50, n.2, p.624–627, 2010. https://doi.org/10.2135/cropsci2009.05.0250.

BROMAN, K. W.; SPEED, T. P. A model selection approach for the identification of quantitative trait loci in experimental crosses. **Journal of the Royal Statistical Society,** v. 64, n.4, p.641–656, 2002. https://doi.org/10.1111/1467-9868.00354

CALUS, M. P. L.; VEERKAMP, R. F. Accuracy of multi-trait genomic selection using different methods. **Genetics Selection Evolution,** v.43, n.26, p.1-14, 2011. https://doi.org/10.1186/1297-9686-43-26.

CHE, X.; XU, S. Significance Test and Genome Selection in Bayesian Shrinkage Analysis. **International Journal of Plant Genomics,** v.2010, p.1-10, 2010. https://doi.org/10.1155/2010/893206.

CRUZ, C. D. Genes: a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum Agronomy**, v.35, n.3, p. 271-76, 2013. http://dx.doi.org/10.4025/actasciagron.v35i3.21251.

DAETWYLER, H. D.; CALUS, M. P. L.; PONG-WONG, R.; DE LOS CAMPOS, G.; HICKEY, J. M. Genomic prediction in animals and plants: simulation of data, validation, reporting, and bench-marking. **Genetics**, v.193, n.2 ,p.347-365, 2013. https://doi.org/10.1534/genetics.112.147983.

DAETWYLER, H. D.; HICKEY, J. M.; HENSHALL, J. M.; DOMINIK, S.; GREDLER, B.; VAN DER WERF, J. H. D.; HAYES, B. J. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. **Animal Production Science,** v.50, n.11-12, p.1004-1010, 2010. https://doi.org/10.1071/AN10096.

DE LOS CAMPOS, G.; HICKEY, J. M.; PONG-WONG, R.; DAETWYLER, H. D.; CALUS, M. P. L. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, v.193, n.2, p.327-345, 2013. https://doi.org/10.1534/genetics.112.143313.

ENDELMAN, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. **The Plant Genome**, v.4, n.3, p.250-255, 2011. https://doi.org/10.3835/plantgenome2011.08.0024.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics.** Addison Wesley Longman, Edinburgh, SC, USA. 1996.

GARRICK, D.J. The nature, scope and impact of genomic prediction in beef cattle in the United States. **Genetics Selection Evolution,** v.43, n.17, p.1-11, 2011. https://doi.org/10.1186/1297-9686-43-17.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. Genomic Selection Using Low-Density Marker Panels. **Genetics,** v.182, n.1, p.343-353, 2009. https://doi.org/10.1534/genetics.108.100289

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M.E. Invited review: Genomic selection in dairy cattle: progress and challenges. **Journal of dairy science**, v.92, n.2, p.433-443, 2009. https://doi.org/10.3168/jds.2008-1646.

HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J. L. Genomic selection for crop improvement. **Crop Science,** v.49, n.1, p.1-12, 2009. https://doi.org/10.2135/cropsci2008.08.0512.

JARQUÍN, D.; KOCAK, K.; POSADAS, L.; HYMA, K.; JEDLICKA, J.; GRAEF, G.; et al. Genotyping by sequencing for genomic prediction in a soybean breeding population. **BMC Genomics,** v.15, n.740, p.1-10, 2014. https://doi.org/10.1186/1471-2164-15-740.

JENA, K. K.; MACKILL, D. J. Molecular markers and their use in marker-assisted selection in rice. **Crop Science**, v.48, n.4, p.1266-1276, 2008. https://doi.org/10.2135/cropsci2008.02.0082.

JIA, Y.; JANNINK, J. L. Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. **Genetics**, v.192, n.4, p.1513–1522, 2012. https://doi.org/10.1534/genetics.112.144246.

LEE, S. H.; VAN DER WERF, J. H. J.; HAYES, B. J.; GODDARD, M. E.; VISSCHER, P. M. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. **Plos Genetics,** v.4, n.10e1000231, p. 1-11, 2008. https://doi.org/10.1371/journal.pgen.1000231.

MEUWISSEN, T. H. E., HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v.157, n.4, p.1819–1829, 2001. Available from: https://www.genetics.org/content/157/4/1819.long.

SPINDEL, J.; BEGUM, H.; AKDEMIR, D.; VIRK, P.; COLLARD, B.; REDOÑA, E. et al. Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. **Plos Genetics,** v.11, n.6, p.1-15, 2015. https://doi.org/10.1371/journal.pgen.1004982.

TEAM, R. C. **R: A language and environment for statistical computing**. 2012.

THOMSON, M. J. High-Throughput SNP Genotyping to Accelerate Crop Improvement. **Plant Breeding Biotechnology**, v.2, n.3, p.195–212, 2014. http://dx.doi.org/10.9787/PBB.2014.2.3.195.

VAZQUEZ, A. I..; ROSA, G. J. M.; WEIGEL, K. A.; DE LOS CAMPOS, G.; GIANOLA, D.; ALLISON, D. B. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. **Journal Dairy Science,** v.93, n.12, p.5942–5949, 2010. https://doi.org/10.3168/jds.2010-3335.

WRAY, N. R.; GODDARD, M. E.; VISSCHER, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. **Genome Research,** v.17, n.1, p.1520-1528, 2007. https://doi.org/10.1101/gr.6665407.