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

# Improving accuracy of genomic prediction in maize single-crosses through different kernels and reducing the marker dataset

Massáine Bandeira e Sousa

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

Piracicaba
2017

Massáine Bandeira e Sousa
Bachelor in Biological Sciences

# Improving accuracy of genomic prediction in maize single-crosses through different kernels and reducing the marker dataset

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

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

Piracicaba
2017

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

Sousa, Massáine Bandeira e
Improving accuracy of genomic prediction in maize single-crosses through different kernels and reducing the marker dataset / Massáine Bandeira e Sousa. - - Piracicaba, 2017.
89 p.

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

1 Seleção genômica 2. Kernel Gaussiano 3. GBLUP 4. Interação genótipo x ambiente I. Título

*To my dear mother Marlene, who always motivating me to follow my dreams, for all love and dedication. To my love Luciano Rogério.*

## ACKNOWLEDGEMENT

Firstly, I would like to thank my mother Marlene and my brother Lucas for all love, understanding, and support during this Ph.D. and in all my life.

An especial thank to my fiance Luciano Rogério. My gratitude for discussing the analisys, review my thesis writing, support my dreams, and for being the reason of all my smiles.

I would like to express my gratitude to my advisor Prof. Roberto Fritsche Neto for the support during my Ph.D. research and for all knowledge shared. Thank you for making me a better scientist. It was a pleasure be part of your group during these four years.

Besides my advisor, my sincere thanks also to Dr. Jose Crossa, who provided me an opportunity to join his team at BSU/CIMMYT as a visiting student. Thank you for friendship and teaching me so many things and for help me to value myself.

I would like to thank Miguel Camargo for their assistance, making possible the development of lab's experiments.

To my labmates, Patric Pinho, Felipe Couto, Miriam Vidotti, Italo Granato, Felipe Matias, Leandro Mendonça, Evellyn Giselly, Danilo Lyra, Giovanni Galli, Julia Morosini, thank you for all knowledge shared and for all the fun we have had in the last four years. An especially thank to Roberta Castilho, Andreza Jardelino, and my biology friends for the friendship, and laughs.

I also thank the labmates during my internship at CIMMYT, for welcome and support.

My sincere thank to the lovely couple, Emily and Carlos Cortez, who was my family during my time abroad.

I would like to thank "Luiz de Queiroz" College of Agriculture for the opportunities granted to me along these years.

I truly thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the financial support, which made possible my studies in Brazil and México.

CONTENTS

## RESUMO

**Aprimorando a acurácia da predição genômica em híbridos de milho através de diferentes kernels e redução do subconjunto de marcadores**

No melhoramento de plantas, a predição genômica (PG) é uma eficiente ferramenta para aumentar a eficiência seletiva de genótipos, principalmente, considerando múltiplos ambientes. Esta técnica tem como vantagem incrementar o ganho genético para características complexas e reduzir os custos. Entretanto, ainda são necessárias estratégias que aumentem a acurácia e reduzam o viés dos valores genéticos genotípicos. Nesse contexto, os objetivos foram: *i*) comparar duas estratégias para obtenção de subconjuntos de marcadores baseado em seus efeitos em relação ao seu impacto na acurácia da seleção genômica; *ii*) comparar a acurácia seletiva de quatro modelos de PG incluindo o efeito de interação genótipo × ambiente (G×A) e dois kernels (GBLUP e Gaussiano). Para isso, foram usados dados de um painel de diversidade de arroz (RICE) e dois conjuntos de dados de milho (HEL e USP). Estes foram avaliados para produtividade de grãos e altura de plantas. Em geral, houve incremento da acurácia de predição e na eficiência da seleção genômica usando subconjuntos de marcadores. Estes poderiam ser utilizados para construção de *arrays* e, consequentemente, reduzir os custos com genotipagem. Além disso, utilizando o kernel Gaussiano e incluindo o efeito de interação G×A há aumento na acurácia dos modelos de predição genômica.


Palavras-chave: Seleção genômica; Kernel Gaussiano; GBLUP; Interação genótipo × ambientes

## ABSTRACT

**Improving accuracy of genomic prediction in maize single-crosses through different kernels and reducing the marker dataset**

In plant breeding, genomic prediction (GP) may be an efficient tool to increase the accuracy of selecting genotypes, mainly, under multi-environments trials. This approach has the advantage to increase genetic gains of complex traits and reduce costs. However, strategies are needed to increase the accuracy and reduce the bias of genomic estimated breeding values. In this context, the objectives were: *i*) to compare two strategies to obtain markers subsets based on marker effect regarding their impact on the prediction accuracy of genome selection; and, *ii*) to compare the accuracy of four GP methods including genotype × environment interaction and two kernels (GBLUP and Gaussian). We used a rice diversity panel (RICE) and two maize datasets (HEL and USP). These were evaluated for grain yield and plant height. Overall, the prediction accuracy and relative efficiency of genomic selection were increased using markers subsets, which has the potential for build fixed arrays and reduce costs with genotyping. Furthermore, using Gaussian kernel and the including G×E effect, there is an increase in the accuracy of the genomic prediction models.


Keywords: Genomic selection; Gaussian kernel; GBLUP; Genotype × environment interaction

# 1. INTRODUCTION

Recent technologies for large-scale genotyping and the development of statistical tools have been used to increase selection in breeding programs. Molecular marker-assisted breeding, including marker-assisted selection (MAS) and genomic prediction (GP), in combination with high-throughput and precise phenotyping, can significantly accelerate the development of new varieties (Xu et al. 2012). MAS is a breeding approach based on significant associations of markers and target traits where only the significant markers are used for selection (Xu et al. 2012). However, quantitative traits are affected by many genes, and the limited of MAS is due to the genetic variance proportion explained by the quantitative trait loci (QTL). In GP, breeding values of progenies can be predicted by regressing phenotypic values on all available markers (Meuwissen et al. 2001).

According to Meuwissen et al. (2001), if a marker is in linkage disequilibrium with a QTL, same markers alleles will be correlated with positive effects on the quantitative trait, across all families, and can be used without the need to establish linkage phase in each family. In addition, GP is based on two phases, training and selection (Jonas and de Koning 2013). In training phase markers effects will be estimated based on a phenotypic data of a training population and then tested in a validation population, in order to verify the prediction accuracy of the model (Newell and Jannink 2014). In selection phase, genomic estimated breeding values (GEBV's) of the segregating population are predicted, in which only the individuals are genotyped. GP utilizes the advantage of high genome coverage based on molecular markers, simultaneously adjusting the genetic effects for all the markers, without using statistical tests. In addition, it allows to predict the performance of new breeding parents in early generations and generate crossing and selections based on the prediction model (Jannink et al. 2010).

Genomic prediction was first proposed in animal breeding (Meuwissen et al. 2001), and then applied to plant breeding (Crossa et al. 2010, 2011; de los Campos et al. 2010). Several genomic prediction studies were applied to simulated and real plant breeding data (Bernardo and Yu 2007; Crossa et al. 2010, 2011; de los Campos et al. 2013; Massman et al. 2013; Perez-Rodriguez et al. 2013; Beyene et al. 2015). In general, these studies showed good prediction accuracies for complex traits. In plant breeding, the GP provides opportunities to increase the genetic gain of complex traits per unit of time and reduce costs (Bassi et al. 2015; Heffner et al. 2009; de Oliveira et al. 2012). In a wheat program, genomic prediction schemes have similar costs to phenotypic selection, but the potential of increasing the gain per unit time should be the real driver (Bassi et al. 2015).

Several factors influence the accuracy of genomic prediction, such as genetic architecture, non-additive effects, models, the presence of G×E interaction, and marker density (Pérez-Elizalde et al. 2015; Spindel et al. 2015; Cuevas et al. 2016; Ma et al. 2016). Many statistical methods have been proposed to estimate the GEBVs in the training population. The most common method used is the genomic best linear unbiased prediction (GBLUP) model, which uses the genetic marker information for computing associations between individuals by the Genomic Relationship Matrix (GRM, G) (Habier et al. 2007). The linear GBLUP model capture additive effects between markers. However, complex traits are affected by non-additive effects due to dominance and genetic interactions (epistasis) and their interaction with environment. Thus, non-parametric and semi-parametric methods modeling the relationship between the phenotype and markers in a GS context were proposed by Gianola et al. (2006). Non-parametric methods are able to capture small epistatic interactions without explicitly modeling them. An example of a semi-parametric method is the Reproducing Kernel Hilbert Space (RKHS), that uses a kernel function to convert the marker matrix into a set of distances between pairs of individuals. Therefore, reducing the size of the parametric space and thus can capture small complex interactions between genes (Heslot et al. 2012; Gianola et al. 2014). Thus, RKHS methods had potential to increase accuracy prediction (Perez-Rodriguez et al. 2013; Pérez-Elizalde et al. 2015).

Another factor that influences the GP is the presence of genotype × environment interaction (G×E) effect. Genomic-enabled prediction models were originally developed for a single trait in a single environment. However, multi-environment plant breeding trials are routinely conducted to estimate and take advantage of genotype × environment interaction (G×E). Therefore, to implement GS strategies in plant breeding, G×E needs to be estimated, modeled, and predicted. Resende Júnior et al. (2012) and Windhausen et al. (2012) observed that environment had a high influence on the prediction accuracy of the model when the effects were validated in different environments from where they were estimated. It shows the impact of G×E in the prediction of genotype performance. Thus, Burgueño et al. (2012) extended the GBLUP methodology to incorporate and model G×E effects. The Bayesian model of Jarquín et al. (2014) is another GBLUP extension that introduces main and interaction markers effects and environmental covariables via covariance structures. Heslot et al. (2014) proposed using crop modeling for assessing genomic G×E. Recently, Lopez-Cruz et al. (2015) proposed a GBLUP prediction model that explicitly models G×E and marker × environment interaction (M×E) where marker effects and genomic values are partitioned into components that are stable across environments (main effects) and others that are environment-specific (interactions). In general,

studies have shown that modeling G×E can give substantial gains in prediction accuracy (Burgueño et al., 2012; Heslot et al., 2014; Jarquín et al., 2014; Crossa et al., 2016). Cuevas et al. (2016) compared methods that applied the G×E interaction GS model of Lopez-Cruz et al. (2015) using a linear kernel (GBLUP) and a nonlinear Gaussian kernel and observed that methods using multi-environment G×E interaction models with Gaussian kernel showed higher prediction ability.

Current applications of GP are typically based on single nucleotide polymorphism (SNP) genotypes called from SNP array data. Several studies had shown that low-density arrays can improve the accuracy of genomic prediction (Moser et al. 2010; Spindel et al. 2015; Hoffstetter et al. 2016; Ma et al. 2016). It was demonstrated that high genotyping density does not always increase accuracy and markers subset sometimes outperform the entire dataset (Zhang et al. 2010; Ma et al. 2016). Standard sets of selected SNPs are a tendency for low-cost of genotyping in plant breeding. Therefore, it can develop alternative genotyping platforms less expensive and more repeatable using selected marker subsets.

With the promise of accurate predictions, the genomic prediction has been quickly implemented by several breeding programs. However, strategies are still needed to make the implementation of genomic prediction reducing costs with genotyping and with more accurate predictions. In this context, the objectives were: *i*) to compare two strategies to obtain markers subset based on marker effect regarding their impact on the prediction accuracy of genome selection; and *ii*) compare the accuracy of four GP methods including genotype × environment interaction and two kernels (GBLUP and Gaussian).

## REFERENCES

Bassi FM, Bentley AR, Charmet G, et al (2015) Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). Plant Sci 242:23–36. doi: 10.1016/j.plantsci.2015.08.021

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090. doi: 10.2135/cropsci2006.11.0690

Beyene Y, Semagn K, Mugo S, et al (2015) Genetic Gains in Grain Yield Through Genomic Selection in Eight Bi-parental Maize Populations under Drought Stress. Crop Sci 55:154–163. doi: 10.2135/cropsci2014.07.0460

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

Crossa J, De Los Campos G, Maccaferri M, et al (2016) Extending the marker x Environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Sci 56:2193–2209. doi: 10.2135/cropsci2015.04.0260

Crossa J, De Los Campos G, Pérez P, et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi: 10.1534/genetics.110.118521

Crossa J, Pérez P, de los Campos G, et al (2011) Genomic Selection and Prediction in Plant Breeding. J Crop Improv 25:239–261. doi: 10.1080/15427528.2011.558767

Cuevas J, Crossa J, Soberanis V, et al (2016) Genomic Prediction of Genotype x Environment Interaction Kernel Regression Models. Plant Genome August:1–20. doi: 10.3835/plantgenome2016.03.0024

de los Campos G, Gianola D, Rosa GJM, et al (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb) 92:295–308. doi: 10.1017/S0016672310000285

de los Campos G, Hickey JM, Pong-Wong R, et al (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345. doi: 10.1534/genetics.112.143313

de Oliveira EJ, de Resende MDV, da Silva Santos V, et al (2012) Genome-wide selection in cassava. Euphytica 187:263–276. doi: 10.1007/s10681-012-0722-0

Gianola D, Fernando RL, Stella A (2006) Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. Genetics 173:1761–1776. doi: 10.1534/genetics.105.049510

Gianola D, Weigel KA, Krämer N, et al (2014) Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS One. doi: 10.1371/journal.pone.0091693

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397. doi: 10.1534/genetics.107.081190

Heffner EL, Sorrells ME, Jannink J (2009) Genomic Selection for Crop Improvement. Crops 1–12. doi: 10.2135/cropsci2008.08.0512

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127:463–480. doi: 10.1007/s00122-013-2231-5

Heslot N, Yang H-P, Sorrells ME, Jannink JL (2012) Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci 52:146. doi: 10.2135/cropsci2011.09.0297

Hoffstetter A, Cabrera A, Huang M, Sneller C (2016) Optimizing Training Population Data and Validation of Genomic Selection for Economic Traits in Soft Winter Wheat. G3 (Bethesda) 6:2919–28. doi: 10.1534/g3.116.032532

Jannink J, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–77. doi: 10.1093/bfgp/elq001

Jarquín D, Crossa J, Lacaze X, et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127:595–607. doi: 10.1007/s00122-013-2243-1

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

Lopez-Cruz M, Crossa J, Bonnett D, et al (2015) Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker x Environment Interaction Genomic Selection Model. G3: Genes|Genomes|Genetics 5:569–82. doi: 10.1534/g3.114.016097

Ma Y, Reif JC, Jiang Y, et al (2016) Potential of marker selection to increase prediction accuracy of genomic selection in soybean (Glycine max L.). Mol Breed 36:1–10. doi: 10.1007/s11032-016-0504-9

Massman JM, Jung HJG, Bernardo R (2013) 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. doi: 10.2135/cropsci2012.02.0112

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. doi: 11290733

Moser G, Khatkar MS, Hayes BJ, Raadsma HW (2010) Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet Sel Evol 42:37. doi: 10.1186/1297-9686-42-37

Newell MA, Jannink J-L (2014) Genomic Selection in Plant Breeding. In: Fleury D, Whitford R (eds) Crop Breeding. Springer, Berlin/Heidelberg, pp 117–130

Pérez-Elizalde S, Cuevas J, Pérez-Rodríguez P, Crossa J (2015) Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction. J Agric Biol Environ Stat 20:512–532. doi: 10.1007/s13253-015-0229-y

Perez-Rodriguez P, Gianola D, Gonzalez-Camacho JM, et al (2013) Comparison Between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat. G3: Genes|Genomes|Genetics 2:1595–1605. doi: 10.1534/g3.112.003665

Resende Júnior MFR, Muñoz P, Acosta JJ, et al (2012) Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. New Phytol 193:617–624. doi: 10.1111/j.1469-8137.2011.03895.x

Spindel J, Begum H, Akdemir D, et al (2015) 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. 1–25. doi: 10.5061/dryad.7369p.Funding

Windhausen VS, Atlin GN, Hickey JM, et al (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 (Bethesda) 2:1427–36. doi: 10.1534/g3.112.003699

Xu Y, Lu Y, Xie C, et al (2012) Whole-genome strategies for marker-assisted plant breeding. Mol Breed 29:833–854. doi: 10.1007/s11032-012-9699-6

Zhang Z, Liu J, Ding X, et al (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5:1–8. doi: 10.1371/journal.pone.0012648

# 2. MARKER SELECTION MAY INCREASE PREDICTION ACCURACY AND REDUCE COSTS OF GENOMIC SELECTION


## ABSTRACT

Different methodologies to obtain subsets of markers have been proposed as an alternative to developing a fixed array of SNP, reducing costs and time with genotyping. Thus, the objectives were: (1) to compare two strategies to obtain markers subsets by markers effects: original and re-estimated effects; (2) to compare the prediction accuracy, bias and relative efficiency of a main genotypic effect (MGE) model fitted with a linear kernel Genomic Best Linear Unbiased Predictor, GBLUP (GB), and a nonlinear kernel Gaussian kernel (GK) using markers subsets. The two model-kernel combinations for each marker subsets were applied in a public dataset of rice diversity panel (RICE) and two hybrids maize datasets (HEL and USP). These were evaluated for grain yield and plant height. Decreasing the number of markers, higher estimates of prediction accuracy were obtained by MGE model with GB and GK kernels. Overall, the relative efficiency of genomic to phenotypic selection was increased using markers subsets by the re-estimated effects method. We concluded that from a high-density panel, it is possible to select the most informative markers to improve accuracy and build a low-cost array to implement genomic selection in breeding programs. The best strategy to obtain markers subsets is the re-estimation of the marker effect, which increases the accuracy and reduces the bias.


Keywords: Gaussian kernel; GBLUP; SNP arrays; Relative efficiency

## 2.1. INTRODUCTION

With the advent of higher-density single-nucleotide polymorphisms (SNP) covering the whole genome of many plant species, breeding values can be predicted by regressing phenotypic values on all available markers (Meuwissen et al. 2001; Crossa et al. 2013). Afterward, genomic estimated breeding values (GEBV) are calculated based on these marker effects, which is the main output of genomic prediction (GP). Therefore, many statistical methods have been proposed to estimate the marker effects in the training population.

The first method was the Ridge Regression Best Linear Unbiased Prediction (RR-BLUP) (Meuwissen et al. 2001). This method is based on the assumption in that all marker effects are normally distributed and have equal variance (Meuwissen et al. 2001). The most common method used, based on the same assumptions, and considered as equivalent is the genomic best linear unbiased prediction (GBLUP) model. This model uses the genetic marker information for computing associations between individuals through the Genomic Relationship

Matrix (GRM, G) (Habier et al. 2007). Although it has shown satisfactory results, some authors proposed modifications in order to increase prediction accuracy. Besides the linear GBLUP kernel, semi-parametric methods to model the relationship between the phenotype and markers in a GP context were proposed by Gianola et al. (2006). One of them is the reproducing kernel Hilbert space (RKHS), a semi-parametric method that uses a kernel function to convert the marker matrix into a set of distances between pairs of individuals (Heslot et al. 2012). Several studies have shown that using RKHS methods to obtain genetic relatedness between individuals has potential to increase accuracy prediction, mainly when genotype by environment interaction effect is added in the models (Cuevas et al. 2016; e Sousa et al. 2017).

Currently, high-density SNP chips provide the best coverage of genomes, but they tend to be expensive. On the other hand, genotyping by sequencing (GBS) is an affordable alternative. However, it has shown relatively high rates of missing data and the quality of genotypes tends to be lower since it depends on the genome-wide sequence read depth (Gorjanc et al. 2015b). Thus, another approach is to develop low density and cost fixed arrays, which is highly efficient in smaller breeding programs (Spindel et al. 2015). The advantages of fixed arrays include robust allele calling, cost-effectiveness per data point and speed of genotyping turn-around (Thomson 2014a). Therefore, marker density has been investigated in some studies, which has shown divergent results regarding the use of low or higher density SNPs datasets (Habier et al. 2009; Weigel et al. 2009; Moser et al. 2010; Crossa et al. 2013; Perez-Rodriguez et al. 2013; Tayeh et al. 2015). Furthermore, higher genotyping density does not always increase accuracy and markers subset sometimes outperform the entire dataset (Zhang et al. 2010; Ma et al. 2016).

In this context, different methodologies to obtain subset of markers have been proposed, such as random (evenly distributed across the genome), marker effect, and markers significantly associated with the quantitative trait loci (QTL) (Resende et al. 2012; Spindel et al. 2015; Hoffstetter et al. 2016a). In general, the use of selected markers subset by their effects or positions has provided an efficient strategy to increase accuracy (Vazquez et al. 2010; Resende et al. 2012; Zhang et al. 2015). For instance, Tayeh et al. (2015) decreased the number of markers from 9824 to 2945 by retaining only a single marker per unique map position. They did not observe a reduction in prediction accuracies in any of the traits evaluated. Moreover, Ma et al. (2016) showed that the marker preselection based on haplotype block analyses is an interesting option to reduce the implementation cost of genomic selection.

In this sense, standard sets of selected SNPs are a tendency for low-cost of genotyping in plant breeding. However, for each targeted trait, it is necessary a specific SNP set, because the selected SNPs may vary with traits. In addition, the efficiency of a trait-specific low-density SNP

chip depends critically on the linkage disequilibrium between the SNPs with large estimated effects and the true causative loci that affect the trait of interest (Wu et al. 2016). In this context, we compared two strategies to obtain markers subsets. The former is based on the original marker effect, obtained in the first estimate, using the whole dataset (original effect - ORI strategy). The latter is based on the re-estimated markers effects, that are re-estimated over the dataset reductions (reestimated effect - REE strategy). Thus, the objectives of our study were: (1) to compare two strategies to obtain markers subsets by markers effects: original and re-estimated effects; (2) to compare the prediction accuracy, bias and relative efficiency of a main genotypic effect (MGE) model fitted with a linear kernel Genomic Best Linear Unbiased Predictor, GBLUP (GB), and a nonlinear kernel Gaussian kernel (GK) using markers subsets.

## 2.2. MATERIAL AND METHODS

### 2.2.1. Phenotypic data

For this study, we considered a rice and two maize datasets, using GY (grain yield, ton $ha^{-1}$) and PH (plant height, cm): *i*) the first data is a public dataset available at Rice Diversity platform (http://www.ricediversity.org) (RICE). We used 270 elite breeding lines (F6–F7) from the International Rice Research Institute (IRRI) irrigated rice breeding program evaluated in a single location in Los Baños, Philippines during three years (2009 to 2011) in dry season (Spindel et al. 2015); *ii*) the second dataset is from Helix Sementes® Company, São Paulo, Brazil (HEL). HEL consisted of 452 maize hybrids obtained by crossing 111 inbred lines in a partial diallel. The experimental trial was carried out at five sites located in Southern, Southeast, and Midwest regions of Brazil, during the first growing season of 2014/15. The experimental design was a randomized block with two replicates per genotype and environment; *iii*) the third maize dataset is from University of São Paulo (USP). Data consists of 739 maize hybrids obtained by crossing 49 inbred lines in a partial diallel. The hybrids were evaluated at Piracicaba and Anhumas, São Paulo, Brazil, in 2016. The hybrids were evaluated using an augmented block design, with two commercial hybrids as checks. In both sites, the hybrids were evaluated under an ideal nitrogen (N) level, with 100 kg N $ha^{-1}$. Each plot had a range of 7 m, with 0.50 m spacing between rows and 0.33 m between plants; there was a phenotypic imbalance in both maize datasets.

### 2.2.2. Genotypic data

The RICE inbred lines were genotyped using about 73 K SNPs with GBS. The HEL and USP parent inbred lines were genotyped with an Affymetrix® Axiom® Maize Genotyping Array of 616 K SNPs (Unterseer et al. 2014). Standard quality control (QC) were applied to the data, removing markers with a *Call Rate* $\geq$ 0.95. The remaining missing data in lines were imputed with Synbreed package (Wimmer et al. 2015) using the algorithms from software Beagle 4.0 (Browning and Browning 2008). In the HEL and USP maize datasets, the hybrid genotypes were obtained by genomic information of the parent inbred lines. Then, in RICE, HEL, and USP datasets markers with Minor Allele Frequency (MAF) $\leq$ 0.05 were removed. The final marker matrix was composed by 39,811, 52,811 and 61,824 SNPs for RICE, HEL and USP datasets, respectively.

### 2.2.3. Statistical models

#### 2.2.3.1. Estimating BLUPs

We used a linear mixed model to calculate best linear unbiased predictions (BLUPs) for rice inbred lines and maize hybrids. For grain yield and plant height, BLUPs were obtained over years for RICE dataset, and over environments for HEL and USP dataset. Estimation of random effects values and variance components were performed by Restricted Maximum Likelihood/Best Linear Unbiased Predictor (REML/BLUP) procedure, by fitting the following model:

$$\boldsymbol{y} = \boldsymbol{Xr} + \boldsymbol{Tg} + \boldsymbol{Bs} + \boldsymbol{Hx} + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{y}$ is the response vector, $\boldsymbol{r}$ is the replicate effect considered as fixed, $\boldsymbol{g}$ is the vector of random effect of genotypes, where $\boldsymbol{g} \sim NID(0, I\sigma_g^2)$, $\boldsymbol{s}$ is the vector of fixed effects of environment; $\boldsymbol{x}$ is the vector of random effects of Genotype by Environment interaction, where $\boldsymbol{x} \sim NID(0, I\sigma_x^2)$ and $\boldsymbol{\varepsilon}$ is the vector of error, where $\boldsymbol{\varepsilon} \sim NID(0, I\sigma_\varepsilon^2)$. The $\boldsymbol{X}$, $\boldsymbol{T}$, **B** and $\boldsymbol{H}$ are incidence matrices. Environment effect for HEL and USP dataset is represented by site and for RICE dataset by year.

The significance of random effects was estimated using the Likelihood Ratio Test (LRT) (Gilmour et al. 2009). The variance components estimated for each model effect were used to calculate entry-mean based heritability ($H^2$), using means from each environment (considering each year or site as an environment).

#### 2.2.3.2. Obtaining markers subsets

We used RR-BLUP (Random Regression – Best Linear Unbiased Predictor) method (Endelman 2011) to obtain markers subsets based on SNP effects:

$$\mathbf{y} = 1\boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

where $\mathbf{y}$ is the vector of BLUPs, $\boldsymbol{\mu}$ is intercept, $\mathbf{Z}$ is a matrix of markers, $\boldsymbol{\beta}$ is a vector of marker effects, and $\boldsymbol{\varepsilon}$ is a residual effect. The following distribution were assumed $\beta \sim N(0, \boldsymbol{I}\sigma_\beta^2)$ and $\varepsilon \sim N(0, \boldsymbol{I}\sigma_\varepsilon^2)$.

Then, we applied two strategies to obtain the markers subsets, the reduction of the number of markers by *i*) their original effect (ORI strategy) and *ii*) re-estimation of markers effects (REE strategy). For each strategy, we first obtained a subset of 10000 markers, and from these markers, we obtained and subsequently evaluated five subsets (5000, 2500, 1000, 500, and 100 SNPs) for each strategy. For ORI strategy, the vector of markers effect was first estimated using all the markers and then ranked in decreasing order by their absolute values and selected 10000 SNPs with highest absolute value. Afterward, 5000 with highest absolute value was selected, this step was repeated until the subset of 100 SNPs was obtained. For REE strategy, the subset containing 10000 SNPs was obtained in the same way of ORI strategy. However, the markers effects were re-estimated, ranked in decreasing order by their absolute values and then, we selected 5000 SNPs with highest absolute value. This procedure was repeated until the subset of 100 SNP was obtained.

#### 2.2.3.3. Main genotypic effect (MGE) model

The six marker subsets described above were used to perform genomic prediction, using the main genotypic effect model. This model fits the data separately from each subset and considers the main effect of genotypes. In matrix notation, the model is written as:

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{\varepsilon} \quad (5)$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the response vector, and $y_i$ represents the observations in the $i^{th}$ genotype $(i = 1, \ldots, n)$ in each markers subset; $\mu$ is the general mean; $\boldsymbol{Z}_u$ is the incidence matrix that connects the random genetic effects with phenotypes; $\boldsymbol{u}$ is the random genetic effects, and $\boldsymbol{\varepsilon}$ the residual random effects. The MGE model (5) assumes that the distribution of the $\boldsymbol{u}$ vector is multivariate normal with mean zero and a covariance matrix $\boldsymbol{K}$, with $\boldsymbol{u} \sim N(0, \sigma_u^2\boldsymbol{K})$, where $\sigma_u^2$ is the genetic variance component of $\boldsymbol{u}$ and $\boldsymbol{K}$ is a symmetric, positive semi-definite matrix, denoting the variance-covariance of the genetic values constructed from the genomic molecular

markers by subsets. The error $\boldsymbol{\varepsilon}$ was assumed to be independent of each other and normally distributed $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma_{\varepsilon}^{2}\boldsymbol{I})$. Therefore, $\boldsymbol{u}$ is an approximation of the true unknown genetic values, and $\boldsymbol{\varepsilon}$ captures the residual genetic effects that were not explained by $\boldsymbol{u}$ plus other non-genetic effects that approximate the errors.

The main genotypic effect model was used with two kernel regression methods for each markers subset.

*Main genotypic effect model with GBLUP (MGE-GB):* in the MGE model (5), matrix $\boldsymbol{K}$ was constructed using the linear kernel $\boldsymbol{K} = \frac{\boldsymbol{XX'}}{p}$ proposed by VanRaden (2007, 2008), where $\boldsymbol{X}$ is the standardized matrix of molecular markers for the individuals, of order $n \times p$, where $p$ is the number of markers by subset.

*Main genotypic effect model with Gaussian kernel (MGE-GK):* the Gaussian kernel was defined as $\boldsymbol{K}(\boldsymbol{x}_i\boldsymbol{x}_{i`}) = \exp(-hd_{ii`}^{2})$, where $d_{ii`}$ is the Euclidean distance between the individuals $i^{\text{th}}$ and $i'^{\text{th}}\ (i = 1, \ldots, n_j)$ given by the markers; $h > 0$ is the bandwidth parameter that controls the rate of decay of $\boldsymbol{K}$ values (Pérez-Rodríguez et al., 2012; Cuevas et al., 2016). We used $\boldsymbol{K}(\boldsymbol{x}_i\boldsymbol{x}_{i`}) = \exp(-hd_{ii`}^{2}/\text{median}(d_{ii`}^{2}))$, where $h = 1$ and the median of the distances is used as a scaling factor (Crossa et al., 2010).

### 2.2.4. Variance components, prediction accuracy, and bias parameters

The MGE model fitted with GB and GK methods were used on the whole RICE, HEL, and USP datasets for all the traits. Before fitting the models, the phenotypic data were centered and standardized (i.e., each phenotypic data point was centered by subtracting the overall mean and standardized by dividing by the sample standard deviation). These analyses were performed to derive estimates of variance components and genomic heritability. Since the data was standardized, the summation of the variance components of one specific model will approximate 1.

The prediction accuracy (PA) was performed separately for each trait and marker subset (all markers and the five markers subsets) for MGE-GB and MGE-GK models, and was assessed using thirty random partitions (replications), with 80% of the hybrids/lines comprising the training set (TRN), and the remaining 20% comprising the testing set (TST). For this validation procedure, all the parameters of the models (including variance components resulting from

residual effects and genetic effects) were re-estimated from TRN data in each of the TRN-TST partitions.

For each TRN-TST partition, models were fitted to the TRN data set, and prediction accuracy was assessed by computing Pearson's correlation between predictions and phenotypes in the TST dataset. The same TRN-TST partitions were used to evaluate the prediction accuracy of each of the models. Thus, thirty correlations were computed for each model and trait.

We used two approaches to measuring the bias of GEBVs for ORI and REE strategies, which were the slope coefficient for the regression of vector of phenotypes on GEBVs (Resende et al. 2012), and reliability (Gorjanc et al. 2015a). For the first method, unbiased models are expected to have a slope coefficient of 1, whereas values greater than 1 indicate a biased underestimation in the GEBVs prediction and values smaller than 1 indicate a biased overestimation of the GEBVs. For the second, reliability (REL) was calculated as $r = 1 - \left(PEV_g/\sigma_g^2\right)$, where $PEV$ is obtained as mean prediction error variance of the validation set ($PEV = PEV_g$), where $PEV_g$ stands for the error related to the genetic effect. We used the mean values of both bias strategies estimated from thirty replications were used in the overall model performance comparison.

### 2.2.5. Relative efficiency of genomic prediction

The relative efficiency of genomic prediction for the all proposed scenarios was compared to the traditional phenotypic selection. For this, two parameters were used: *i*) the relative efficiency per cycle ($RE_c$), which was calculated by $r/\sqrt{H^2}$, were $r$ is the accuracy of model and $H$ is the phenotypic heritability of trait (Hoffstetter et al. 2016b); *ii*) the coincidence of selection, which was estimated based on the top 10% and 20% individuals identified by TRN-TST and the individuals with the best phenotypic performance for each trait.

### 2.2.6. Software

We used the rrBLUP-R package (Endelman, 2011) to run the RR-BLUP model. The BGLR package (de los Campos and Pérez-Rodríguez, 2016) was used to carry out the MGE models, using a total of 30,000 MCMC iterations, 5,000 for burn-in, and five for thinning. Appendix B has the R codes used to obtained markers subsets by ORI and REE strategies, and how we fitted the MGE model with GBLUP and Gaussian kernels.

## 2.3. RESULTS

### 2.3.1. Phenotypic data

We observed for both traits, plant height (PH) and grain yield (GY), in the RICE, HEL, and USP dataset, genetic variability according to the likelihood ratio test based on the joint analysis (Appendix A, Table A1). Similarly, random effects of Genotype × Year (RICE) and interaction Genotype × Site (HEL and USP) were significant, except for PH in USP dataset. This significant effect suggests the differential performance of genotypes between sites or year for the tested traits. Entry-mean based heritability was 0.89 (PH) and 0.78 (GY) for RICE, 0.87 (PH) and 0.77 (GY) for HEL, and 0.83 (PH) and 0.59 (GY) for USP dataset (Additional Table S1), reflecting good accuracy of phenotypic evaluation.

### 2.3.2. Variance components of markers subset

Variance components (residual and genetic) and genomic heritability were obtained by MGE model through GBLUP and GK kernels (MGE-GB and MGE-GK).

*RICE dataset*

Using markers subsets obtained by REE strategy, for PH and GY trait, the estimated residual variance components were smaller than those obtained by ORI strategy (Figures 1A and 1D). The results taken through of the MGE-GK and MGE-GB models indicated that the use of small markers subset induced a larger reduction in the estimated residual variance. The variance component of genetic effects for each subset increased when MGE-GK model used rather than MGE-GB model (Figures 1B and 1E). Furthermore, by the REE strategy, the genetic effects were slightly higher than ORI strategy, except for 500 markers dataset. The genomic heritability ranged from 0.44 (all markers MGE-GB model) to 0.90 (markers subset by REE strategy). When only 500 to 5000 markers were used to compute the GB or GK-matrix, the values ranged from 0.75 to 0.90. For PH, the results showed similar tendency with GY (Figure 1E).

*HEL dataset*

The results of residual variance showed, for PH and GY, that MGE-GK tends to fit better the data than MGE-GB. It is because no consistent difference was found using all or subset of markers through the former method (Figures 2A and 2D). For GY, the genetic variance for MGE-GB (range from 0.20 to 0.30) was always smaller than MGE-GK (range from 1.20 to

0.60) (Figure 2E). Genetic variance was higher, around 1.30 to MGE-GK and 0.80 to MGE-GB, using all markers for PH (Figure 2B). However, using markers subsets by MGE-GK models, the genetic variances were always higher (around 0.70) than those obtained by MGE-GB (range from 0.50 to 0.30). Hence, genomic heritabilities by the former were greater than by the latter for all subsets of markers and both traits (Figures 2C and 2F).

*USP dataset*

For GY, we observed higher values for residual variance using MGE-GB than by MGE-GK models (Figure 3D). Moreover, the genetic variance was smaller, mainly through MGE-GB than by MGE-GK (Figure 3E). Using all markers, we can see slightly higher values than using markers subset. The genomic heritability was lower using the former (range from 0.20 to 0.10) than the latter (range from 0.30 to 0.45) (Figure 3F). The residual and genetic variance and heritability for PH followed the same trend of GY (Figure 3A, 3B, and 3C).

### 2.3.3. Prediction accuracy and bias of GEBVs of markers subset

*RICE dataset*

Results for GY showed that the prediction accuracy (PA) using markers subsets was better than using all markers (Figure 4). The PA using all markers was 0.297 for MGE-GB and 0.291 for MGE-GK. However, when the markers subsets were used to predict, we observed a substantial increase in PA. For instance, for MGE-GB and MGE-GK models, the increase using 500 markers was 222% and 213%, respectively. This subset had the better PA and was obtained by REE strategy. Overall, comparing the two approaches, we verified REE was better than ORI. The PA by MGE-GB ranged from 0.622 (100 markers) to 0.759 (1000 markers) via ORI strategy and ranged from 0.819 (100 markers) to 0.959 (500 markers) via REE strategy. In general, the MGE model using GB kernel is slightly better than GK kernel. For PH, the results are similar with GY (Figure 4). The PA using all markers was low, where for MGE-GB model was 0.382 and for MGE-GK was 0.406. The increase in PA using markers subset was 156% and 132% using MGE-GB and MGE-GK models, respectively. The marker subset that showed the best PA (0.978 by MGE-GB and 0.945 by MGE-GK) was with 500 markers obtained by REE strategy.

The coefficient of regression (slope) of phenotype on GEBVs was calculated as a measurement of the bias of each strategy-model-subset (Appendix A, Table A2). Ideally, a value of $\beta = 1$ indicates no bias in the prediction. For GY and PH, comparing ORI or REE approaches and MGE-GB or MGE-GK models the slopes for MGE-GB model with ORI

strategy were close to one, indicating no consistent bias in the prediction. In general, the fewer the markers, the smaller the bias. Furthermore, the value of β derived from MGE-GK model with REE strategy was higher for all traits (average across traits equal to 1.49). Moreover, the use of subsets with fewer markers resulted in slightly higher reliabilities for both traits (Appendix A, Table A2). Overall, for PH the REE strategy produced slightly better values of reliability than ORI approach.

*HEL dataset*

The results obtained for GY indicated that using markers subsets showed an increase in prediction accuracy (Figure 5). The PA using all markers was intermediate for MGE-GB (0.548) and MGE-GK (0.674). We observed an increase in PA using markers subsets. For instance, using just 500 markers, the PA was 0.676 (MGE-GB), and 5000 markers were 0.721 (MGE-GK), which represents an improvement of 23% and 7% for MGE-GB and MGE-GK models, respectively. Considering the comparison between the two strategies, the REE was better when the prediction used MGE-GK, where the PA ranged from 0.676 (100 markers) to 0.721 (5000 markers). For PH, the PA was high (>0.81) for both models using all and subsets of markers (Figure 5). The increase using markers subset was 8% and 4% for MGE-GB and MGE-GK models, respectively.

The coefficient of regression (slope) for all scenarios was close to one, indicating no consistent bias in the prediction. In general, the reliabilities of the models for GY showed that MGE-GK (0.794 to 0.860) showed higher reliabilities than MGE-GB model (0.664 to 0.809) (Appendix A, Table A2). On the other hand, for PH, all scenarios had similar reliabilities (> 0.809).

*USP dataset*

For both traits, all scenarios presented similar accuracies (Figure 6). The slope for GY and PH traits, for both models and subsets were close to one, indicating no consistent bias in the prediction (Appendix A, Table A2).

### 2.3.4. Relative efficiency of markers subsets to genomic prediction

*RICE dataset*

Considering 10% of an intensity of selection (IS), we observed a higher increase in the coincidence of selection (CS) when markers subsets were used (Table 1 and Figure 7A-7B). For instance, the CS values for GY by MGE-GB model ranged from 26% (all markers) to 81% (500 markers) and by MGE-GK range from 29% (all markers) to 70% (500 markers). The REE strategy selects markers that provide greater coincidence of genomic with phenotypic selection. Under 20% as IS, the CS showed the same tendency of that in 10% IS. However, the values were slightly higher, when MGE-GB model was used.

Regarding the relative efficiency (REc) per cycle, MGE-GB using 2500, 1000, and 500 markers, and MGE-GK using 500 markers subset, obtained by REE strategy, have values equal or greater than 1, showing the efficiency of genomic prediction over the phenotypic selection.

*HEL dataset*

Following the trend described above, when markers subsets are used, we observed an increase in the CS via MGE-GB (Table 1 and Figure 7C-7D). However, this increase is similar with those obtained using MGE-GK. For instance, for PH, the values through MGE-GB ranged from 42% (all markers) to 48% (500 markers) and by MGE-GK from 44% (all markers) to 47% (100 markers). The values considering 20% as IS were slightly higher than by selecting 10%.

Regarding the relative efficiency (REc) per cycle for PH and GY, the values were less than 1, showing that, in these cases, the efficiency of genomic prediction did not exceed phenotypic selection (Table 1 and Figure 7C-7D).

*USP dataset*

Under 10% as IS, for GY, there was a slight increase (4%) in the CS, considering markers subset rather than all markers by MGE-GB model (Table 1 and Figure 7E-7F). MGE-GK prone to produce lower values of CS using markers subset. In general, the CS values were similar for both models, except when all markers are used. Using 20% as IS, there was a higher coincidence via MGE-GK. On the other hand, for PH, the results reveal that the coincidences of selection are similar among the models, methods and intensities of selection.

Considering the relative efficiency (REc) per cycle for PH (0.74 to 0.77) and GY (0.65 to 0.71), the values were less than 1, showing that, in these cases, the efficiency of genomic prediction did not exceed phenotypic selection (Table 1 and Figure 7E-7F).

## 2.4. DISCUSSION

Several studies have documented the benefits of using markers subset to improve the prediction accuracy in genomic selection analyses (Moser et al. 2010; Resende et al. 2012; Spindel et al. 2015; Tayeh et al. 2015; Hoffstetter et al. 2016a). In this research, two strategies were used to obtain markers subsets where the prediction accuracy, using each subset or all markers, was compared by a main genetic effect (MGE) model using GB (GBLUP) and GK (Gaussian) kernels. We found that, by selecting subsets of SNPs, accuracy is considerably increased. Furthermore, the REE strategy is the best to obtain markers subsets, due to the re-estimation of the marker effect, increasing accuracy and reducing bias of GEBVs. According to Porto-Neto et al. (2015), SNP selection would take into account different marker allele frequencies between individuals, and the missing heritability can be largely recovered, leading to improvements of accuracies. In general, many sequence variants associated with complex traits have small effects and low repeatability (Bian and Holland 2017). The ORI and REE strategies allow eliminating many redundant markers with small effect, increasing accuracy. This procedure allows reducing the multicollinearity problem between markers, which occurs especially because markers in close positions are expected to be highly correlated. Similarly, Resende et al. (2012) also used an approach of re-estimating markers effect and found higher performance compared to Bayesian methods for rust disease resistance and wood density traits in Loblolly pine.

Several studies showed that selecting markers from a high density genotyping, can be an effective alternative to genomic prediction (Spindel et al. 2015, 2016; Tayeh et al. 2015; Ma et al. 2016). Another way would be to weight the genomic relationship matrix with previously selected SNPs from genome-wide association studies (GWAS) (Su et al. 2014; Spindel et al. 2016). Zhang et al. (2015) showed that the G-matrix weighted with individual SNPs with strong and robust association signals can effectively improve genomic prediction. VanRaden et al. (2017), observed an increase in accuracy add markers, selected with highest significance test, largest absolute effect, or largest genetic variance contributed by the locus, in high-density SNP chips for animal breeding. However, we used only markers with higher effects to build the GRM and concluded that this approach is an efficient and easy way to be implemented in breeding programs.

Recently for a rice breeding program, Yu et al. (2014) developed accurate fixed arrays with 6 K SNPs from a high density genotyping, selecting representative SNPs obtained by genetic diversity studies and markers located inside genes of important traits, which are evenly distributed on the 12 chromosomes. Furthermore, Spindel et al. (2016) using genomic prediction models that incorporate *de novo* GWAS, with the same rice dataset used in our work, found that approximately 5000 SNPs were effective for prediction, suggesting the possible use of smaller fixed SNP arrays.

In our results, eliminating markers with small effect, from 5000 to 500 SNPs, is a good strategy and has the potential for build a trait-specific low-density SNP chip, reducing costs of genotyping in a breeding program.

Considering a subset with 500 markers obtained by ORI and REE strategies, we observed a wide distribution of them on all chromosomes of maize and rice (Appendix A, Figure A1, A2, and A3). In the RICE dataset, the re-estimation of markers effects provided a wider distribution of markers across the genome, than HEL and USP datasets, allowing a more pronounced increase in PA. Similar methods based on selected markers evenly distributed across the genome was reported in the literature (Spindel et al. 2015; Tayeh et al. 2015; Moser et al. 2010). For instance, Spindel et al. (2015) observed that subsets either distributed evenly throughout the genome was the most important contributor to PA and had less accuracy variance than chosen at random. Moreover, Moser et al. (2010) showed that a chip containing 3000 evenly spaced markers could provide approximately 90% of the accuracy achieved with a high-density.

Based on the maize hybrids datasets, we conclude that using Gaussian kernel we could obtain higher prediction accuracies compared to the linear kernel (Figure 5 and 6). The Gaussian kernel can capture non-additive effects while GBLUP only additive effects (Gianola et al. 2014; Cuevas et al. 2016b). Maize hybrids present high heterosis due to non-additive effects (dominance and epistasis), explaining the results achieved. However, in the RICE dataset, the GBLUP model showed slightly higher prediction accuracies than GK in most scenarios. Furthermore, the smaller number of markers used in genomic prediction, the smaller the difference of PA observed between GBLUP and GK. Therefore, a small number of markers is not able to capture non-additive effects between markers, especially in the complex trait as grain yield. Moreover, the heterosis present in maize hybrids may have contributed to the best result of using GK.

Regarding the bias parameters, we found slight differences among models and strategies (Appendix A, Table A2). According to Porto-Neto et al. (2015), while accuracy affects the ranking of individuals, bias affects the range of estimated breeding values. The authors worked with animal data and found lower bias using SNPs selected from related individuals. We found similar results, where the differences between strategies and kernels were larger in GEBVs bias and prediction accuracy when applied to a population with more genetic diversity. In addition, Ma et al. (2015) analyzed same strategies to reduce the bias of predicted genetic trend and affirmed that this parameter should be taken into consideration in the validation of genomic predictions. Nonetheless, for RICE dataset the markers subsets obtained by ORI strategy

evaluated by GBLUP kernel, had performed better than GK concerning reducing bias of GEBVs.

Comparing the coincidence of the genomic with the phenotypic selection and relative efficiency per cycle (REc), we verified more coincidence using markers subsets (Table 1). Although the coincidence was not 100%, the genomic selection is still a promising alternative, mainly because it has the advantage of reducing the time of a breeding cycle (Bassi et al. 2015). Overall, our results showed that from with a high-density panel, it is possible to select the most informative markers to improve accuracy and build a low-cost array to implement genomic selection in breeding programs. The best strategy to obtain markers subsets is the re-estimation of the marker effect, which increases the accuracy and reduces the bias.

## 2.5. CONCLUSION

From with a high-density panel, it is possible to select the most informative markers to improve accuracy and build a low-cost array to implement genomic selection in breeding programs. The best strategy to obtain markers subsets is the re-estimation of the marker effect, which increases the accuracy and reduces the bias.

## REFERENCES

Bassi FM, Bentley AR, Charmet G, et al (2015) Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). Plant Sci 242:23–36. doi: 10.1016/j.plantsci.2015.08.021

Bian Y, Holland JB (2017) Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. Heredity (Edinb) 118:585–593. doi: 10.1038/hdy.2017.4

Browning BL, Browning SR (2008) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210–223. doi: 10.1016/j.ajhg.2009.01.005

Crossa J, Pérez P, Hickey J, et al (2013) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb) 112:48–60. doi: 10.1038/hdy.2013.16

Cuevas J, Crossa J, Montesinos-Lopez O, et al (2016a) Bayesian Genomic Prediction with Genotype × Environment Interaction Kernel Models. G3 (Bethesda) g3.116.035584. doi: 10.1534/g3.116.035584

Cuevas J, Crossa J, Soberanis V, et al (2016b) Genomic Prediction of Genotype x Environment Interaction Kernel Regression Models. Plant Genome August:1–20. doi: 10.3835/plantgenome2016.03.0024

de los Campos G, Perez-Rodriguez P (2016) BGLR: Bayesian generalized linear regression. R package version 1.0.5.

e Souza MB, Cuevas J, Couto EG de O, et al (2017) Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype × Environment Interaction. G3: Genes|Genomes|Genetics g3.117.042341. doi: 10.1534/g3.117.042341

Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome J 4:250–255. doi: 10.3835/plantgenome2011.08.0024

Gianola D, Fernando RL, Stella A (2006) Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. Genetics 173:1761–1776. doi: 10.1534/genetics.105.049510

Gianola D, Weigel KA, Krämer N, et al (2014) Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS One. doi: 10.1371/journal.pone.0091693

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0. VSN International, Hemel Hempstead

Gorjanc G, Bijma P, Hickey JM (2015a) Reliability of pedigree-based and genomic evaluations in selected populations. Genet Sel Evol 47:65. doi: 10.1186/s12711-015-0145-1

Gorjanc G, Cleveland MA, Houston RD, Hickey JM (2015b) Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet Sel Evol 47:12. doi: 10.1186/s12711-015-0102-z

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397. doi: 10.1534/genetics.107.081190

Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. Genetics 182:343–353. doi: 10.1534/genetics.108.100289

Heslot N, Yang H-P, Sorrells ME, Jannink JL (2012) Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci 52:146. doi: 10.2135/cropsci2011.09.0297

Hoffstetter A, Cabrera A, Huang M, Sneller C (2016a) Optimizing Training Population Data and Validation of Genomic Selection for Economic Traits in Soft Winter Wheat. G3 (Bethesda) 6:2919–28. doi: 10.1534/g3.116.032532

Hoffstetter A, Cabrera A, Sneller C (2016b) Identifying quantitative trait loci for economic traits in an elite soft red winter wheat population. Crop Sci 56:547–558. doi: 10.2135/cropsci2015.06.0332

Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. Genetics 201:759–768. doi: 10.1534/genetics.115.177907

Ma P, Lund MS, Nielsen US, et al (2015) Single-step genomic model improved reliability and reduced the bias of genomic predictions in Danish Jersey. J Dairy Sci 98:9026–9034. doi: 10.3168/jds.2015-9703

Ma Y, Reif JC, Jiang Y, et al (2016) Potential of marker selection to increase prediction accuracy of genomic selection in soybean (Glycine max L.). Mol Breed 36:1–10. doi: 10.1007/s11032-016-0504-9

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. doi: 11290733

Moser G, Khatkar MS, Hayes BJ, Raadsma HW (2010) Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet Sel Evol 42:37. doi: 10.1186/1297-9686-42-37

Perez-Rodriguez P, Gianola D, Gonzalez-Camacho JM, et al (2013) Comparison Between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat. G3: Genes|Genomes|Genetics 2:1595–1605. doi: 10.1534/g3.112.003665

Pérez-Rodríguez P, Gianola D, González-Camacho JM, et al (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 (Bethesda) 2:1595–605. doi: 10.1534/g3.112.003665

Porto-Neto LR, Barendse W, Henshall JM, et al (2015) Genomic correlation: harnessing the benefit of combining two unrelated populations for genomic selection. Genet Sel Evol 47:84. doi: 10.1186/s12711-015-0162-0

Resende MFR, Munoz P, Resende MD V., et al (2012) Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). Genetics 190:1503–1510. doi: 10.1534/genetics.111.137026

Spindel J, Begum H, Akdemir D, et al (2015) 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. 1–25. doi: 10.5061/dryad.7369p.Funding

Spindel JE, Begum H, Akdemir D, et al (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity (Edinb) 116:395–408. doi: 10.1038/hdy.2015.113

Su G, Christensen OF, Janss L, Lund MS (2014) Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J Dairy Sci 97:6547–6559. doi: 10.3168/jds.2014-8210

Tayeh N, Klein A, Le Paslier M-C, et al (2015) Genomic Prediction in Pea: Effect of Marker Density and Training Population Size and Composition on Prediction Accuracy. Front Plant Sci 6:1–11. doi: 10.3389/fpls.2015.00941

Thomson MJ (2014a) High-Throughput SNP Genotyping to Accelerate Crop Improvement. Plant Breed Biotechnol 2:195–212. doi: 10.9787/PBB.2014.2.3.195

Thomson MJ (2014b) High-Throughput SNP Genotyping to Accelerate Crop Improvement. Plant Breed Biotechnol 2:195–212. doi: 10.9787/PBB.2014.2.3.195

Unterseer S, Bauer E, Haberer G, et al (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15:823. doi: 10.1186/1471-2164-15-823

VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. J Dairy Sci 91:4414–4423. doi: 10.3168/jds.2007-0980

VanRaden PM (2007) Genomic Measures of Relationship and Inbreeding. Interbull Annu Meet Proc 37:33–36. doi: 10.1007/s13398-014-0173-7.2

VanRaden PM, Tooker ME, O'Connell JR, et al (2017) Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol 49:32. doi: 10.1186/s12711-017-0307-4

Vazquez a I, Rosa GJM, Weigel K a, et al (2010) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J Dairy Sci 93:5942–5949. doi: 10.3168/jds.2010-3335

Weigel K a, de los Campos G, González-Recio O, et al (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J Dairy Sci 92:5248–5257. doi: 10.3168/jds.2009-2092

Wimmer AV, Auinger H, Albrecht T, et al (2015) synbreed: Framework for the analysis of genomic prediction data using R. 1–43.

Wu XL, Xu J, Feng G, et al (2016) Optimal design of low-density SNP arrays for genomic prediction: Algorithm and applications.

Xia J, Wu Y, Fang H, et al (2015) Improving the Efficiency of Genomic Selection in Chinese Simmental beef cattle. bioRxiv 22673. doi: 10.1101/022673

Yu H, Xie W, Li J, et al (2014) A whole-genome SNP array (RICE6K) for genomic breeding in rice. Plant Biotechnol J 12:28–37. doi: 10.1111/pbi.12113

Zhang Z, Erbe M, He J, et al (2015) Accuracy of Whole Genome Prediction Using a Genetic Architecture Enhanced Variance-Covariance Matrix. G3: Genes|Genomes|Genetics 5:615–627. doi: 10.1534/g3.114.016261

Zhang Z, Liu J, Ding X, et al (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5:1–8. doi: 10.1371/journal.pone.0012648

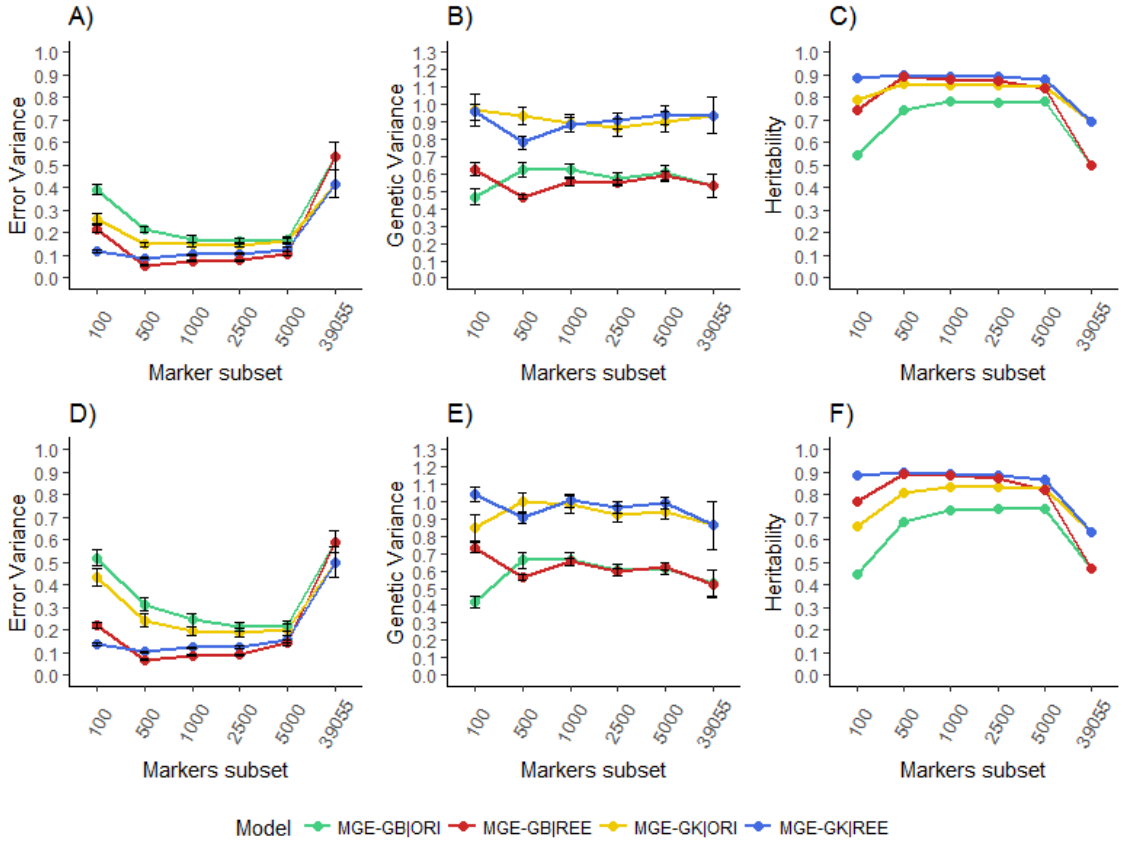


**Figure 1**. RICE dataset. Variance components and genomic heritability through MGE-GB and MGE-GK models considering markers subsets (horizontal axis) obtained by original effect (ORI) and reestimated effect (REE) strategies, for plant height (A) Residual variance, (B) genetic variance, and (C) heritability; and for grain yield (D) Residual variance, (E) genetic variance, and (F) heritability.

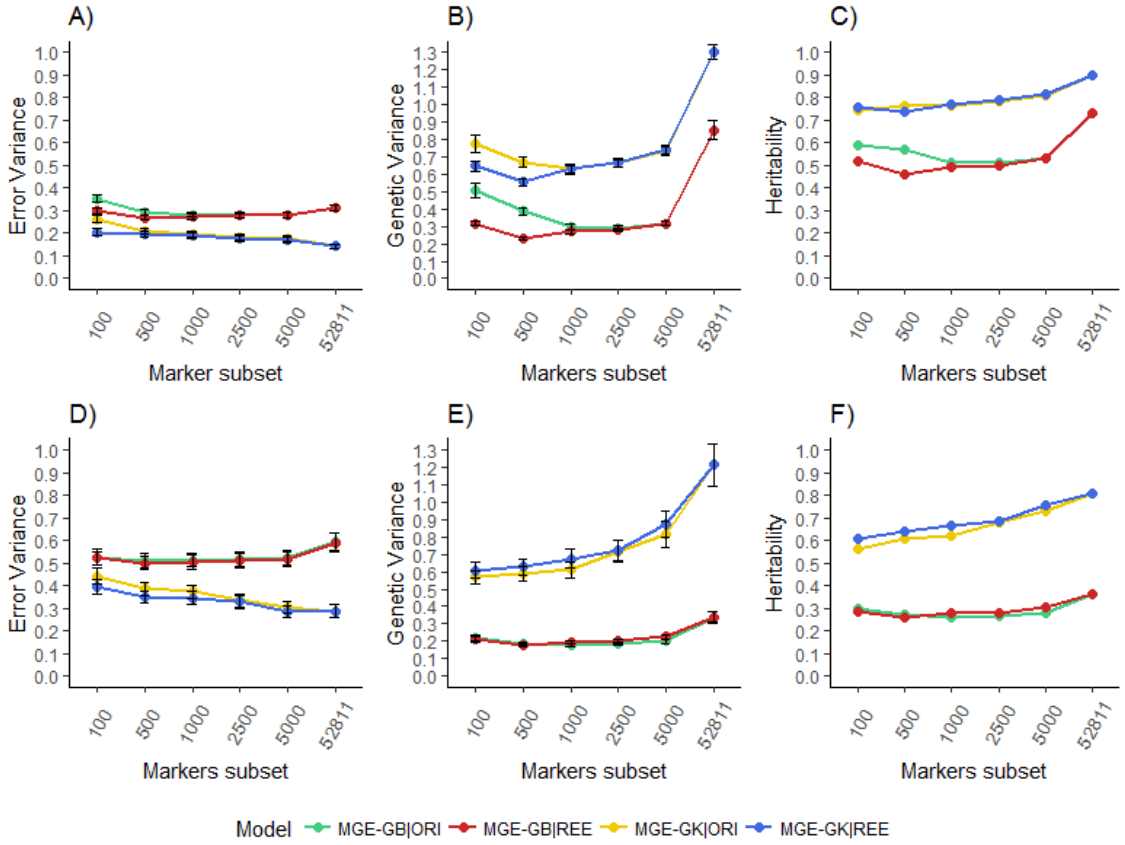


**Figure 2**. HEL dataset. Variance components and genomic heritability through MGE-GB and MGE-GK models considering markers subsets (horizontal axis) obtained by original effect (ORI) and reestimated effect (REE) strategies, for plant height (A) Residual variance, (B) genetic variance, and (C) heritability; and for grain yield (D) Residual variance, (E) genetic variance, and (F) heritability.

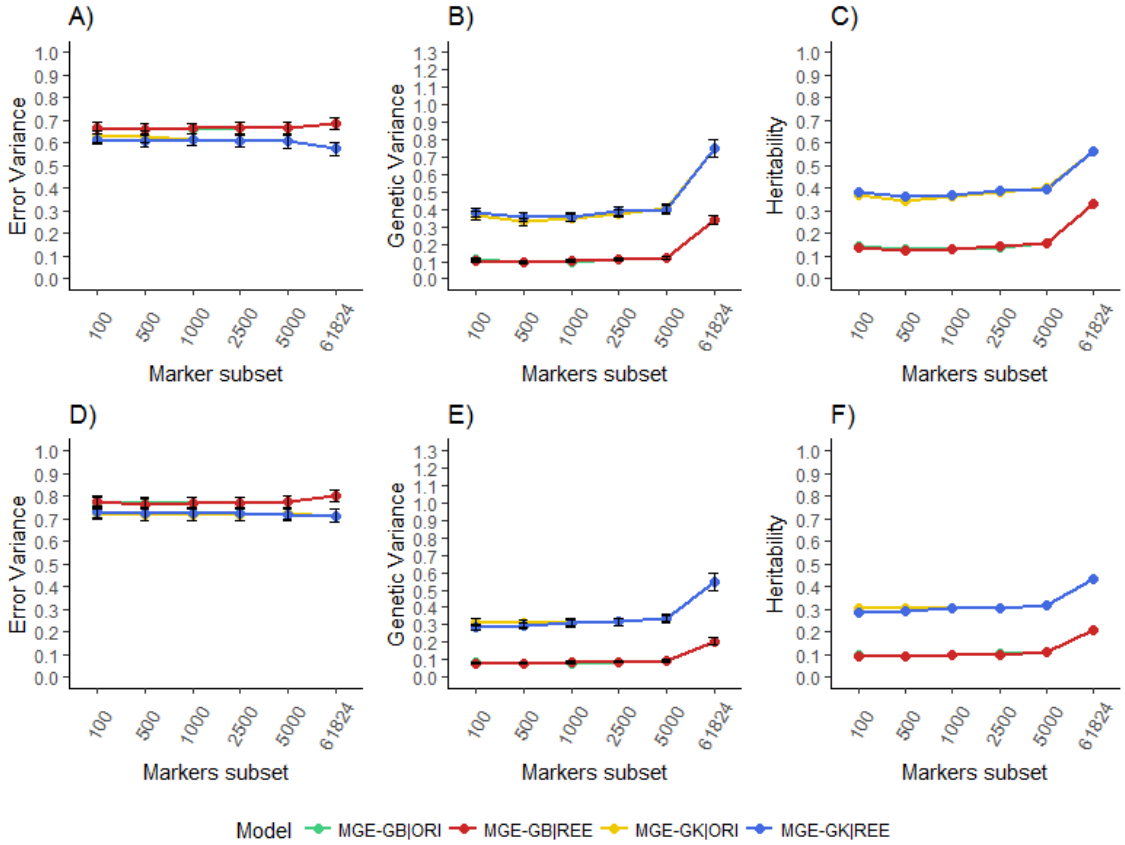


**Figure 3**. USP dataset. Variance components and genomic heritability through MGE-GB and MGE-GK models considering markers subsets (horizontal axis) obtained by original effect (ORI) and reestimated effect (REE) strategies, for plant height (A) Residual variance, (B) genetic variance, and (C) heritability; and for grain yield (D) Residual variance, (E) genetic variance, and (F) heritability.

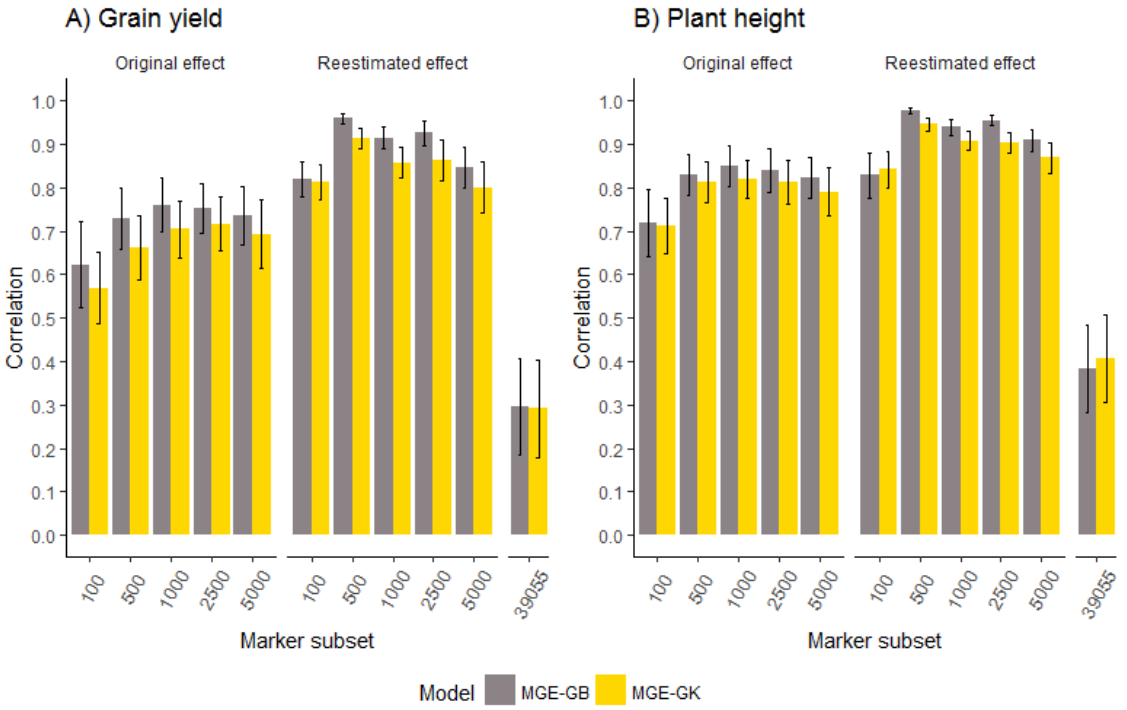


**Figure 4**. RICE dataset. Mean correlation between observed and predictive values (average from thirty random TRN-TST replications) through MGE-GB and MGE-GK models considering markers subsets (horizontal axis) obtained by original effect (ORI) and reestimated effect (REE) strategies for grain yield and plant height.

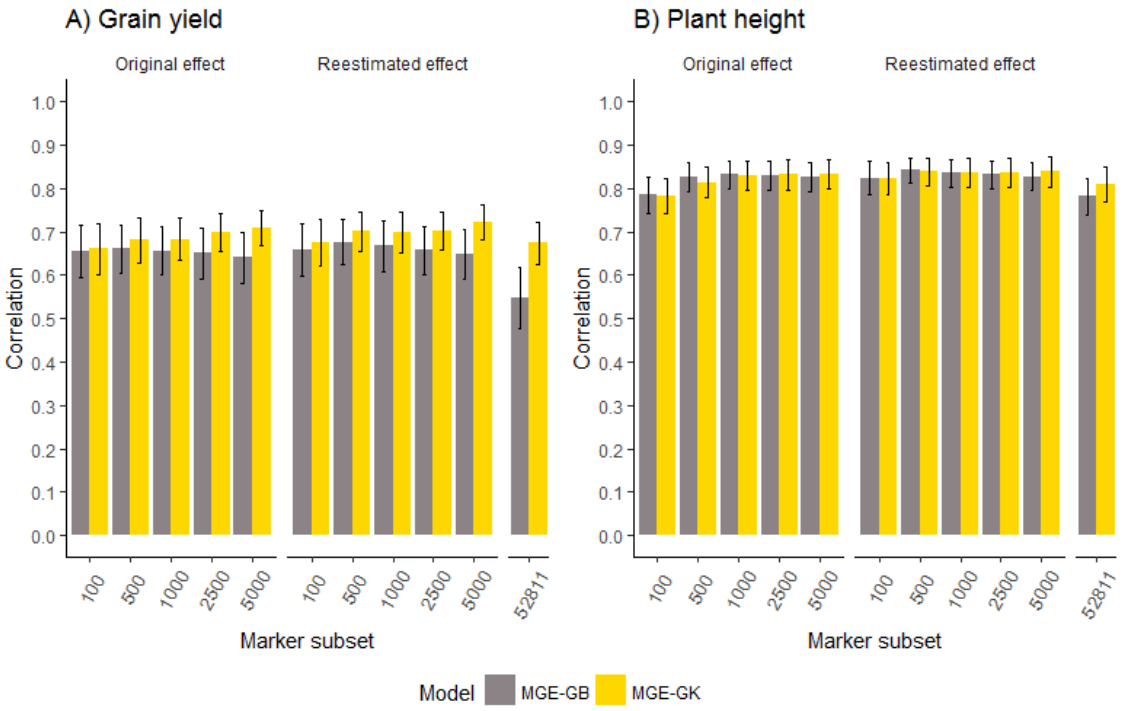


**Figure 5**. HEL dataset. Mean correlation between observed and predictive values (average from thirty random TRN-TST replications) through MGE-GB and MGE-GK models considering markers subsets (horizontal axis) obtained by original effect (ORI) and reestimated effect (REE) strategies for grain yield and plant height.

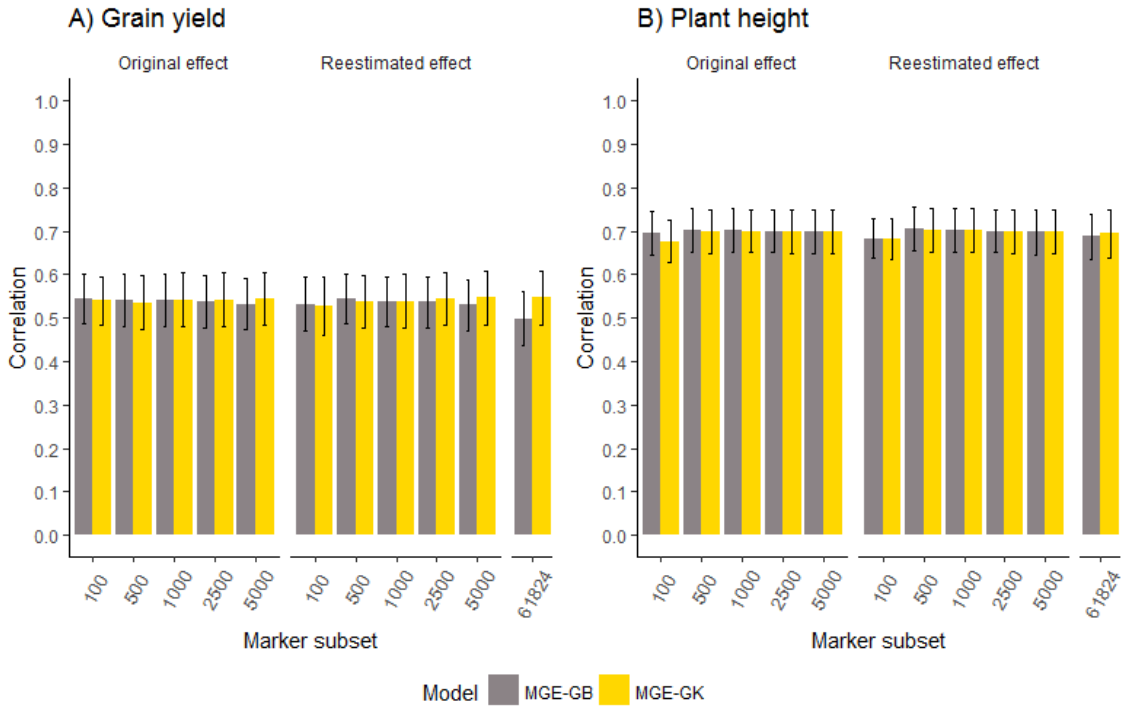


**Figure 6**. USP dataset. Mean correlation between observed and predictive values (average from thirty random TRN-TST replications) through MGE-GB and MGE-GK models considering markers subsets (horizontal axis) obtained by original effect (ORI) and reestimated effect (REE) strategies for grain yield and plant height.

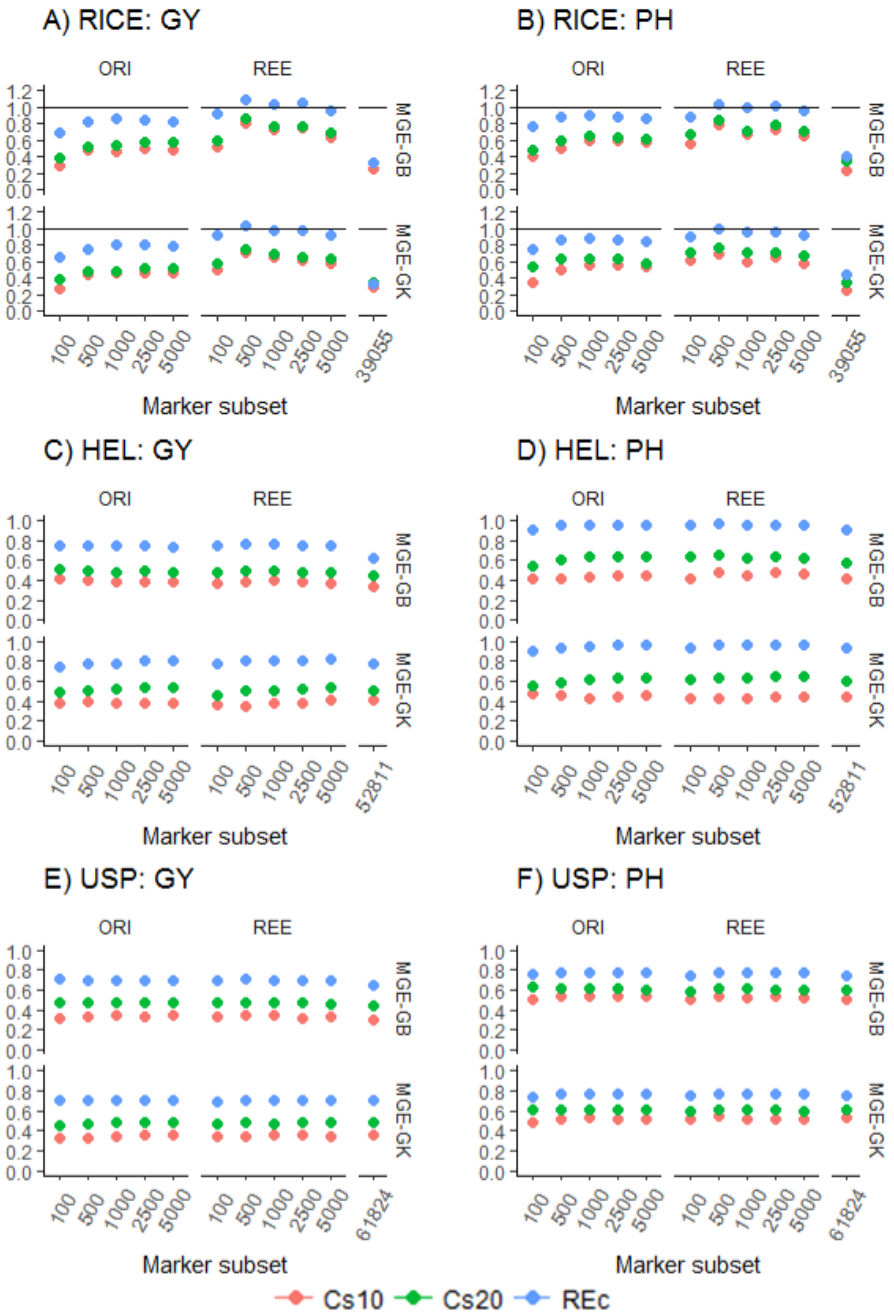


**Figure 7**. Coincidence of selection at 10% (Cs10) and 20% (Cs20) of intensity of selection, and relative efficiency ($RE_c$) for a cycle of genomic prediction through MGE-GB and MGE-GK models considering markers subsets obtained by original effect (ORI) and reestimated effect (REE) strategies for plant height and grain yield in RICE, HEL and USP datasets.

**Table 1.** Coincidence of selection at 10% (Cs10) and 20% (Cs20) of intensity of selection, and relative efficiency ($RE_c$) for a cycle of genomic prediction through MGE-GB and MGE-GK models considering markers subsets obtained by original effect (ORI) and reestimated effect (REE) strategies for plant height and grain yield in RICE, HEL and USP datasets

| | | | Grain yield | | | | | | | | | | | Plant height | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Model | Parameter | Original effect | | | | | | Reestimated effect | | | | | Original effect | | | | | | Reestimated effect | | | | |
| | | | All* | 5000 | 2500 | 1000 | 500 | 100 | 5000 | 2500 | 1000 | 500 | 100 | All* | 5000 | 2500 | 1000 | 500 | 100 | 5000 | 2500 | 1000 | 500 | 100 |
| RICE dataset | MGE-GB | Cs10 | 0.26 | 0.49 | 0.51 | 0.47 | 0.48 | 0.29 | 0.63 | 0.75 | 0.73 | 0.81 | 0.53 | 0.24 | 0.57 | 0.59 | 0.59 | 0.51 | 0.41 | 0.65 | 0.73 | 0.68 | 0.79 | 0.55 |
| | | Cs20 | 0.33 | 0.58 | 0.58 | 0.54 | 0.52 | 0.39 | 0.70 | 0.77 | 0.77 | 0.86 | 0.60 | 0.35 | 0.62 | 0.64 | 0.65 | 0.60 | 0.49 | 0.71 | 0.79 | 0.72 | 0.84 | 0.67 |
| | | $RE_c$ | 0.34 | 0.83 | 0.85 | 0.86 | 0.83 | 0.70 | 0.96 | 1.05 | 1.03 | 1.09 | 0.93 | 0.40 | 0.87 | 0.89 | 0.90 | 0.88 | 0.76 | 0.96 | 1.01 | 1.00 | 1.04 | 0.88 |
| | MGE-GK | Cs10 | 0.29 | 0.46 | 0.45 | 0.45 | 0.43 | 0.27 | 0.58 | 0.61 | 0.65 | 0.70 | 0.50 | 0.25 | 0.53 | 0.55 | 0.56 | 0.49 | 0.34 | 0.58 | 0.65 | 0.60 | 0.69 | 0.61 |
| | | Cs20 | 0.34 | 0.51 | 0.51 | 0.48 | 0.47 | 0.38 | 0.63 | 0.64 | 0.69 | 0.74 | 0.58 | 0.35 | 0.58 | 0.62 | 0.63 | 0.62 | 0.53 | 0.67 | 0.70 | 0.70 | 0.76 | 0.70 |
| | | $RE_c$ | 0.33 | 0.78 | 0.81 | 0.80 | 0.75 | 0.64 | 0.91 | 0.98 | 0.97 | 1.03 | 0.92 | 0.43 | 0.84 | 0.86 | 0.87 | 0.86 | 0.75 | 0.92 | 0.96 | 0.96 | 1.00 | 0.89 |
| HEL dataset | MGE-GB | Cs10 | 0.33 | 0.38 | 0.39 | 0.39 | 0.40 | 0.41 | 0.37 | 0.39 | 0.40 | 0.39 | 0.37 | 0.42 | 0.44 | 0.45 | 0.43 | 0.41 | 0.41 | 0.46 | 0.47 | 0.45 | 0.48 | 0.41 |
| | | Cs20 | 0.44 | 0.47 | 0.49 | 0.48 | 0.50 | 0.51 | 0.48 | 0.48 | 0.49 | 0.49 | 0.48 | 0.58 | 0.63 | 0.64 | 0.63 | 0.61 | 0.54 | 0.62 | 0.64 | 0.62 | 0.65 | 0.64 |
| | | $RE_c$ | 0.62 | 0.73 | 0.74 | 0.75 | 0.75 | 0.75 | 0.74 | 0.75 | 0.76 | 0.77 | 0.75 | 0.90 | 0.95 | 0.95 | 0.96 | 0.95 | 0.90 | 0.95 | 0.96 | 0.96 | 0.97 | 0.95 |
| | MGE-GK | Cs10 | 0.41 | 0.38 | 0.37 | 0.37 | 0.40 | 0.37 | 0.41 | 0.37 | 0.37 | 0.34 | 0.36 | 0.44 | 0.45 | 0.44 | 0.43 | 0.45 | 0.47 | 0.44 | 0.44 | 0.43 | 0.42 | 0.43 |
| | | Cs20 | 0.51 | 0.53 | 0.53 | 0.52 | 0.51 | 0.49 | 0.53 | 0.52 | 0.51 | 0.51 | 0.46 | 0.60 | 0.63 | 0.63 | 0.62 | 0.59 | 0.55 | 0.64 | 0.64 | 0.63 | 0.63 | 0.62 |
| | | $RE_c$ | 0.77 | 0.81 | 0.80 | 0.78 | 0.78 | 0.75 | 0.82 | 0.80 | 0.80 | 0.80 | 0.77 | 0.93 | 0.96 | 0.96 | 0.95 | 0.94 | 0.90 | 0.96 | 0.96 | 0.96 | 0.96 | 0.94 |
| USP dataset | MGE-GB | Cs10 | 0.30 | 0.34 | 0.33 | 0.34 | 0.33 | 0.32 | 0.33 | 0.32 | 0.34 | 0.34 | 0.33 | 0.51 | 0.53 | 0.54 | 0.54 | 0.53 | 0.50 | 0.52 | 0.54 | 0.52 | 0.54 | 0.50 |
| | | Cs20 | 0.44 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.46 | 0.48 | 0.47 | 0.48 | 0.48 | 0.60 | 0.60 | 0.61 | 0.62 | 0.62 | 0.63 | 0.60 | 0.60 | 0.61 | 0.61 | 0.59 |
| | | $RE_c$ | 0.65 | 0.69 | 0.70 | 0.70 | 0.70 | 0.71 | 0.69 | 0.70 | 0.70 | 0.71 | 0.69 | 0.75 | 0.77 | 0.77 | 0.77 | 0.77 | 0.76 | 0.77 | 0.77 | 0.77 | 0.77 | 0.75 |
| | MGE-GK | Cs10 | 0.36 | 0.36 | 0.35 | 0.34 | 0.33 | 0.33 | 0.34 | 0.35 | 0.35 | 0.34 | 0.34 | 0.53 | 0.52 | 0.52 | 0.53 | 0.52 | 0.49 | 0.51 | 0.51 | 0.52 | 0.55 | 0.52 |
| | | Cs20 | 0.48 | 0.48 | 0.48 | 0.48 | 0.47 | 0.46 | 0.48 | 0.48 | 0.47 | 0.48 | 0.47 | 0.61 | 0.61 | 0.61 | 0.61 | 0.61 | 0.61 | 0.60 | 0.61 | 0.61 | 0.61 | 0.60 |
| | | $RE_c$ | 0.71 | 0.71 | 0.71 | 0.71 | 0.70 | 0.70 | 0.71 | 0.71 | 0.70 | 0.70 | 0.69 | 0.76 | 0.77 | 0.77 | 0.77 | 0.77 | 0.74 | 0.77 | 0.77 | 0.77 | 0.77 | 0.75 |

*39,811, 52,811, 61,824 SNPs for RICE, HEL, and USP datasets, respectively

# APPENDICES

## Appendix A

**Table A1.** Wald test for fixed effects and Likelihood ratio test (LRT) for random effects estimated via REML/BLUP and narrow-sense heritability for plant height and grain yield for RICE, HEL, and USP dataset

| Dataset | Effect | Variation source | Plant height[R] | Grain yield[C] |
|---|---|---|---|---|
| RICE | Fixed [W] | Year | 391,14*** | 896,98*** |
| | | Repetition within Year | 17,60*** | 25,60*** |
| | Random [T] | Genotype | 182,06*** | 16,52*** |
| | | Genotype × Year | 96,76*** | 68,00*** |
| | | Heritability | 0,89 | 0,78 |
| HEL | Fixed | Site | 339.00*** | 2055.30*** |
| | | Block within Site | 3.00 [NS] | 2.20 [NS] |
| | Random | Genotype | 445.78*** | 309,18*** |
| | | Genotype × Site | 49.04*** | 348.61*** |
| | | Heritability | 0,87 | 0,77 |
| USP | Fixed | Site[T] | 66.00 ** | 297.20** |
| | | Block within Site | 0 [NS] | 0 [NS] |
| | Random | Genotype | 533.36** | 155.00** |
| | | Genotype × Site | 1.67 [NS] | 18.51** |
| | | Heritability | 0.83 | 0.59 |

[T] Random effects significant at ***$P = 0{,}001$, **$P = 0{,}01$ and [NS] – non-significant by LRT, evaluated by $\chi^2$ test with one degree of freedom; [W] Fixed effects significant at ***$P = 0{,}001$, **$P = 0{,}01$ and [NS] – non-significant by Wald statistic; [R] – evaluated in three years for RICE, in three sites for HEL and in two sites for USP; [C] – evaluated in three years for RICE, in five sites for HEL and in two sites for USP

**Table A2.** Bias and reliability (REL) (average of 30 random TRN-TST partitions), through MGE-GB and MGE-GK models considering markers, subsets obtained by original effect (ORI) and reestimated effect (REE) strategies with the standard deviation in parenthesis for plant height and grain yield in RICE, HEL, and USP datasets

| Model | Strategy | Subset | RICE dataset | | | | HEL dataset | | | | USP dataset | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Plant height | | Grain yield | | Plant height | | Grain yield | | Plant height | | Grain yield | |
| | | | Bias | REL | Bias | REL | Bias | REL | Bias | REL | Bias | REL | Bias | REL |
| MGE-GB | ORI | All | 0.91 (0.29) | 0.59 (0.03) | 0.81 (0.36) | 0.57 (0.03) | 1.05 (0.10) | 0.91 (0.01) | 1.04 (0.18) | 0.74 (0.02) | 0.99 (0.11) | 0.93 (0.01) | 1.034 (0.19) | 0.85 (0.01) |
| | | 5000 | 1.10 (0.17) | 0.78 (0.01) | 1.05 (0.16) | 0.75 (0.02) | 1.04 (0.08) | 0.83 (0.01) | 1.04 (0.16) | 0.67 (0.02) | 0.98 (0.10) | 0.82 (0.01) | 1.03 (0.17) | 0.70 (0.01) |
| | | 2500 | 1.07 (0.16) | 0.80 (0.01) | 1.02 (0.15) | 0.77 (0.01) | 1.04 (0.08) | 0.82 (0.01) | 1.04 (0.16) | 0.66 (0.02) | 0.98 (0.10) | 0.81 (0.01) | 1.03 (0.17) | 0.69 (0.01) |
| | | 1000 | 1.06 (0.14) | 0.83 (0.01) | 1.02 (0.16) | 0.81 (0.01) | 1.04 (0.09) | 0.83 (0.01) | 1.03 (0.16) | 0.68 (0.01) | 0.98 (0.10) | 0.81 (0.01) | 1.03 (0.17) | 0.69 (0.01) |
| | | 500 | 1.03 (0.14) | 0.84 (0.01) | 0.98 (0.15) | 0.82 (0.01) | 1.05 (0.09) | 0.87 (0.01) | 1.03 (0.16) | 0.70 (0.01) | 0.98 (0.10) | 0.82 (0.01) | 1.03 (0.17) | 0.71 (0.01) |
| | | 100 | 1.00 (0.25) | 0.84 (0.01) | 0.98 (0.21) | 0.83 (0.02) | 1.04 (0.10) | 0.92 (0.01) | 1.03 (0.17) | 0.81 (0.01) | 0.98 (0.10) | 0.95 (0.00) | 1.03 (0.17) | 0.80 (0.01) |
| | REE | 5000 | 1.21 (0.14) | 0.80 (0.01) | 1.19 (0.15) | 0.77 (0.01) | 1.04 (0.08) | 0.83 (0.01) | 1.05 (0.16) | 0.69 (0.01) | 0.98 (0.10) | 0.82 (0.01) | 1.03 (0.17) | 0.70 (0.02) |
| | | 2500 | 1.24 (0.09) | 0.83 (0.00) | 1.24 (0.10) | 0.82 (0.00) | 1.04 (0.08) | 0.82 (0.01) | 1.06 (0.16) | 0.68 (0.02) | 0.98 (0.10) | 0.81 (0.01) | 1.03 (0.17) | 0.69 (0.02) |
| | | 1000 | 1.18 (0.08) | 0.84 (0.00) | 1.25 (0.10) | 0.83 (0.00) | 1.04 (0.08) | 0.82 (0.01) | 1.04 (0.17) | 0.68 (0.01) | 0.98 (0.10) | 0.80 (0.01) | 1.02 (0.17) | 0.68 (0.01) |
| | | 500 | 1.14 (0.05) | 0.87 (0.00) | 1.17 (0.09) | 0.87 (0.00) | 1.04 (0.08) | 0.81 (0.01) | 1.05 (0.15) | 0.68 (0.01) | 0.98 (0.10) | 0.79 (0.01) | 1.03 (0.17) | 0.68 (0.01) |
| | | 100 | 0.99 (0.13) | 0.87 (0.01) | 0.96 (0.08) | 0.89 (0.01) | 1.04 (0.08) | 0.87 (0.01) | 1.04 (0.16) | 0.78 (0.01) | 0.98 (0.10) | 0.88 (0.01) | 1.04 (0.17) | 0.76 (0.01) |
| MGE-GK | ORI | All | 1.08 (0.31) | 0.72 (0.02) | 0.96 (0.46) | 0.69 (0.03) | 1.11 (0.09) | 0.90 (0.00) | 1.06 (0.14) | 0.86 (0.01) | 0.99 (0.12) | 0.87 (0.01) | 1.01 (0.15) | 0.85 (0.01) |
| | | 5000 | 1.29 (0.22) | 0.81 (0.01) | 1.24 (0.23) | 0.80 (0.01) | 1.07 (0.09) | 0.86 (0.01) | 1.04 (0.15) | 0.83 (0.01) | 0.97 (0.11) | 0.83 (0.01) | 0.99 (0.15) | 0.80 (0.01) |
| | | 2500 | 1.26 (0.21) | 0.82 (0.01) | 1.21 (0.19) | 0.81 (0.01) | 1.06 (0.09) | 0.85 (0.01) | 1.04 (0.15) | 0.81 (0.01) | 0.97 (0.11) | 0.82 (0.01) | 0.99 (0.18) | 0.79 (0.01) |
| | | 1000 | 1.22 (0.20) | 0.83 (0.00) | 1.16 (0.20) | 0.82 (0.01) | 1.07 (0.09) | 0.84 (0.01) | 1.04 (0.16) | 0.79 (0.01) | 0.98 (0.11) | 0.81 (0.01) | 0.99 (0.16) | 0.79 (0.01) |
| | | 500 | 1.19 (0.21) | 0.84 (0.00) | 1.08 (0.21) | 0.82 (0.01) | 1.08 (0.10) | 0.84 (0.01) | 1.04 (0.16) | 0.79 (0.01) | 0.98 (0.11) | 0.81 (0.01) | 0.99 (0.16) | 0.78 (0.01) |
| | | 100 | 1.08 (0.26) | 0.84 (0.01) | 1.02 (0.25) | 0.81 (0.01) | 1.08 (0.10) | 0.86 (0.01) | 1.04 (0.17) | 0.82 (0.01) | 0.98 (0.11) | 0.86 (0.01) | 0.99 (0.15) | 0.81 (0.01) |
| | REE | 5000 | 1.54 (0.24) | 0.83 (0.00) | 1.52 (0.24) | 0.82 (0.01) | 1.06 (0.08) | 0.86 (0.01) | 1.07 (0.14) | 0.84 (0.01) | 0.98 (0.11) | 0.82 (0.01) | 0.99 (0.15) | 0.80 (0.01) |
| | | 2500 | 1.58 (0.19) | 0.83 (0.00) | 1.59 (0.19) | 0.83 (0.00) | 1.06 (0.09) | 0.85 (0.01) | 1.04 (0.15) | 0.81 (0.01) | 0.97 (0.11) | 0.82 (0.01) | 0.99 (0.15) | 0.79 (0.01) |
| | | 1000 | 1.54 (0.18) | 0.84 (0.00) | 1.65 (0.18) | 0.83 (0.00) | 1.06 (0.08) | 0.84 (0.01) | 1.05 (0.15) | 0.80 (0.01) | 0.98 (0.11) | 0.82 (0.01) | 0.98 (0.16) | 0.78 (0.01) |
| | | 500 | 1.48 (0.15) | 0.84 (0.00) | 1.56 (0.15) | 0.84 (0.00) | 1.06 (0.08) | 0.83 (0.01) | 1.05 (0.15) | 0.80 (0.01) | 0.98 (0.11) | 0.81 (0.01) | 0.99 (0.16) | 0.77 (0.01) |
| | | 100 | 1.22 (0.15) | 0.85 (0.00) | 1.20 (0.10) | 0.86 (0.00) | 1.05 (0.09) | 0.85 (0.01) | 1.04 (0.16) | 0.81 (0.01) | 0.98 (0.11) | 0.83 (0.01) | 1.00 (0.18) | 0.79 (0.01) |

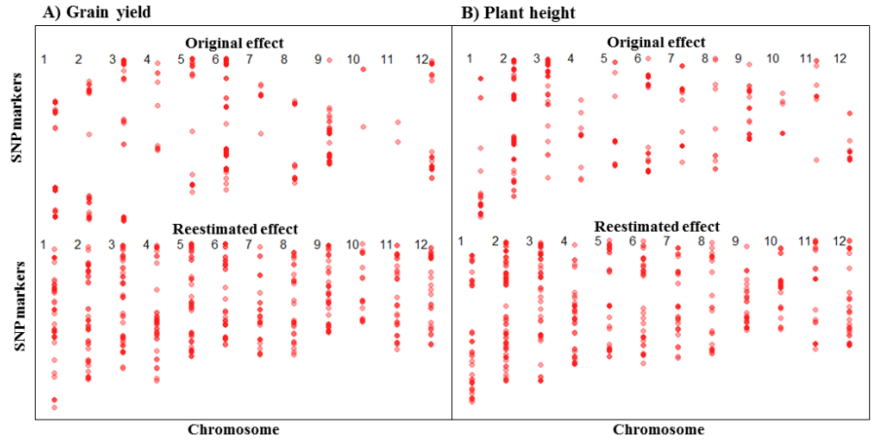


**Figure A1.** RICE dataset. Genome distribution by chromosome of 500 SNP markers obtained by the original effect (ORI) and reestimated effect (REE) methods for A) grain yield and B) plant height.

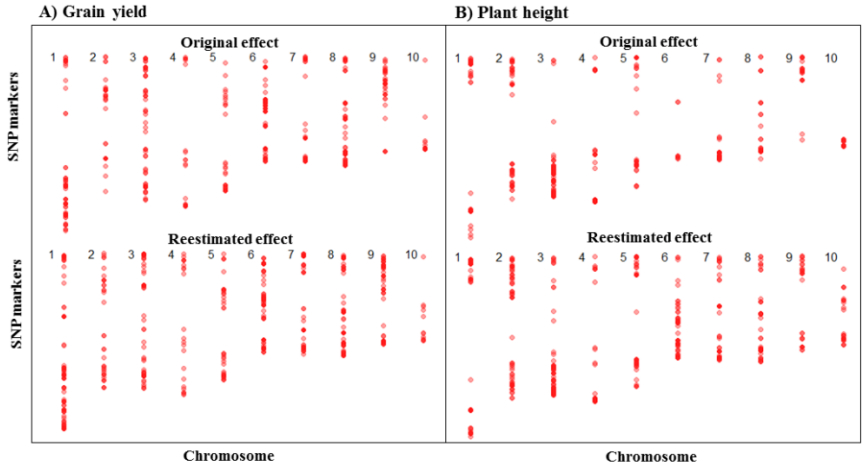


**Figure A2.** HEL dataset. Genome distribution by chromosome of 500 SNP markers obtained by the original effect (ORI) and reestimated effect (REE) methods for A) grain yield and B) plant height.

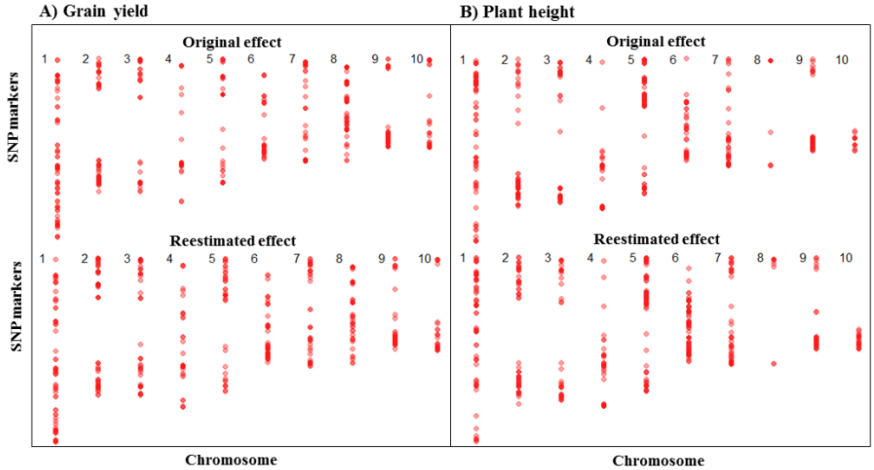


**Figure A3.** USP dataset. Genome distribution by chromosome of 500 SNP markers obtained by the original effect (ORI) and reestimated effect (REE) methods for A) grain yield and B) plant height.

## Appendix B

**Box B1**. The simplified scripts that can be used to obtain the markers by ORI and REE strategy through rrBLUP R package.

```
Y <- scale(Y,center=TRUE,scale=TRUE)
Zend = Z                # SNP genotype -1, 0 and 1
strategy <- "ORI"      # choose ORI or REE
markers <- c(52811, 10000, 5000, 2500, 1000, 500, 100)   # Choose number of markers by subset
library(rrBLUP)
if (strategy == "ORI"){
RR <- as.matrix((mixed.solve(Y,Zend))$u)
RRabs <- as.matrix(RR[order(abs(RR),decreasing = T),])
Zred <- list()
for(r in 1:length(markers)) {
  RRredabs <- head(RRabs,markers[r])
  Zred[[r]] <- Zend[,match(rownames(RRredabs),colnames(Zend))]
}}
if (strategy == "REE"){
  Zred <- list()
  g <- length(markers)
  for(r in 1:g) {
    RR <- as.matrix((mixed.solve(Y,Zend))$u)
    RRabs <- as.matrix(RR[order(abs(RR),decreasing = T),])
    RRredabs <- head(RRabs,markers[r])
    Zred[[r]] <- Zend[,match(rownames(RRredabs[[r]]),colnames(Zend))]
}}
```

The example uses the following R-objects: Z (n×p) is the genomic marker matrix; Y (n×1) is the numeric and standardized vector with phenotypic values.

**Box B2**. The simplified scripts that can be used to fit through MGE-GB or MGE-GK models

```
# Fitting MGE model
kernel <- "GB"     # Choose kernel GK or GB
partis <- 30   #number of replications
percTST <- 0.20
d <- length(Y)
nTST <- round(percTST*d)
COR <- matrix(NA,nrow=1,ncol=(partis))
COR_subset <- matrix(NA,nrow=length(markers),ncol=(partis))
for (n in 1:length(markers)) {
  if (kernel == "GK"){
  X = scale(Zred[[n]], center=TRUE, scale=TRUE)
  dist<-as.matrix(dist(X))^2
  K<-exp(-dist/median(dist))  }
if (kernel == "GB"){
    X <- Zred[[n]] + 1
    X <- scale(X, center=TRUE, scale=TRUE)
    K <- (X%*%t(X))/ncol(X)  }
 ETA <- list(list(K=K,model='RKHS'))
  set.seed(12345)
  for(i in 1:partis){
    nTST <- round(percTST*d)
    tst <-sample(1:d,size=nTST,replace=FALSE)
    YNA <- Y
    YNA[tst]<-NA
    yNA <- as.vector(YNA)
    fm <- BGLR(y=yNA,ETA=ETA,nIter=30000,burnIn=5000,thin=2)
    COR[i] <- cor(Y[tst],fm$yHat[tst]) }
COR_subset[n,] <- COR}
```

# 3. GENOMIC-ENABLED PREDICTION IN MAIZE USING KERNEL MODELS WITH GENOTYPE × ENVIRONMENT INTERACTION


## ABSTRACT

Multi-environment trials are routinely conducted in plant breeding to select candidates for the next selection cycle. In this study, we compare the prediction accuracy of four developed genomic-enabled prediction models: (1) single-environment, main genotypic effect model (SM), (2) multi-environment, main genotypic effects model (MM), (3) multi-environment, single variance G×E deviation model (MDs), and (4) multi-environment, environment-specific variance G×E deviation model (MDe). Each of these four models were fitted using two kernel methods: a linear kernel Genomic Best Linear Unbiased Predictor, GBLUP (GB), and a non-linear kernel Gaussian kernel (GK). The eight model-method combinations were applied to two extensive Brazilian maize data sets (HEL and USP data sets), having different numbers of maize hybrids evaluated in different environments for grain yield (GY), plant height (PH) and ear height (EH). Results show that the MDe and the MDs models fitted with the Gaussian kernel (MDe-GK, and MDs-GK) had the highest prediction accuracy. For GY in the HEL data set, the increase in prediction accuracy of SM-GK over SM-GB ranged from 9% to 32%. For the MM, MDs, and MDe models, the increase in prediction accuracy of GK over GB ranged from 9% to 49%. For GY in the USP data set, the increase in prediction accuracy of SM-GK over SM-GB ranged from 0% to 7%. For the MM, MDs, and MDe models, the increase in prediction accuracy of GK over GB ranged from 34% to 70%. For traits PH and EH, gains in prediction accuracy of models with GK compared to models with GB were smaller than those achieved in GY. Also these gains in prediction accuracy decreased when a more difficult prediction problem was studied.



Keywords: Genomic-enabled prediction; Maize multi-environment trials; Genotype × environment interaction (G×E); Genomic Best Linear Unbiased Predictor (GBLUP) linear kernel; Gaussian non-linear kernel; Genomic selection

Published on April 28, 2017 as doi:10.1534/g3.117.042341


## 3.1. INTRODUCTION

Genomic selection (GS) arose from the need to improve the prediction of complex traits based on dense marker information. It was first proposed in animal breeding (Meuwissen et al., 2001) and then applied to plant breeding (de los Campos et al., 2009; Crossa et al., 2010; 2011). Using GS, genomic breeding values are estimated as the sum of marker effects for unphenotyped individuals in the testing population. Several genomic prediction models were developed and applied to simulated and real plant breeding data (Bernardo and Yu, 2007;

Massman et al., 2013; de los Campos et al., 2009, 2013; Crossa et al., 2010, 2011; Pérez-Rodríguez et al., 2013; Beyene et al., 2015). In general, these studies showed good prediction accuracies for complex traits evaluated by means of random cross-validation partitions of the data.

Genomic-enabled prediction models were originally developed for a single trait in a single environment. However, multi-environment plant breeding trials are routinely conducted to estimate and take advantage of genotype × environment interaction (G×E). Therefore, to implement GS strategies in plant breeding, G×E needs to be estimated, modeled, and predicted. When genetic marker information is used for computing associations between individuals through the Genomic Relationships Matrix (GRM) **G** (Habier et al., 2007), this model is also referred to as Genomic Best Linear Unbiased Predictor (GBLUP) (VanRaden 2007, 2008). Burgueño et al. (2012) extended the GBLUP methodology to incorporate and model G×E effects. The Bayesian model of Jarquín et al. (2014) is another GBLUP extension that introduces main and interaction effects of markers and environmental co-variables via covariance structures. Heslot et al. (2014) proposed using crop modeling for assessing genomic G×E. In general, studies have shown that modeling G×E can give substantial gains in prediction accuracy (Burgueño et al., 2012; Heslot et al., 2014; Jarquín et al., 2014; Crossa et al., 2016a, 2016b; Cuevas et al., 2016a, 2016b).

Recently, López-Cruz et al. (2015) proposed a GBLUP prediction model that explicitly models G×E and marker × environment interaction (M×E) where marker effects and genomic values are partitioned into components that are stable across environments (main effects) and others that are environment-specific (interactions). The model of López-Cruz et al. (2015) has advantages and disadvantages; the advantages are: (i) it can be easily implemented using existing software for GS, for example, BGLR (de los Campos and Pérez-Rodríguez, 2016); (ii) it can be implemented using both shrinkage methods (a ridge-regression type estimator) and variable (marker) selection methods. In this case, the M×E model can be employed not only for GS but also for genome-wide association analyses to identify genomic regions that contribute to stability and to interaction effects (Crossa et al., 2016b). Furthermore, in terms of reducing the models' residual variance, the M×E model outperformed the more traditional single-environment and across-environment models for complex traits (López-Cruz et al., 2015; Crossa et al., 2016b). However, the M×E model is more efficient when applied to sets of environments that have positive correlations. This limitation arises because the genetic covariance between any pair of environments is represented by the variance of the main effect, which forces the covariance between pairs of environments to be positive (López-Cruz et al., 2015).

VanRaden (2008) first suggested models using a standard linear kernel, where the GBLUP is a linear model characterized by parameters related to additive quantitative genetics theory. However, complex traits are affected by non-linearity effects between genotypes and phenotypes due to complex interactions among genes (i.e., epistasis) and their interaction with the environment. Gianola et al. (2006) proposed nonparametric and semi-parametric methods to model the relationship between the phenotype and markers that are available within the GS framework. The nonparametric methods are capable of accounting for small complex epistatic interactions without explicitly modeling them.

Therefore, the semi-parametric Reproducing Kernel Hilbert Space (RKHS) reduces the dimensions of the parametric space and also captures small complex interactions among markers (Gianola and van Kaam, 2008; de los Campos et al., 2010; Morota and Gianola, 2014; Gianola et al., 2006, 2014). The RKHS method uses a kernel function to convert the marker matrix into a set of distances between pairs of individuals (Heslot et al., 2012). Recently, Jiang and Reif (2015) formulated a model with explicit epistatic effects of markers and proved that this model is equivalent to RKHS with Gaussian kernel, thus demonstrating that this model captures epistatic effects among markers. Several studies confirmed the advantage of using RKHS regression to increase prediction accuracy by capturing non-additive variation (de los Campos et al., 2010; Heslot et al., 2012; Morota et al., 2013; Pérez-Rodríguez et al., 2013; Morota and Gianola, 2014).

Recently, Cuevas et al. (2016a) compared methods that applied the G×E interaction GS model of López-Cruz et al. (2015) using a linear kernel (GBLUP) and a non-linear Gaussian kernel with the bandwidth estimated by an empirical Bayes method proposed by Pérez-Elizalde et al. (2015), and a Kernel Averaging method, or multi Gaussian kernels (de los Campos et al., 2010). Cuevas et al. (2016a) evaluated these methods using single-environment and multi-environment G×E interaction models to show the higher prediction ability of the Gaussian kernel models with the G×E model vs the linear kernel with the G×E model. The most flexible Gaussian kernels captured major and complex effects of markers in addition to their interaction effects.

In genomic-enabled prediction, linear models consider genetic values as linear combinations of marker effects; therefore, GBLUP with G×E can also be fitted with the non-linear Gaussian kernel. Similarly, the model of Jarquín et al. (2014) can also be used with the linear GBLUP kernel and with the non-linear Gaussian kernel (GK). However, so far, the model of Jarquín et al. (2014) has not been used with the Gaussian kernel (GK), and its prediction accuracy has not been compared with the G×E with the GBLUP kernel of López-Cruz et al.

(2015) or with the model of Jarquín et al. (2014) across environments or when including G×E with the GBLUP kernel.

Therefore, the main objectives of this research were: (1) to study the prediction accuracy of the multi-environment, single variance G×E deviation model (MDs) of Jarquín et al. (2014) used with the Gaussian kernel (GK) method (MDs-GK) and the prediction accuracy of the multi-environment, environment-specific variance G×E deviation model (MDe) of López-Cruz et al. (2015) fitted with the Gaussian kernel (GK) method (MDe-GK), and to compare them with the prediction accuracy of their counterpart models using the linear kernel GBLUP (GB) method (MDs-GB and MDe-GB), and (2) to compare the accuracy of the four previous models with the accuracy of the single-environment, main genotypic effect model (SM), and the multi-environment, main genotypic effect (MM) of Jarquín et al. (2014) using GB and GK methods (SM-GB, SM-GK, MM-GB, and MM-GB).

Two data sets of Brazilian maize hybrids (HEL and USP) genotyped with dense molecular markers and phenotyped in different environments with different numbers of hybrids per environment and different traits were used to compare the prediction accuracy of those eight model-methods.

## 3.2. MATERIALS AND METHODS

### 3.2.1. Phenotypic experimental data

This study considered two hybrid maize data sets. The first data set is from the Helix Seeds Company (HEL), while the second is from the University of São Paulo (USP). The HEL data set consists of 452 maize hybrids obtained by crossing 111 pure lines (inbreds); the hybrids were evaluated in 2015 at five Brazilian sites: Nova Mutum (NM), (13º 05’ S, 56º 05’ W, 460 meters above sea level) and Sorriso (SO) (12º 32' S, 55º 42' W, 365 meters above sea level) in the state of Mato Grosso; Pato de Minas (PM) (18º 34' S, 46º 31' W, 832 meters above sea level) and Ipiaçú (IP) (18º 41' S, 49º 56' W, 452 meters above sea level) in the state of Minas Gerais; and Sertanópolis (SE) (23º 03' S, 51º 02' W, 361 meters above sea level) in the state of Parana. The experimental design was a randomized block with two replicates per genotype and environment. The phenotypic and genomic data on inbred lines are credited to Helix Seeds Ltda Company.

Data from USP consist of 740 maize hybrids obtained by crossing 49 inbred lines. The hybrids were evaluated at Piracicaba and Anhumas, São Paulo, Brazil, in 2016. The hybrids were evaluated using an augmented block design, with two commercial hybrids as checks to correct for

micro-environmental variation. At each site, two levels of nitrogen (N) fertilization were used: Ideal N (IN) and Low N (LN) for a total of four artificial environments (P-IN, P-LN, A-IN, and A-IN). The experiment conducted under ideal N conditions received 100 kg $ha^{-1}$ of N (30 kg $ha^{-1}$ at sowing and 70 kg $ha^{-1}$ in a coverage application) at the V8 plant stage, while the experiment with low N received 30 kg/ha of N at sowing. Each plot was 7 m in length, with 0.50 m spacing between rows and 0.33 m between plants.

There was an imbalance in both data sets because not all hybrids were evaluated in all locations (incomplete field trials). The main traits of the two data sets were GY (grain yield in ton $ha^{-1}$), PH (plant height in cm) and EH (ear height in cm). Plant height was measured from the ground to the flag leaf and EH from the ground to the base of the ear. The empirical distribution of the three evaluated traits (GY, PH, and EH) were symmetric in most of the environments (Fig. 1). For the HEL data set, PH and EH were evaluated in three environments. Broad sense heritability (repeatability) were computed using the standard formula based on plot means

$$\sigma_h^2/[\sigma_h^2 + \sigma_{hs}^2/s + \sigma_e^2/sr$$

where $\sigma_h^2$ is the variance of the hybrids, $\sigma_{hs}^2$ is the variance of the hybrids × location interaction and $\sigma_e^2$ is the residual error variance and *s* and *r* are the number of environments and replicates in each environment, respectively.

### 3.2.2. Genotypic data

The USP and HEL parent lines were genotyped with an Affymetrix® Axiom® Maize Genotyping Array of 616 K SNPs (Unterseer et al., 2014). Standard quality controls (QC) were applied to the data, removing markers with a *Call Rate* $\geq$ 0.95. The remaining missing data in lines were imputed with the Synbreed package (Wimmer et al., 2015) using the algorithms from the software Beagle 4.0 (Browning and Browning, 2008). The hybrid genotypes were obtained by genomic information of the parent inbred lines. Markers with Minor Allele Frequency (MAF) of $\leq$ 0.05 were removed. After applying QC, 52,811 and 54,113 SNPs were available to make the predictions in the HEL and USP data sets, respectively.

### 3.2.3. Data availability

The phenotypic and genotypic data for the maize hybrids included in this study can be found at http://hdl.handle.net/11529/10887. Each HEL and USP data set contains the data corresponding to each GB and GK kernel, as well as phenotypic data for each trait (PH, EH, and GY) in each environment.

### 3.2.4. Statistical models

Statistical models for genomic predictions taking into account G×E were proposed by Jarquín et al. (2014) and López-Cruz et al. (2015). The Jarquín model incorporates genetic information from molecular markers or from pedigree (Pérez-Rodríguez et al., 2015), and/or from environmental covariates, whereas the López-Cruz model decomposes the marker effect across all environments and for each specific environment (interaction).

In this study, four statistical prediction models were fitted to both data sets to study their prediction accuracy using random cross-validation schemes (Table 1). We also compared the prediction accuracy of two kernel regression methods in the four models. The models were: a single-environment, main genotypic effect model (SM), a multi-environment, main genotypic effect model (MM) (Jarquín et al., 2014), a multi-environment, single variance G×E deviation model (MDs) (Jarquín et al., 2014) and a multiple-environment environment-specific variance G×E deviation model (MDe) (López-Cruz et al., 2015). The two kernel regression methods were: the linear kernel GBLUP (GB) method used by Jarquín et al. (2014) and López-Cruz et al. (2015), and the Gaussian kernel (GK) method proposed by Cuevas et al. (2016a).

#### 3.2.4.1. The single-environment, main genotypic effect model (SM)

The SM model fits the data from each environment separately and takes into account the main effect of genotypes. In matrix notation the model is written as:

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the response vector, and $y_i$ represents the observations in the $i^{\text{th}}$ line $(i = 1, \ldots, n)$ in each environment; $\mu$ is the general mean; $\boldsymbol{Z}_u$ is the incidence matrix that connects the random genetic effects with phenotypes; $\boldsymbol{u}$ is the random genetic effects for each environment, and $\boldsymbol{\varepsilon}$ the residual random effects for each environment. The SM model (1) assumes that the distribution of the $\boldsymbol{u}$ vector is multivariate normal with mean zero and a

covariance matrix $\sigma_{u_j}^2\boldsymbol{K}$, that is, $\boldsymbol{u}\sim N(\boldsymbol{0},\sigma_u^2\boldsymbol{K})$, where $\sigma_{u_j}^2$ is the genetic variance component of $\boldsymbol{u}$ in the $j^{\text{th}}$ environment and $\boldsymbol{K}$ is a symmetric, positive semi-definite matrix, denoting the variance-covariance of the genetic values constructed from the genomic molecular markers. It is also assumed that the errors $\boldsymbol{\varepsilon}$ in each environment are independent with homogeneous variance, $\sigma_\varepsilon^2$, and distributed as $\boldsymbol{\varepsilon}\sim\boldsymbol{N}(\boldsymbol{0},\sigma_{e_j}^2\boldsymbol{I})$ (where $\boldsymbol{I}$ is the identity matrix, and $\sigma_{e_j}^2$ is the residual variance in the $j^{\text{th}}$ environment). Therefore, $\boldsymbol{u}$ is an approximation of the true unknown genetic values, and $\boldsymbol{\varepsilon}$ captures the residual genetic effects that were not explained by $\boldsymbol{u}$ plus other non-genetic effects that approximate the errors. Note that matrix $\boldsymbol{K}$ may have different dimensions depending on the number of lines evaluated in each environment.

#### 3.2.4.2. The single-environment, main genotypic effect model with GBLUP (SM-GB) and Gaussian Kernel (SM-GK)

For SM model (1), matrix $\boldsymbol{K}$ can be constructed using the linear kernel $\boldsymbol{K}=\frac{\boldsymbol{XX'}}{p}$ (de los Campos et al., 2013) proposed by VanRaden (2007, 2008) for estimating the GBLUP, where $\boldsymbol{X}$ is the standardized matrix of molecular markers for the individuals, of order $n\times p$, where $p$ is the number of markers (Table 1).

When markers have a more complex function, the Gaussian kernel (GK) has proved to be more effective for capturing complex cryptic interactions between markers, thereby improving the prediction accuracy of the model (Cuevas et al., 2016a). The entries of the Gaussian kernel are computed as $\boldsymbol{K}(\boldsymbol{x}_i\boldsymbol{x}_{i`})=\exp(-hd_{ii`}^2)$, where $d_{ii`}$ is the Euclidean distance between the individuals $i$th and $i'^{\text{th}}$ $(i=1,\ldots,n_j)$ given by the markers; $h>0$ is the bandwidth parameter that controls the rate of decay of $\boldsymbol{K}$ values (Pérez-Rodríguez et al., 2012; Cuevas et al., 2016a). In this study, GK takes the form $\boldsymbol{K}(\boldsymbol{x}_i\boldsymbol{x}_{i`})=\exp(-hd_{ii`}^2/\text{median}(d_{ii`}^2))$, where $h=1$ and the median of the distances is used as a scaling factor (Crossa et al., 2010). de los Campos et al. (2010) described the theory of the Gaussian kernel in the context of the kernel averaging method for the RKHS.

#### 3.2.4.3. The multi-environment, main genotypic effect model (MM)

One multi-environment model considers the main fixed effects of environments, as well as the random genetic effects across environments

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_E\boldsymbol{\beta}_E + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{\varepsilon} \quad (2)$$

where $\boldsymbol{y} = \left(\boldsymbol{y}_{n_i1}, \boldsymbol{y}_{n_i2}, \ldots, \boldsymbol{y}_{n_is}\right)'$ is the response vector and $\boldsymbol{y}_{n_is}$ is the vector of observations of the lines $(i = 1, \ldots, n_i)$ in the $j^{\text{th}}$ environment $(j = 1, \ldots, s)$. The fixed effects of the environment for the data used in this study are modeled with the incidence matrix of the environments $\boldsymbol{Z}_E$, where the parameters to be estimated are the intercept for each environment ($\boldsymbol{\beta}_E$) (other fixed effects can be incorporated into the model) and incidence matrix $\boldsymbol{Z_u}$ connects genotypes with phenotypes for each environment. It is assumed that the random vector of genetic effects $\boldsymbol{u}$ across environments follows a multivariate normal with mean zero and a covariance matrix $\sigma^2_{u_0}\boldsymbol{K}$, that is, $\boldsymbol{u}\sim N(\mathbf{0}, \sigma^2_{u_0}\boldsymbol{K})$, where $\sigma^2_{u_0}$ is the variance of the main genetic effects across environments and $\boldsymbol{\varepsilon}\sim\boldsymbol{N}(\mathbf{0}, \sigma^2_e\boldsymbol{I}$. In this case, the resulting model is equivalent to the across-environment GBLUP model of López-Cruz et al. (2015) and Jarquín et al. (2014). When in the MM model of (2) $\boldsymbol{u}$ is used with $\boldsymbol{K} = \frac{\boldsymbol{XX'}}{p}$ (GBLUP), the model is the GBLUP across environments (MM-GB), whereas if $\boldsymbol{u}$ is used with the Gaussian kernel, the model is the GK across environments (MM-GK) (Table 1).

#### 3.2.4.4. Multi-environment, single variance genotype × environment deviation model (MDs)

This model extends (2) by adding the random interaction effect of the environments with the genetic information of the lines $(ue_{ij})$

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_E\boldsymbol{\beta}_E + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{ue} + \boldsymbol{\varepsilon} \quad (3)$$

It is assumed that the vector of random effects of the interaction $\boldsymbol{ue}$ has a multivariate normal distribution, $\boldsymbol{ue}\sim\boldsymbol{N}(\mathbf{0}, [\boldsymbol{Z}_u\boldsymbol{K}\boldsymbol{Z}'_u] \circ [\boldsymbol{Z}_E\boldsymbol{Z}'_E]\sigma^2_{ue})$, where $(\circ)$ is the Haddamar product operator and denotes the element-to-element product between two matrices in the same order (Jarquín et al., 2014), $\sigma^2_{ue}$ is the variance component of the interaction, the $\boldsymbol{K}$ matrix is defined as before and $\boldsymbol{u}$ is the vector of the main genetic effects that follows a multivariate normal with mean zero and a covariance matrix $\sigma^2_{u_0}\boldsymbol{K}$, that is, $\boldsymbol{u}\sim N\left(\mathbf{0}, \sigma^2_{u_0}\boldsymbol{K}\right)$, where $\sigma^2_{u_0}$ is the variance of the main genetic effects and $\boldsymbol{\varepsilon}\sim\boldsymbol{N}(\mathbf{0}, \sigma^2_e\boldsymbol{I})$.

Therefore, under these conditions

$$[\boldsymbol{Z}_u\boldsymbol{K}\boldsymbol{Z}_u'] \circ [\boldsymbol{Z}_E\boldsymbol{Z}_E'] = \begin{bmatrix} \boldsymbol{K}_1 & \cdots & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{K}_j & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \cdots & \boldsymbol{K}_m \end{bmatrix}$$

where $\boldsymbol{K}_j$ represents the kernel constructed from the molecular markers of the lines in the $j^{th}$ environment. Just as in model (2), matrix $\boldsymbol{K}$ is used in the variance-covariance for $\boldsymbol{u}$ of model (3) and is also a component of the variance-covariance of $\boldsymbol{ue}$. Then, kernel matrix $\boldsymbol{K}$ can be constructed using the linear kernel (MDs-GB) or the Gaussian Kernel (GK) (MDs-GK) (Table 1).

#### 3.2.4.5. Multi-environment, environment-specific variance genotype × environment deviation model (MDe)

The model of López-Cruz et al. (2015) separates the genetic effects of markers into the main marker effects across all environments and the specific marker effects in each environment. Thus, this model in matrix notation is

$$\boldsymbol{y} = \mu\boldsymbol{1} + \boldsymbol{Z}_E\boldsymbol{\beta}_E + \boldsymbol{Z}_u\boldsymbol{u}_0 + \boldsymbol{u}_E + \boldsymbol{\varepsilon} \quad (4)$$

For the MDe model (4) of the vector, $\boldsymbol{u}$ represents the main effect of markers (across all environments) with a variance-covariance structure similar to those used in models (2) and (3), that is, $\boldsymbol{u}_0 \sim N(0, \sigma^2_{u_0}\boldsymbol{K})$. However, as pointed out by López-Cruz et al. (2015) and Cuevas et al. (2016a), $\sigma^2_{u0}$ is common to all s environments and the borrowing of information among environments is generated through kernel matrix $\boldsymbol{K}$. On the other hand, $\boldsymbol{u}_E$ represents the specific effect of the markers in environments (or the effects of the interaction) with a variance-covariance structure different from model (3), that is, $\boldsymbol{u}_E \sim N(\boldsymbol{0}, \boldsymbol{K}_E)$, where $\boldsymbol{K}_E$ is:

$$\boldsymbol{K}_E = \begin{bmatrix} \sigma^2_{u_{E_1}}\boldsymbol{K}_1 & \cdots & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \sigma^2_{u_{E_j}}\boldsymbol{K}_j & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \cdots & \sigma^2_{u_{E_m}}\boldsymbol{K}_m \end{bmatrix} =$$

$$\begin{bmatrix} \sigma^2_{u_{E_1}}\boldsymbol{K}_1 & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} + \cdots + \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sigma^2_{u_{E_j}}\boldsymbol{K}_j & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} + \cdots$$

$$+ \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \sigma^2_{u_{E_s}}\boldsymbol{K}_s \end{bmatrix}$$

Note that matrix $\boldsymbol{K}_E$ can be expressed as the sum of $s$ matrices and the effects given by $\boldsymbol{u}_{Ej}$ are specific for the $j$th environment; thus, it has a normal distribution with mean and variance equal to $\mathbf{0}$, except for the $j$th environment, which has a variance-covariance matrix $\sigma^2_{\boldsymbol{uE}_j}\boldsymbol{K}_j$. The importance of these two terms ($\boldsymbol{u}$ and $\boldsymbol{u}_E$) of MDe model (4) is given by the corresponding variance components that are estimated from the data. Kernel matrix $\boldsymbol{K}$ is used in the components of $\boldsymbol{u}$, while kernel matrix $\boldsymbol{K}_E$ is used in the component of $\boldsymbol{u}_E$; both, $\boldsymbol{K}$ and $\boldsymbol{K}_E$ can be used with linear kernel (MDe-GB) or with the Gaussian kernel (MDe-GK) (Table 1).

### 3.2.5. Estimating variance components using full data analyses

The four models —SM, MM, MDs, and MDe fitted with GBLUP and GK methods— were used on the entire HEL and USP data sets for all the traits. To fit the models, the phenotypic data were centered and standardized (i.e., each phenotypic data point was centered by subtracting the overall mean of all environments and standardized by dividing by the sample standard deviation across all environments). These analyses were performed to derive estimates of variance components. The posterior variance components resulting from the residual effects, genetic main effect, and genetic environment-specific effects of the four models described above for three traits and for environments in HEL and USP data sets were computed and reported. Since the data were standardized, the summation of the variance components approximates 1.

### 3.2.6. Assessing prediction accuracy by random cross-validation

The prediction accuracy of the single-environment model-methods was assessed using 50 random partitions organized in 5 folds with 10 random partitions each, with 80% of the hybrids comprising the training (TRN) set and the remaining 20% of the individuals comprising

the testing (TST) set. This was performed separately in each environment as the single-environment models were fitted separately for each environment. For this validation procedure, all the parameters of the models (including variance components resulting from residual effects, and genetic effects) were re-estimated from TRN data in each of the TRN-TST partitions.

For the multi-environment models, the prediction accuracy of the model-methods for predicting the pattern of missing values in each environment was generated using two different cross-validation (CV) designs (Burgueño et al., 2012). The random cross-validation 1 design (CV1) mimics the prediction problem faced by breeders when newly developed lines have not been evaluated in any environment; in this case, 20% of the lines were not observed (unphenotyped) in all the environments and had to be predicted. The random cross-validation 2 design (CV2) mimics a prediction problem when lines are tested in incomplete field trials (sparse testing), where some lines are evaluated in some environments but not included in other environments. For these approaches, we assigned 80% of the observations to TRN and 20% to TST. None of the lines to be predicted in the TST are in the TRN set. For this validation procedure, all the parameters of the models (including variance components resulting from residual effects, genetic main effects, genetic × environment interactions effects, and environment-specific effects) were re-estimated from TRN data in each of the 50 random TRN-TST partitions. Note that CV2 design can only be applied to multi-environment model-methods (MM, MDs and MDE) but not to single-environment model-methods (SM). Also for SM the random cross-validation is a CV1 but applied to only one individual environment (site).

For each TRN-TST partition, models were fitted to the TRN data set and prediction accuracy was assessed by computing Pearson's product–moment correlation between predictions and phenotypes in the TST data set within environments. The same TRN-TST partitions were used to assess the prediction accuracy of each of the models. Thus, 50 correlations were computed for each model and trait; the mean and standard deviations of these 50 correlations are reported.

Therefore, adjusting models for full data, making inference on the parameters, and assessing predictions in TRN-TST partitions were based on 30,000 samples collected from the posterior and predictive distributions after discarding 5000 samples as burn-in.

### 3.2.7. Software

The aforementioned models can be implemented using the R (R Core Team, 2016) package Bayesian generalized linear regression (BGLR) (de los Campos and Pérez-Rodríguez,

2016; Pérez-Rodríguez and de los Campos, 2014 ). Appendix C gives the R codes used in the models included in this study when the numbers of individuals in different environments vary. It should be pointed out that in previous studies using the Jarquín et al. (2014) and López-Cruz et al. (2015) models, the number of entries in each environment was the same; thus, the model fitting and the random cross-validation partition scheme were simpler than in this case, where there are different numbers of entries in different environments.

## 3.3. RESULTS

### 3.3.1. Descriptive Statistics

Box-plots of GY, PH and EH in each environment for the two data sets are depicted in Fig. 1. Most of the distributions of the traits in environments had a symmetric distribution. For the HEL data set, environments NM and PM produced better GY, while environment SE had the lowest GY (Fig. 1A). Traits PH and EH showed similar trends but they were measured in only three environments: IP, PM, and SE. Data from USP had environments with soil nutritional stress, and the results showed that environments with ideal nitrogen conditions (P-IN and A-IN) had better potential for all traits (Fig. 1A).

All pair-wise environments for all traits in the HEL and USP data sets had high and positive correlations (Tables 2 and 3). As expected, for PH and EH, the correlations were higher than for the complex trait GY. In all environments, GY, PH and EH showed high heritability in the HEL data set: GY ranged from 0.60 to 0.86, PH varied from 0.72 to 0.91, and EH ranged from 0.69 to 0.89. For the USP data set, heritability values were high for PH and EH in all environments (from 0.60 to 0.85 and 0.66 to 0.91, respectively). For the USP data set, trait GY had an intermediate value of 0.42 for sites Anhumas and Piracicaba with ideal nitrogen (IN) and lower values for Piracicaba and Anhumas under low nitrogen conditions (LN) (0.22 and 0.19, respectively).

### 3.3.2. Estimating variance components

The distribution of the residuals after fitting the eight model-method combinations to the two data sets was approximately normal (data not shown). Phenotypic (and marker) data were standardized; thus, summation of variance components approximated 1 for GB models.

Deviations may be due to the high degree of imbalance in the number of hybrids in different environments, especially for data set HEL.

#### 3.3.2.1. HEL data set

*Single-environment*

For the SM model for the three traits in each environment, the estimated residual variance components for the GK method were smaller than those for the GB method (Table 4). In contrast, the variance component of genetic effects for each environment increased for the SM-GK model-method as compared to the genetic effects of model-method SM-GB.

*Multi-environment*

For the MM model, the residual variance components for MM-GK were smaller than the residuals for MM-GB for all traits, whereas for the genetic effects, the opposite occurred: the variance component of MM-GK was higher than that of MM-GB. For the MDs models, the residual variance components for MDs-GK were always smaller than the residual variances for MDs-GB, whereas the opposite occurred for the genetic main effect and genetic interaction effect (Table 4).

The results of the MM and MDs models indicated that the inclusion of the interaction term (G×E) induced a larger reduction in the estimated residual variance for GY; for traits PH and EH, these reductions in residual variance were smaller than those found for GY from the MM and MDs models. For the MDe model, the residual variance components of MDe-GK were smaller than those of the MDe-GB for all traits, whereas the variance components for the genetic main effect and genetic environment specific effect were higher for the GK models than for the GB model. The genetic main effect values are related to the correlations between environments, which are positive and range from medium to high. For the variance component associated with the genetic interaction effect, the values are lower because they decreased G×E in environments with a positive correlation. In general, the fit of MDs is slightly worse than the fit of MDe models because the genetic interaction effect component has a small effect on all traits, but mainly on EH and PH (Table 4).

#### 3.3.2.2. USP data set

*Single-environment*

For all traits, the variance components of genetic effects were higher when using the GK than the GB method in all environments and traits for the single-environment, main genotypic effect model (SM) (Table 5). The residual variance for SM-GK was always smaller than the residual for SM-GB for all traits.

*Multi-environment*

The residual variance components of the MM and MDs models for all traits were very similar, as were most of the variance components for the genetic main effects. For all the environments and traits, the residuals from the GK were smaller than the residuals from the GB for the MM, MDs, and MDe models. The variance components of the genetic main effects of MDs-GB and MDe-GB were consistently smaller than the genetic main effect of model-methods MDs-GK and MDe-GK. Also, the variance components of the genetic environment specific effects of the GK models were higher than those of the GB models for all traits (Table 5).

### 3.3.3. Prediction accuracy of the models with GBLUP and GK methods

Results of the prediction accuracy of the eight model-method combinations for the random cross-validation CV2 of the two data sets are given in Figs. 2-5 and Tables 6 and 7. Results of the prediction accuracy of the eight model-method combinations for random cross-validation CV1 of the two data sets are given in Tables D1 and D2 in Appendix D. Note that for the single-environment models, prediction accuracy was assessed by 50 random partitions of the data into 80% TRN and 20% TST.

#### 3.3.3.1. HEL data set

*Single-environment*

Results for CV2 for GY showed that the prediction accuracy of the SM-GK model-method was higher than for SM-GB in all environments (Fig. 2). For GY in environment IP, prediction accuracies were 0.64 for SM-GB and 0.73 for SM-GK, while in environment PM, prediction accuracies were 0.68 for SM-GB and 0.74 for SM-GK (Table 6). The percent change in accuracy for single-environment GB vs. single-environment GK for GY ranged from 9 to 32%

for the PM and NM environments, respectively. For trait PH, prediction accuracies were high: >0.65 for SM-GB and >0.69 for SM-GK. For trait EH, prediction accuracies were >0.64 for SM-GB and >0.67 for SM-GK; the increase in prediction accuracy ranged from 3 to 6% for traits EH and PH (Table 6).

*Multi-environment*

For GY, the best model-methods for CV2 were MDs-GK and MDe-GK for the five environments, followed by MDs-GB in the IP environment, MM-GK in the NM, PM and SO environments and by MDs-GB and MDe-GB in the SE environment (Fig. 3). For trait GY, the percent change in accuracy for MM-GB vs MM-GK was high, ranging from 9% in PM to 48% in SE environment; for MDs-GB vs MDs-GK, the % change ranged from 13% in PM to 47% in the SO environment; and for MDe-GB vs MDe-GK, the % change ranged from 13% in PM to 49% in SE environment (Table 6). These results show that for trait GY, the increase in prediction accuracy was high using the GK method. For less complex traits PH and EH, the percent change of GK over GB for all models showed a small increase in prediction accuracy, with the percent change ranging from 1% to 6%; there was no difference for trait EH in the IP environment.

For GY, MDs-GK and MDe-GK were, for CV2, the models that had the best prediction accuracy in all environments, with values ranging from 0.58 in SE to 0.81 in the IP environment. These values were higher for a complex trait like GY (Fig. 3). For the PH trait, MDs-GK and MDe-GK were the best models, with a prediction accuracy of 0.81-0.82 for all environments; for trait EH, these values ranged from 0.77 in IP to 0.81 in PM. In general, MDs models had similar results across environments.

Prediction accuracy for random cross-validation CV1 decreased (Table D1, Appendix D) as compared with those computed for CV2 for all traits and models. For example, for environment IP, MM-GB in CV2 gave a correlation of 0.56, whereas for CV1, prediction accuracy was 0.48, that is, an 8% decrease in accuracy, whereas for environment NM, the decrease is about 10%. Similar decreases in accuracy from CV1 to CV2 were found for other traits and model-method combinations. However, for GY, models with GK showed a greater increase in prediction accuracy over models with GB under CV1 (Table D1, Appendix D) than the increases achieved by those models under CV2 (Table 6). The % change in prediction accuracy of CV1 vs CV2 for models with GK vs GB did not change for traits PH and EH.

*Summary of results*

Prediction accuracy of trait GY in the 5 testing environments increased from 9% to 48% for models with method GK compared to models with method GB. For GY, increases in prediction accuracy of environments from multi-environment models MM, MDs, and MDe with GK vs MM, MDs, and MDe with GB tended to be higher than those achieved by the single-environment model (SM-GK vs SM-GB). For traits PH and EH in 3 environments, the increase in prediction accuracy of method GK over method GB ranged only from 0% to 6% and increases in prediction accuracy from multi-environment models MM, MDs, and MDe with GK vs MM, MDs, and MDe with GB tended to be lower than those achieved by the single-environment model (SM-GK vs SM-GB).

Although prediction accuracy for all model-method combinations decreased in CV1 vs CV2, the % change in the prediction accuracy of models with GK vs models GB did not change and even increased for GY.

#### 3.3.3.2. USP data set

*Single-environment*

Table 7 shows the correlations between observed and predicted values for all models in the USP data set. In general, for traits GY, PH and EH, all prediction accuracies of the single-environment models were similar using the GB or GK methods (Fig. 4). In these cases, the percentage change in accuracy for SM-GB vs SM-GK was 7% (GY) in environment P-LN and 5% (PH) in environment A-LN. For GY, the prediction accuracy in the A-IN environment (0.40 for SM-GB and SM-GK) was greater than in other environments. Gains in prediction accuracy of SM-GK over SM-GB for trait GY in this data set are much lower than those achieved in the HEL data set.

*Multi-environment*

Between the GB and GK methods, there were large increases in prediction accuracy using the GK method, especially for traits GY and PH (Fig. 5). For trait GY, the MM model gave a high percentage change in accuracy for GB vs GK ranging from 34 to 59%. In the MDs model, the percentage change ranged from 35 to 65%, and for the MDe model, the percentage change ranged from 37 to 70% (Table 7).

For PH, the increase using the GK method was similar to those found for the GY trait. The percentage change for the MM model with GB vs GK ranged from 38% (A-IN and P-IN)

to 57% (A-LN). For the MDs models, the percentage change ranged from 37% (A-IN) to 61% (A-LN), and for the MDe models, the percentage change ranged from 37% (A-IN) to 51% (A-LN). For trait EH, the increase was smaller, ranging from 15 to 23% for the GBLUP vs the GK method. The MM and MDs models showed similar correlations when compared using the same kernel method. Models MDs-GK and MDe-GK had a clear and sustainable increase in prediction accuracy over their counterparts using the GB kernel, MDs-GB and MDe-GB.

The prediction accuracy of random cross-validation CV1 decreased (Table D2, Appendix D) compared to the accuracy obtained with CV2 for all traits and models. In general, compared with other models, MDs and MDe decreased the accuracy of CV1 vs CV2 more; this was more pronounced for GY in some environments than in other environments. Also, the decrease in the accuracy of CV1 vs CV2 was higher for GY than for PH and EH.

*Summary of results*

The prediction accuracy of trait GY in 4 environments increased from 34% to 70% for multi-environment models MM, MDs, and MDe with GK vs MM, MDs, and MDe with GB. These gains in accuracy were much higher than those achieved by single-environment SM-GK over SM-GB (from 0% to 7%). Similar patterns were found for traits PH and EH, where the increase in prediction accuracy of method GK over method GB ranged from -2% to 5% for the single-environment model (SM) and drastically increased for multi-environment models MM, MDs, and MDe with GK vs MM, MDs, and MDe with GB, ranging from 15% to 57%.

Prediction accuracy for all model-method combinations decreased in CV1 vs CV2 and the % change in accuracy of models with GK vs models with GB decreased for all three traits. The increase in accuracy of models with GK vs models with GB did not occur under the CV1 design.

## 3.4. DISCUSSION

Several studies have documented the benefits of using a non-linear Gaussian kernel in multi-environment models for capturing small complex interactions among markers and increasing prediction accuracy (Gianola et al., 2006, 2014; Crossa et al., 2010; Pérez-Rodríguez et al., 2013; Pérez-Elizalde et al., 2015; Cuevas et al., 2016a, 2016b). In this research, genome-enabled prediction accuracy was studied in two different hybrid maize data sets using multi-environment models with the GBLUP and Gaussian kernel methods. In this context, we proposed including the Gaussian kernel in the model of Jarquin et al. (2014), which accounts for

environment information without and with interaction terms. Prediction accuracy was obtained using traits with different genetic architecture and heritability values that ranged from low to high.

### 3.4.1. Prediction accuracy differences in datasets, methods, cross-validation designs and G×E

We compared the single-environment model with two G×E models. The first G×E model can accommodate environmental covariables (Jarquín et al., 2014) that were not available in this study; the second G×E model (López-Cruz et al., 2015) accommodates environment main effects and environment-specific variances; both models used the GB and Gaussian kernel methods proposed by Cuevas et al. (2016a). These two G×E models have demonstrated good prediction accuracy when used in genomic-enabled prediction studies (López-Cruz et al., 2015; Zhang et al., 2015; Crossa et al., 2016a; Saint Pierre et al., 2016). We found that GK methods improved the prediction ability of all single-environment and multi-environment models for CV2 design for all traits in HEL and USP datasets. Prediction of all traits with multi-environment models incorporating the G×E term and the GK method gave better prediction accuracy (especially for complex traits such as GY) than the other model-methods. The increases in prediction accuracy using models with GK under random cross-validation 2 (CV2) are also reflected under random cross-validation 1 (CV1) for HEL data set, but to a lesser extent due to the difficulty of borrowing information of unobserved (unphenotyped) lines in all environments. Similar decreases in prediction accuracy were found by López-Cruz et al. (2015) when attempting to predict wheat lines in untested (unobserved) environments under the CV1 random partition scheme.

Furthermore, for trait GY, differences in CV1 vs CV2 for GB and GK methods for the two data sets might also be due to other factors such: (1) differences in dataset repeatability, higher in HEL data sets (0.81, 0.60, 0.86, 0.69, and 0.78 for each of the 5 environments) than in USP (0.22, 0.42, 0.19, and 0.42 for each of the four environments) (Tables 2 and 3, respectively) due to better quality of the replicated trials in HEL than the unreplicated trials in USP; (2) more opportunity to borrow information from the close relatives hybrids in HEL (genetic diversity =0.175) than from the less related hybrids included in USP dataset (genetic diversity =0.372). For $h$=1 (bandwidth parameter) the Gaussian kernel is a direct function of the squared Euclidean distance between hybrids based on markers ($d_{ii'}$), thus for a set of unrelated hybrids $d_{ii'} \rightarrow$ large and $\boldsymbol{K} \rightarrow \mathbf{I}$ (identity matrix), that is, the Gaussian kernel weight more heavily the within (own)

hybrid information compared to the between hybrid (from relatives) information than GB (the marker information is of no use) (de los Campos et al., 2010). On the other hand, for a set of related hybrids then $d_{ii'} \rightarrow 0$ and $\boldsymbol{K} \rightarrow \mathbf{1}$ (matrix of ones), that is, marker information given by the Gaussian kernel weight equally or less the own hybrid information compared to information from relatives given by GB. Therefore, we speculate that for related hybrids of HEL dataset GK gives heavier weights to the within hybrid as well as between hybrids than the GB method does, that is why for GY, CV1 and CV2 for the multi-environment models gave very good increase in prediction accuracy over the GB, (3) for CV2 G×E modulates both values (within and between hybrid information); hybrids in USP (doubled the genetic diversity) less related than those from HEL data set, then it is expected negligible values between hybrids, and similar values assign by GK and GB to within hybrid information; for CV2, predictions accuracy depends, to a great extent, on the correlations between environments.

In summary, for CV2, for the prediction accuracy of GY in the HEL dataset, the method GK weights the own and the between hybrids performance as well as the relationship between environment (G×E modeled by MDs y MDe), whereas for USP, GK weights only the own hybrid performance and the relationship between environment becomes more important that in the HEL data set.

In the studies of López-Cruz et al. (2015) and Cuevas et al. (2016a), the data were balanced in the sense that all the individuals were included at the same time in all the environments. On the contrary, in this study, we had a great deal of imbalance because different numbers of maize hybrids were included in different environments; this was more pronounced in the HEL data set than in the USP data set and different R scripts were necessary for implementing the model-method combinations.

### 3.4.2. Prediction accuracy using linear and non-linear kernel methods

According to Gianola et al. (2014), Gaussian kernel has better predictive ability and a more flexible structure than GBLUP. Another point is that GK can capture non-additive effects between markers. Jiang and Reif (2015) evaluated maize data sets and investigated whether the prediction accuracy across connected bi-parental families can be increased by modeling additive × additive epistasis; the authors found that the prediction accuracy of RKHS (including epistasis) was superior to that of GBLUP (ignoring epistasis). Our study clearly shows a large increase in predictive ability when the G×E model and the GK method are combined for all traits, but mainly for GY (both data sets) and PH (USP data set). This indicates that to increase predictive

ability, it is important to consider the non-additive effect (i.e., epistasis) as the genetic relatedness across connected populations.

There are different choices for computing kernel functions: for example, linear kernel matrices incorporate only additive effects of the markers, polynomials kernels of different orders might incorporate different degrees of marker interactions, and the Gaussian kernel function uses complex epistatic marker interaction (Akdemir and Jannink, 2015). In genomic selection, additive kernels of the GBLUP type are employed when predicting breeding values, whereas when attempting to predict genetic values, Gaussian kernels would be more appropriate. Akdemir and Jannink (2015) demonstrated that epistatic marker effects in local regions of the chromosome with low recombination are stable through generations and offer the opportunity to exploit epistasis for improving genomic-enabled prediction accuracy; the authors defined local kernels in regions of the genome and calculated separate kernels for each region. The results shown in this study clearly showed the benefit of exploiting these local epistatic effects captured in the Gaussian Kernel and their interaction with environments.

The flexible structure of the MDs and MDe models is important, especially when combined with kernels that capture non-additive effects, as previously proven. In general, in the two maize data sets, the MDs-GK and MDe-GK models had better prediction accuracy than other models. The increase was not higher due to the occurrence of intermediate-to-high positive correlations between the analyzed environments, resulting in low G×E interaction. These results corroborate those obtained by López-Cruz et al. (2015) and Cuevas et al. (2016a), who observed that the intensity of the environmental correlation is related to the proportion of the genomic variance explained by the genetic main effects of markers across environments and genetic-specific effects of markers in environments.

### 3.4.3. Better fit of the G×E Gaussian kernel models

For GY of the HEL data set, the better fit of the model-method MDe-GB over the MDs-GB and MM-GB models is evident in their residual variance components: 0.227, 0.230, and 0.336, respectively. Similarly, for GY, the better fit of model MDe-GK over the MDs-GK and MM-GK models is also evident in their residual variance components: 0.093, 0.107, and 0.273, respectively. These trends in the residuals of the models are also found in traits PH and EH. The residuals of GK models were lower than those of the GBLUP models, indicating the better fit of the non-linear vs the linear kernel methods.

For GY of the USP data set, similar trends in the residuals of model-method MDe-GB vs the MDs-GB and MM-GB models were found: 0.840, 0.836, and 0.864, respectively. Similar clear trends in the residuals of these models were found for traits PH and EH. These patterns are also clear for the residuals of GK model-methods MDe-GK, MDs-GK, and MM-GK for trait EH but not for trait PH.

### 3.4.4. Prediction accuracy using multi-environment models

In this study, all pairwise correlations between environments were high and positive. This is important, because the G×E model of López-Cruz et al. (2015) has the limitation of better and more efficient prediction when applied to subsets of environments that have positive and similar correlations (Crossa et al., 2016b; Cuevas et al., 2016a). In positively correlated environments, the main marker effects are the most influential components when predicting genetic values; their variance component is high, and this produces better prediction accuracy than the single-environment model (López-Cruz et al., 2015; Cuevas et al., 2016a). Environments with intermediate or high positive correlations indicate little G×E interaction. Thus, the model reacts by reducing the specific marker effect. A correlation that is negative or close to zero implies strong G×E interaction, making it difficult to predict one environment based on information from another. With this, the model reduces the main effect of the marker and increases the specific effects. To work around the limitations of the G×E models, Cuevas et al. (2016b) developed multi-environment Bayesian genomic models that allow an arbitrary genetic covariance structure between environments, because an unstructured covariance matrix was used and its parameters were estimated from the data.

In this study, the prediction of multi-environment models was assessed by applying the cross-validation strategy called CV2 in Burgueño et al. (2012) and (Cuevas et al., 2016a), where some lines are represented in some environments but not in others. This CV2 validation strategy performed better than cross-validation strategy CV1 (where lines have not been evaluated in any field trials) proposed by Burgueño et al. (2012), when applied in multi-environment models (Jarquín et al., 2014; López-Cruz et al., 2015; Crossa et al., 2016b; Saint Pierre et al., 2016).

When introducing interaction effects, models MDs and the MDe showed increases in prediction accuracy for GY in most environments for the HEL and USP data sets over the single-environment model. Furthermore, models with GK had superior prediction accuracy than GB models. For traits PH and EH in both data sets, not much increase in prediction accuracy of MDs and MDe models was achieved over the single-environment model. These results were

expected because the genetic architecture of traits PH and EH is less complex than that of GY and less influenced by environmental factors. When main marker effects and interaction effects are introduced in the model using covariance structures this improves prediction accuracy.

The increase in accuracy when the interaction term for complex traits was included concurs with the findings of Crossa et al. (2016a), where increases in prediction accuracy were achieved by including dense molecular markers and G×E in a set of Mexican and Iranian landraces. Zhang et al. (2015) also used multi-environment models incorporating G×E and obtained an increase in prediction accuracy of several maize bi-parental populations. Similarly, Saint Pierre et al. (2016) evaluated spring wheat lines and introduced other environmental covariates in the Jarquin's model, showing that the prediction models gave better predictions using random cross-validation. Recently, results of extensive analyses of spring wheat trials across international environments conducted for several years in South and West Asia, North Africa and Mexico showed a consistent increase in the genomic prediction accuracy of the MDs-GB model over the MM-GB and SM-GB models (Sukumaran et al., 2017). One limitation of the data used in this study is that only one year was available for assessing the genomic-enabled accuracy of the various models.

Our results show that predictions with medium-to-high accuracy for genomic selection programs can be expected in environments with low-to-high heritability. Despite the low heritability of GY in P-LN and A-LN environments, the results were similar to results in other environments that have medium-to-high heritability. But in general, for the USP data set, prediction accuracy of all target traits under stress conditions was lower than under ideal nitrogen conditions, while in stress environments, heritability was lower than under ideal nitrogen conditions. These results agree with those obtained by Zhang et al. (2015), who found lower genomic prediction accuracy under water stress conditions than under ideal conditions, especially for complex traits such as GY.

## 3.5. CONCLUSION

Incorporating the Gaussian kernel method into the model of Jarquín et al. (2014) increased prediction accuracy as compared to the model used with the linear kernel GBLUP; these results were found in both data sets with the models used in this study: MM across environments, MDs and MDe. Other results of the application of eight model-method combinations between four models (SM, MM, MDs, and MDE) and two kernel methods (GBLUP and Gaussian kernel) in two extensive maize data sets show that (1) genomic models

incorporating G×E interaction had higher prediction accuracy than single-environment models; and (2) models with non-linear Gaussian kernel had higher prediction accuracy than models with linear kernel GBLUP. The model-method combinations with the highest prediction accuracy were MDe-GK and MDs-GK. Results of this study indicated that by employing appropriate statistical genomic-enabled prediction models the researchers and plant breeders can improve the prediction of hybrids that were not evaluated in several environments. Random cross-validation 2 (CV2) mimics sparse (incomplete) testing that allow to safe resources to the breeding program while improving prediction accuracy. Further research is required to compare the genomic-enabled G×E kernel prediction models used in this study with the models recently developed by Cuevas et al. (2016b), who studied genomic-enabled G×E models by means of Kronecker products applied to unstructured covariance matrices. It is also necessary to develop efficient computing software for fitting the structures considered in the models of this study while maintaining the inferential advantages of the Markov Chain Monte Carlo.

## REFERENCES

Akdemir, D., and J-L. Jannink. 2015. Locally epistatic genomic relationship matrices for genomic association and prediction. Genetics 199.3 857–871.

Bernardo, R., and J. Yu. 2007. Prospects for genome-wide selection for quantitative traits in maize. Crop Sci. 47(3): 1082–1090.

Beyene, Y., K. Semagn, S. Mugo, A. Tarekegne, R. Babu, B. Meisel, P. Sehabiague, D. Makumbi, C. Magorokosho, S. Oikeh, J. Gakunga, M. Vargas, M. Olsen, B.M. Prasanna, M. Banziger, and J. Crossa. 2015. Genetic Gains in Grain Yield Through Genomic Selection in Eight Bi-parental Maize Populations under Drought Stress. Crop Sci. 55(1): 154–163.

Browning, B.L., and S.R. Browning. 2008. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84(2): 210–223.

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

Crossa, J., D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño, C. Saint-Pierre, P. Vikram, C. Sansaloni, C. Petroli, D. Akdemir, C. Sneller, M. Reynolds, M. Tattaris, T. Payne, C. Guzman, R.J. Peña, P. Wenzl, and S. Singh. 2016a. Genomic Prediction of Gene Bank Wheat Landraces. G3: Genes|Genomes|Genetics 6(7): 1819–1834.

Crossa, J., G. de los Campos, M. Maccaferri, R. Tuberosa, J. Burgueño, and P. Pérez-Rodríguez. 2016b. Extending the marker × environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Sci. 56(5): 2193–2209.

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

Crossa, J., P. Pérez, G. de los Campos, G. Mahuku, S. Dreisigacker, and C. Magorokosho. 2011. Genomic Selection and Prediction in Plant Breeding. J. Crop Improv. 25(3): 239–261.

Cuevas, J., J. Crossa, O. Montesinos-Lopez, J. Burgueno, P. Pérez-Rodríguez, and G. de los Campos. 2016a. Bayesian Genomic Prediction with Genotype × Environment Interaction Kernel Models. G3: Genes|Genomes|Genetics 7:41–53.

Cuevas, J., J. Crossa, V. Soberanis, S. Pérez-Elizalde, P. Pérez-Rodríguez, G. de los Campos, O.A. Montesinos-López, and J. Burgueño. 2016b. Genomic Prediction of Genotype × Environment Interaction Kernel Regression Models. The Plant Genome 9(3):1–20. doi:10.3835/plantgenome2016.03.0024

de los Campos, G., D. Gianola, G.J.M. Rosa, K.A. Weigel, and J. Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92(4): 295–308.

de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2): 327–345.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1): 375–385.

de los Campos, G., and P. Pérez-Rodríguez. 2016. BGLR: Bayesian generalized linear regression. R package version 1.0.5. : https://CRAN.R-project.org/package=BGLR.

Gianola, D., and J. B. van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289-2303.

Gianola, D., R.L. Fernando, and A. Stella. 2006. Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. Genetics 173(3): 1761–1776.

Gianola, D., K.A. Weigel, N. Krämer, A. Stella, and C.C. Schön. 2014. Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS One 9(4).

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4): 2389–2397.

Heslot, N., D. Akdemir, M.E. Sorrells, and J.L. Jannink. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor. Appl. Genet. 127(2): 463–480.

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.L. Jannink. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. 52(1): 146.

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt, J. Lorgeou, F. Piraux, L. Guerreiro, P. Pérez, M. Calus, J. Burgueño, and G. de los Campos. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127(3): 595–607.

Jiang, Y., and J.C. Reif. 2015. Modeling epistasis in genomic selection. Genetics 201(2): 759–768.

López-Cruz, M., J. Crossa, D. Bonnett, S. Dreisigacker, J. Poland, J.-L. Jannink, R.P. Singh, E. Autrique, and G. de los Campos. 2015. Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3: Genes|Genomes|Genetics. 5(4): 569–82.

Massman, J.M., H.J.G. Jung, and R. Bernardo. 2013. Genome-wide selection vs marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci. 53(1): 58–66.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819–1829.

Morota, G., and D. Gianola. 2014. Kernel-based whole-genome prediction of complex traits: A review. Front. Genet. 5: 1–13. doi:10.3389/fgene.2014.00363

Morota, G., M. Koyama, G.J.M. Rosa, K.A. Weigel, and D. Gianola. 2013. Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. Genet. Sel. Evol. 45(1): 17.

Pérez-Elizalde, S., J. Cuevas, P. Pérez-Rodríguez, and J. Crossa. 2015. Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction. J. Agric. Biol. Environ. Stat. 20(4): 512–532.

Pérez Rodríguez, P., Crossa, J., Bondalapati, K., De Meyer, G., Pita, F., and de los Campos, G. 2015. A pedigree reaction norm model for prediction of cotton (*Gossypium* sp.) yield in multi-environment trials. Crop Science 55, 1143-1151 doi: 10.2135/cropsci2014.08.0577

Pérez-Rodríguez, P., and G. de los Campos. 2014. Genome-Wide Regression & Prediction with the BGLR Statistical Package. Genetics 198: 483-495.

Pérez-Rodríguez, P., D. Gianola, J.M. Gonzalez-Camacho, J. Crossa, Y. Manes, and S. Dreisigacker. 2013. Comparison Between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat. G3: Genes|Genomes|Genetics 2(12): 1595–1605.

Pérez-Rodríguez, P., D. Gianola, J.M. González-Camacho, J. Crossa, Y. Manès, and S. Dreisigacker. 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3: Genes|Genomes|Genetics. 2(12): 1595–605.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Saint Pierre, C., J. Burgueño, J. Crossa, G. Fuentes Dávila, P. Figueroa López, E. Solís Moya, J. Ireta Moreno, V.M. Hernández Muela, V.M. Zamora Villa, P. Vikram, K. Mathews, C. Sansaloni, D. Sehgal, D. Jarquín, P. Wenzl, and S. Singh. 2016. Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. Sci. Rep. 6: 27312.

Sukumaran, S., Crossa, J., Jarquín, D., Lopes,M., and Reynolds, M.P. 2017. Genomic Prediction with Pedigree and Genotype × Environment Interaction in Spring Wheat Grown in South and West Asia, North Africa, and Mexico. G3: Genes|Genomes|Genetics 7:481-495. doi: 10.1534/g3.116.036251

Unterseer, S., E. Bauer, G. Haberer, M. Seidel, C. Knaak, M. Ouzunova, T. Meitinger, T.M. Strom, R. Fries, H. Pausch, C. Bertani, A. Davassi, K.F. Mayer, and C.-C. Schön. 2014. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15(1): 823.

VanRaden, P.M. 2007. Genomic Measures of Relationship and Inbreeding. Interbull Annu. Meet. Proc. 37: 33–36.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91(11): 4414–4423.

Wimmer, A.V., H. Auinger, T. Albrecht, C. Schoen, L. Schaeffer, M. Erbe, U. Ober, Y. Badke, and P. Vandehaar. 2015. synbreed: Framework for the analysis of genomic prediction data using R. (1): 1–43.

Zhang, X., P. Pérez-Rodríguez, K. Semagn, Y. Beyene, R. Babu, M.A. López-Cruz, F. San Vicente, M. Olsen, E. Buckler, J.-L. Jannink, B.M. Prasanna, and J. Crossa. 2015. Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity 114(3): 291–9.

**Table 1.** Model, kernel method, and abbreviation of the model-method combination

| Model* | Kernel method | Abbreviation |
|---|---|---|
| Single-environment, main genotypic effect model (SM) $\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{\varepsilon}$ (1) | GBLUP (GB) | SM-GB |
| | Gaussian Kernel (GK) | SM-GK |
| Multi-environment, main genotypic effect (MM) $\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_E\boldsymbol{\beta}_E + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{\varepsilon}$ (2) | GBLUP (GB) | MM-GB |
| | Gaussian Kernel (GK) | MM-GK |
| Multi-environment, single variance G×E deviation model (MDs) $\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_E\boldsymbol{\beta}_E + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{ue} + \boldsymbol{\varepsilon}$ (3) | GBLUP (GB) | MDs-GB |
| | Gaussian Kernel (GK) | MDs-GK |
| Multi-environment, environment-specific variance G×E deviation model (MDe) $\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_E\boldsymbol{\beta}_E + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{u}_E + \boldsymbol{\varepsilon}$ (4) | GBLUP (GB) | MDe-GB |
| | Gaussian Kernel (GK) | MDe-GK |

* Model-methods are described in the Material and Methods section.

**Table 2**. Heritability and phenotypic correlations among five environments for grain yield and three environments for plant height and ear height for the HEL data set

| | Grain Yield | | | | |
|---|---|---|---|---|---|
| Site | Ipiaçú | Nova Mutum | Pato de Minas | Sertanópolis | Sorriso |
| Nova Mutum | 0.46 | - | - | - | - |
| Pato de Minas | 0.51 | 0.44 | - | - | - |
| Sertanópolis | 0.29 | 0.36 | 0.30 | - | - |
| Sorriso | 0.43 | 0.48 | 0.39 | 0.38 | - |
| | Lower diagonal: Plant Height; upper diagonal: Ear Height | | | | |
| Ipiaçú | - | - | 0.74 | 0.73 | - |
| Pato de Minas | 0.73 | - | - | 0.73 | - |
| Sertanópolis | 0.73 | - | 0.74 | - | - |
| Trait | Heritability | | | | |
| Grain Yield | 0.81 | 0.60 | 0.86 | 0.69 | 0.78 |
| Plant Height | 0.72 | - | 0.86 | 0.91 | - |
| Ear Height | 0.69 | - | 0.81 | 0.89 | - |

**Table 3.** Heritability and phenotypic correlations among four environments for grain yield, plant height and ear height for the USP data set

| | Grain Yield | | | |
|---|---|---|---|---|
| Environment | Piracicaba-LN* | Piracicaba-IN | Anhumas-LN | Anhumas-IN |
| Piracicaba-IN | 0.54 | - | - | - |
| Anhumas-LN | 0.31 | 0.35 | - | - |
| Anhumas-IN | 0.43 | 0.47 | 0.47 | - |
| | Lower diagonal: Plant Height; upper diagonal: Ear Height | | | |
| Piracicaba-LN | - | 0.80 | 0.71 | 0.78 |
| Piracicaba-IN | 0.75 | - | 0.69 | 0.78 |
| Anhumas-LN | 0.68 | 0.67 | - | 0.71 |
| Anhumas-IN | 0.76 | 0.78 | 0.70 | - |
| Trait | Heritability | | | |
| Grain Yield | 0.22 | 0.42 | 0.19 | 0.42 |
| Plant Height | 0.72 | 0.85 | 0.60 | 0.84 |
| Ear Height | 0.73 | 0.87 | 0.66 | 0.91 |

*LN (low nitrogen); IN (ideal nitrogen).

**Table 4.** HEL data set. Estimates of different variance components for single-environment, main genotypic effect GBLUP kernel (SM-GB), single-environment, main genotypic effect Gaussian kernel (SM-GK), multi-environment, main genotypic effect GBLUP kernel (MM-GB), multi-environment, main genotypic effect Gaussian kernel (MM-GK), multi-environment, single variance G×E deviation model GBLUP kernel (MDs-GB), multi-environment, single variance G×E deviation model Gaussian kernel (MDs-GK), multi-environment, environment-specific variance G×E deviation model GBLUP kernel (MDe-GB), multi-environment, environment-specific variance G×E deviation model Gaussian kernel (MDe-GK) for three traits: grain yield, plant height and ear height

| **Single-environment, main genotypic effect (SM)** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Component* | Environment | Grain yield | | Plant height | | Ear height | |
| | | SM-GB | SM-GK | SM-GB | SM-GK | SM-GB | SM-GK |
| Residual ($\sigma^2_{e_j}$) | Ipiaçú | 0.476 (0.05) | 0.259 (0.05) | 0.397 (0.05) | 0.243 (0.05) | 0.419 (0.05) | 0.246 (0.05) |
| | Nova Mutum | 0.715 (0.06) | 0.431 (0.07) | - | - | - | - |
| | Pato de Minas | 0.440 (0.03) | 0.221 (0.03) | 0.332 (0.03) | 0.151 (0.02) | 0.330 (0.03) | 0.163 (0.02) |
| | Sertanópolis | 0.761 (0.07) | 0.467 (0.07) | 0.341 (0.05) | 0.198 (0.03) | 0.387 (0.03) | 0.225 (0.03) |
| | Sorriso | 0.742 (0.07) | 0.467 (0.07) | - | - | - | - |
| Genetic effect ($\sigma^2_{u_j}$) | Ipiaçú | 0.442 (0.10) | 0.965 (0.13) | 0.794 (0.17) | 1.231 (0.23) | 0.726 (0.18) | 1.221 (0.20) |
| | Nova Mutum | 0.391 (0.12) | 1.177 (0.27) | - | - | - | - |
| | Pato de Minas | 0.462 (0.09) | 1.106 (0.15) | 0.884 (0.16) | 1.331 (0.15) | 0.903 (0.16) | 1.278 (0.15) |
| | Sertanópolis | 0.290 (0.09) | 1.156 (0.27) | 0.855 (0.14) | 1.224 (0.17) | 0.736 (0.16) | 1.199 (0.17) |
| | Sorriso | 0.299 (0.09) | 1.132 (0.25) | - | - | - | - |
| **Multi-environment, main genotypic effects (MM)** | | | | | | | |
| | | MM-GB | MM-GK | MM-GB | MM-GK | MM-GB | MM-GK |
| Residual ($\sigma^2_{\varepsilon}$) | | 0.336 (0.01) | 0.273 (0.01) | 0.324 (0.02) | 0.259 (0.01) | 0.299 (0.01) | 0.222 (0.01) |
| GME ($\sigma^2_{u_0}$) | | 0.129 (0.03) | 0.316 (0.04) | 0.758 (0.12) | 0.769 (0.08) | 0.802 (0.13) | 0.846 (0.08) |
| **Multi-environment, single variance G×E deviation (MDs)** | | | | | | | |
| | | MDs-GB | MDs-GK | MDs-GB | MDs-GK | MDs-GB | MDs-GK |
| Residual ($\sigma^2_{\varepsilon}$) | | 0.230 (0.00) | 0.107 (0.01) | 0.274 (0.01) | 0.164 (0.01) | 0.294 (0.01) | 0.188 (0.01) |
| GME ($\sigma^2_{u_o}$) | | 0.122 (0.03) | 0.375 (0.04) | 0.798 (0.13) | 0.844 (0.08) | 0.763 (0.13) | 0.764 (0.08) |
| GIE ($\sigma^2_{ue}$) | | 0.111 (0.01) | 0.244 (0.24) | 0.04 (0.01) | 0.124 (0.03) | 0.051 (0.01) | 0.155 (0.01) |
| **Multi-environment, environment-specific variance G×E deviation (MDe)** | | | | | | | |
| | | MDe-GB | MDe-GK | MDe-GB | MDe-GK | MDe-GB | MDe-GK |
| Residual ($\sigma^2_{\varepsilon}$) | | 0.227 (0.01) | 0.093 (0.01) | 0.293 (0.01) | 0.161 (0.01) | 0.274 (0.01) | 0.162 (0.01) |
| GME ($\sigma^2_{u_o}$) | | 0.070 (0.02) | 0.264 (0.03) | 0.747 (0.12) | 0.834 (0.08) | 0.798 (0.13) | 0.837 (0.08) |
| Genetic environment specific effect ($\sigma^2_{uE_j}$) | Ipiaçú | 0.278 (0.07) | 0.658 (0.11) | 0.046 (0.02) | 0.096 (0.04) | 0.040 (0.02) | 0.09 (0.03) |
| | Nova Mutum | 0.024 (0.01) | 0.103 (0.03) | - | - | - | - |
| | Pato de Minas | 0.320 (0.07) | 0.792 (0.10) | 0.036 (0.01) | 0.091 (0.03) | 0.034 (0.02) | 0.090 (0.03) |
| | Sertanópolis | 0.018 (0.01) | 0.052 (0.02) | 0.080 (0.03) | 0.220 (0.07) | 0.054 (0.03) | 0.210 (0.07) |
| | Sorriso | 0.023 (0.01) | 0.077 (0.02) | - | - | - | - |

*Genetic main effect (GME); Genetic Interaction effect (GIE)

**Table 5.** USP data set. Estimates of different variance components for single-environment, main genotypic effect GBLUP kernel (SM-GB), single-environment, main genotypic effect Gaussian kernel (SM-GK), multi-environment, main genotypic effect model GBLUP kernel (MM-GB), multi-environment, main genotypic effect Gaussian kernel (MM-GK), multi-environment, single variance G×E deviation model GBLUP kernel (MDs-GB), multi-environment, single variance G×E deviation model Gaussian kernel (MDs-GK), multi-environment, environment-specific variance G×E deviation model GBLUP kernel (MDe-GB), multi-environment, environment-specific variance G×E deviation model Gaussian kernel (MDe-GK) for three traits: grain yield, plant height and ear height

| **Single-environment, main genotypic effect (SM)** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Component* | Environment** | Grain yield | | Plant height | | Ear height | |
| | | SM-GB | SM-GK | SM-GB | SM-GK | SM-GB | SM-GK |
| Residual ($\sigma^2_{e_j}$) | P-LN | 0.878 (0.04) | 0.797 (0.05) | 0.749 (0.04) | 0.656 (0.04) | 0.503 (0.03) | 0.377 (0.03) |
| | P-IN | 0.851 (0.05) | 0.786 (0.05) | 0.685 (0.04) | 0.594 (0.04) | 0.492 (0.03) | 0.390 (0.03) |
| | A-LN | 0.873 (0.05) | 0.817 (0.05) | 0.812 (0.04) | 0.726 (0.05) | 0.627 (0.03) | 0.513 (0.04) |
| | A-IN | 0.780 (0.04) | 0.699 (0.05) | 0.654 (0.03) | 0.543 (0.04) | 0.494 (0.03) | 0.375 (0.03) |
| Genetic effect ($\sigma^2_{u_j}$) | P-LN | 0.132 (0.04) | 0.438 (0.12) | 0.257 (0.07) | 0.619 (0.13) | 0.556 (0.13) | 0.955 (0.15) |
| | P-IN | 0.149 (0.05) | 0.420 (0.10) | 0.341 (0.08) | 0.683 (0.14) | 0.537 (0.12) | 0.829 (0.15) |
| | A-LN | 0.124 (0.04) | 0.366 (0.09) | 0.188 (0.05) | 0.523 (0.14) | 0.397 (0.10) | 0.808 (0.07) |
| | A-IN | 0.232 (0.06) | 0.550 (0.12) | 0.383 (0.09) | 0.747 (0.14) | 0.549 (0.12) | 0.922 (0.08) |
| **Multi-environment, main genotypic effects (MM)** | | | | | | | |
| | | MM-GB | MM-GK | MM-GB | MM-GK | MM-GB | MM-GK |
| Residual ($\sigma^2_{\varepsilon}$) | | 0.864 (0.02) | 0.656 (0.02) | 0.705 (0.02) | 0.319 (0.01) | 0.519 (0.01) | 0.279 (0.01) |
| GME ($\sigma^2_{u_0}$) | | 0.167 (0.04) | 1.39 (0.16) | 0.341 (0.07) | 3.131 (0.22) | 0.56 (0.12) | 1.861 (0.14) |
| **Multi-environment, single variance G×E deviation (MDs)** | | | | | | | |
| | | MDs-GB | MDs-GK | MDs-GB | MDs-GK | MDs-GB | MDs-GK |
| Residual ($\sigma^2_{\varepsilon}$) | | 0.836 (0.02) | 0.536 (0.02) | 0.699 (0.02) | 0.304 (0.01) | 0.509 (0.01) | 0.264 (0.01) |
| GME ($\sigma^2_{u_0}$) | | 0.170 (0.04) | 1.965 (0.18) | 0.343 (0.08) | 3.09 (0.21) | 0.547 (0.12) | 1.819 (0.13) |
| GIE ($\sigma^2_{ue}$) | | 0.024 (0.01) | 0.094 (0.02) | 0.008 (0.00) | 0.057 (0.01) | 0.019 (0.01) | 0.073 (0.01) |
| **Multi-environment, environment-specific variance G×E deviation (MDe)** | | | | | | | |
| | | MDe-GB | MDe-GK | MDe-GB | MDe-GK | MDe-GB | MDe-GK |
| Residual ($\sigma^2_{\varepsilon}$) | | 0.840 (0.02) | 0.536 (0.02) | 0.705 (0.02) | 0.305 (0.01) | 0.517 (0.01) | 0.261 (0.01) |
| GME ($\sigma^2_{u_0}$) | | 0.170 (0.04) | 1.966 (0.19) | 0.340 (0.08) | 3.168 (0.22) | 0.558 (0.12) | 1.894 (0.13) |
| Genetic environment specific effect ($\sigma^2_{uE_j}$) | P-LN | 0.013 (0.01) | 0.110 (0.04) | 0.006 (0.00) | 0.042 (0.01) | 0.006 (0.00) | 0.043 (0.01) |
| | P-IN | 0.016 (0.01) | 0.081 (0.03) | 0.005 (0.00) | 0.045 (0.01) | 0.006 (0.00) | 0.039 (0.01) |
| | A-LN | 0.014 (0.01) | 0.101 (0.04) | 0.004 (0.00) | 0.058 (0.02) | 0.008 (0.00) | 0.09 (0.03) |
| | A-IN | 0.020 (0.02) | 0.087 (0.03) | 0.004 (0.00) | 0.037 (0.01) | 0.006 (0.00) | 0.039 (0.01) |

*Genetic main effect (GME); Genetic Interaction effect (GIE); **Environments: Anhumas ideal N (A-IN), Anhumas low N (A-LN), Piracicaba ideal N (P-IN) and Piracicaba low N (P-LN).

**Table 6.** HEL data set. Mean correlation (for 50 random partitions CV2) for models single-environment, main genotypic effects model with GBLUP kernel method (SM-GB) and single-environment, main genotypic effects model with Gaussian kernel method (SM-GK), multi-environment, main genotypic effect model GBLUP kernel (MM-GB), multi-environment, main genotypic effect Gaussian kernel (MM-GK), multi-environment, single variance G×E deviation model GBLUP kernel (MDs-GB), multi-environment, single variance G×E deviation model Gaussian kernel (MDs-GK), multi-environment, environment-specific variance G×E deviation model GBLUP kernel (MDe-GB), multi-environment, environment-specific variance G×E deviation model Gaussian kernel (MDe-GK) for three traits, grain yield, plant height, and ear height (standard deviation in parentheses)

| Model / Environment* | Single environment, main genotypic effect model (SM) | | % Change | Multi-environment, main genotypic effect model (MM) | | % Change | Multi-environment, main, single variance G×E deviation model (MDs) | | % Change | Multi-environment, environment-specific variance G×E deviation model (MDe) | | % Change |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GB | GK | | GB | GK | | GB | GK | | GB | GK | |
| Grain yield | | | | | | | | | | | | |
| IP | 0.64 (0.08) | 0.73 (0.06) | 14 | 0.56 (0.09) | 0.64 (0.08) | 14 | 0.68 (0.09) | 0.81 (0.05) | 19 | 0.66 (0.09) | 0.80 (0.05) | 21 |
| NM | 0.37 (0.09 | 0.49 (0.09) | 32 | 0.47 (0.08) | 0.58 (0.07) | 23 | 0.47 (0.09) | 0.63 (0.07) | 34 | 0.43 (0.10) | 0.62 (0.06) | 44 |
| PM | 0.68 (0.05) | 0.74 (0.04) | 9 | 0.64 (0.07) | 0.70 (0.04) | 9 | 0.69 (0.05) | 0.78 (0.03) | 13 | 0.69 (0.05) | 0.78 (0.03) | 13 |
| SE | 0.36 (0.10) | 0.46 (0.08) | 28 | 0.25 (0.12) | 0.37 (0.11) | 48 | 0.40 (0.10) | 0.58 (0.09) | 45 | 0.39 (0.10) | 0.58 (0.09) | 49 |
| SO | 0.40 (0.10) | 0.52 (0.07) | 30 | 0.38 (0.08) | 0.53 (0.07) | 39 | 0.47 (0.09) | 0.69 (0.04) | 47 | 0.47 (0.08) | 0.68 (0.04) | 45 |
| Plant height | | | | | | | | | | | | |
| IP | 0.65 (0.08) | 0.69 (0.07) | 6 | 0.77 (0.06) | 0.80 (0.05) | 4 | 0.78 (0.06) | 0.81 (0.05) | 4 | 0.77 (0.06) | 0.80 (0.05) | 4 |
| PM | 0.76 (0.04) | 0.79 (0.05) | 4 | 0.79 (0.04) | 0.81 (0.04) | 3 | 0.80 (0.04) | 0.82 (0.04) | 3 | 0.80 (0.04) | 0.82 (0.04) | 3 |
| SE | 0.74 (0.05) | 0.76 (0.05) | 3 | 0.78 (0.04) | 0.80 (0.03) | 3 | 0.79 (0.05) | 0.82 (0.03) | 4 | 0.79 (0.05) | 0.82 (0.03) | 4 |
| Ear height | | | | | | | | | | | | |
| IP | 0.64 (0.08) | 0.67 (0.07) | 6 | 0.76 (0.05) | 0.76 (0.04) | 0 | 0.76 (0.05) | 0.77 (0.04) | 1 | 0.76 (0.05) | 0.77 (0.04) | 1 |
| PM | 0.76 (0.03) | 0.78 (0.04) | 4 | 0.78 (0.04) | 0.79 (0.03) | 1 | 0.79 (0.03) | 0.81 (0.03) | 3 | 0.79 (0.03) | 0.81 (0.03) | 3 |
| SE | 0.69 (0.05) | 0.72 (0.05) | 3 | 0.76 (0.05) | 0.78 (0.04) | 3 | 0.75 (0.05) | 0.78 (0.04) | 4 | 0.75 (0.05) | 0.79 (0.04) | 5 |

*Environments: Ipiaçú (IP), Nova Mutum (NM), Pato de Minas (PM), Sertanópolis (SE) and Sorriso (SO).

**Table 7.** USP data set. Mean correlation (for 50 random partitions CV2) for models single-environment, main genotypic effects model with GBLUP kernel method (SM-GB) and single-environment, main genotypic effects model with Gaussian kernel method (SM-GK), multi-environment, main genotypic effect model GBLUP kernel (MM-GB), multi-environment, main genotypic effect Gaussian kernel (MM-GK), multi-environment, single variance G×E deviation model GBLUP kernel (MDs-GB), multi-environment, single variance G×E deviation model Gaussian kernel (MDs-GK), multi-environment, environment-specific variance G×E deviation model GBLUP kernel (MDe-GB), multi-environment, environment-specific variance G×E deviation model Gaussian kernel (MDe-GK) for three traits, grain yield, plant height, and ear height (standard deviation in parentheses)

| Model / Environment* | Single environment, main genotypic effect model (SM) | | % Change | Multi-environment, main genotypic effect model (MM) | | % Change | Multi-environment, main, single variance G×E deviation model (MDs) | | % Change | Multi-environment, environment-specific variance G×E deviation model (MDe) | | % Change |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GB | GK | | GB | GK | | GB | GK | | GB | GK | |
| Grain yield | | | | | | | | | | | | |
| P-LN | 0.27 (0.04) | 0.29 (0.06) | 7 | 0.32 (0.05) | 0.51 (0.05) | 59 | 0.34 (0.05) | 0.56 (0.06) | 65 | 0.33 (0.05) | 0.56 (0.06) | 70 |
| P-IN | 0.32 (0.04) | 0.33 (0.06) | 3 | 0.35 (0.07) | 0.53 (0.07) | 51 | 0.36 (0.07) | 0.55 (0.06) | 53 | 0.36 (0.07) | 0.55 (0.06) | 53 |
| A-LN | 0.29 (0.05) | 0.31 (0.07) | 7 | 0.35 (0.08) | 0.50 (0.06) | 43 | 0.34 (0.07) | 0.51 (0.06) | 50 | 0.34 (0.07) | 0.51 (0.06) | 50 |
| A-IN | 0.40 (0.05) | 0.40 (0.07) | 0 | 0.44 (0.07) | 0.59 (0.05) | 34 | 0.43 (0.07) | 0.58 (0.06) | 35 | 0.43 (0.07) | 0.59 (0.06) | 37 |
| Plant height | | | | | | | | | | | | |
| P-LN | 0.45 (0.04) | 0.45 (0.07) | 0 | 0.53 (0.07) | 0.79 (0.03) | 49 | 0.52 (0.07) | 0.79 (0.03) | 52 | 0.51 (0.07) | 0.79 (0.03) | 55 |
| P-IN | 0.52 (0.04) | 0.51 (0.06) | -2 | 0.57 (0.05) | 0.79 (0.03) | 39 | 0.57 (0.05) | 0.79 (0.03) | 39 | 0.56 (0.05) | 0.79 (0.03) | 41 |
| A-LN | 0.37 (0.05) | 0.39 (0.07) | 5 | 0.46 (0.07) | 0.72 (0.04) | 57 | 0.46 (0.07) | 0.72 (0.04) | 57 | 0.45 (0.07) | 0.72 (0.05) | 61 |
| A-IN | 0.54 (0.04) | 0.55 (0.05) | 2 | 0.60 (0.05) | 0.83 (0.02) | 38 | 0.60 (0.06) | 0.82 (0.02) | 37 | 0.60 (0.06) | 0.82 (0.02) | 37 |
| Ear height | | | | | | | | | | | | |
| P-LN | 0.67 (0.03) | 0.67 (0.04) | 0 | 0.71 (0.05) | 0.84 (0.03) | 18 | 0.70 (0.05) | 0.84 (0.03) | 20 | 0.70 (0.05) | 0.84 (0.03) | 20 |
| P-IN | 0.68 (0.03) | 0.68 (0.04) | 0 | 0.72 (0.04) | 0.83 (0.02) | 15 | 0.71 (0.04) | 0.82 (0.02) | 15 | 0.71 (0.04) | 0.82 (0.03) | 15 |
| A-LN | 0.57 (0.03) | 0.58 (0.05) | 2 | 0.61 (0.05) | 0.75 (0.03) | 23 | 0.61 (0.05) | 0.75 (0.03) | 23 | 0.61 (0.05) | 0.75 (0.03) | 23 |
| A-IN | 0.69 (0.03) | 0.69 (0.04) | 0 | 0.72 (0.03) | 0.84 (0.02) | 17 | 0.71 (0.03) | 0.84 (0.02) | 18 | 0.71 (0.03) | 0.84 (0.02) | 18 |

*Environments: Anhumas ideal N (A-IN), Anhumas low N (A-LN), Piracicaba ideal N (P-IN) and Piracicaba low N (P-LN)

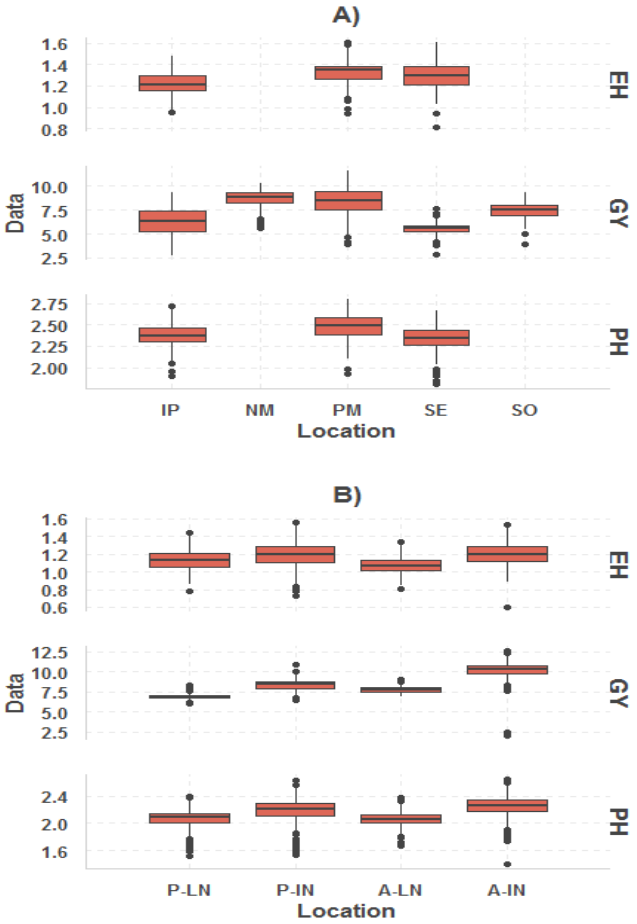


**Figure 1**. Box plot of grain yield (Mg ha$^{-1}$) (GY), plant height (cm) (PH), and ear height (cm) (EH) for: (A) HEL data set and (B) USP data set. Environments for the HEL data set are IP: Ipiaçú, NM: Nova Mutum, PM: Pato de Minas, SE: Sertanópolis, and SO: Sorriso. Environments for the USP data set are: Pir-LN: Piracicaba low nitrogen, Pir-IN: Piracicaba ideal nitrogen, Anh-LN: Anhumas low nitrogen, and Anh-IN: Anhumas ideal nitrogen.

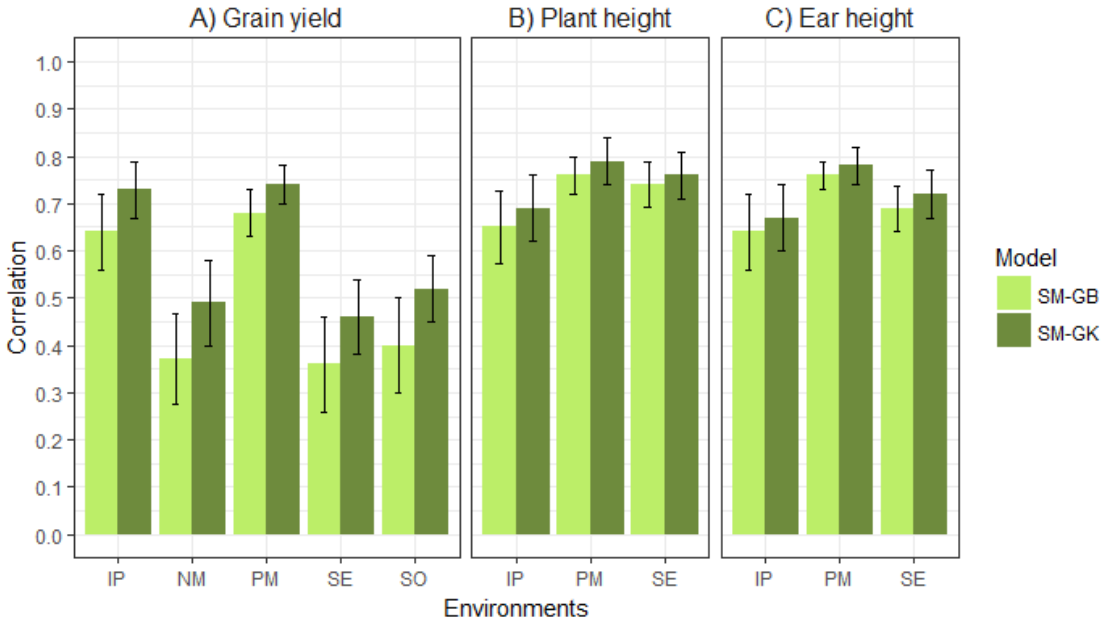


**Figure 2**. HEL data set. Correlation between phenotypes and prediction values (average of 50 random partitions) for single-environment, main genotypic effect model with GBLUP kernel method (SM-GB) and single-environment, main genotypic effect model with Gaussian kernel method (SM-GK) for: (A) five environments (horizontal axis) for grain yield, (B) three environments for plant height and (C) three environments for ear height. Environments are: Ipiaçú (IP), Nova Mutum (NM), Pato de Minas (PM), Sertanópolis (SE) and Sorriso (SO). Error bars show standard deviations.

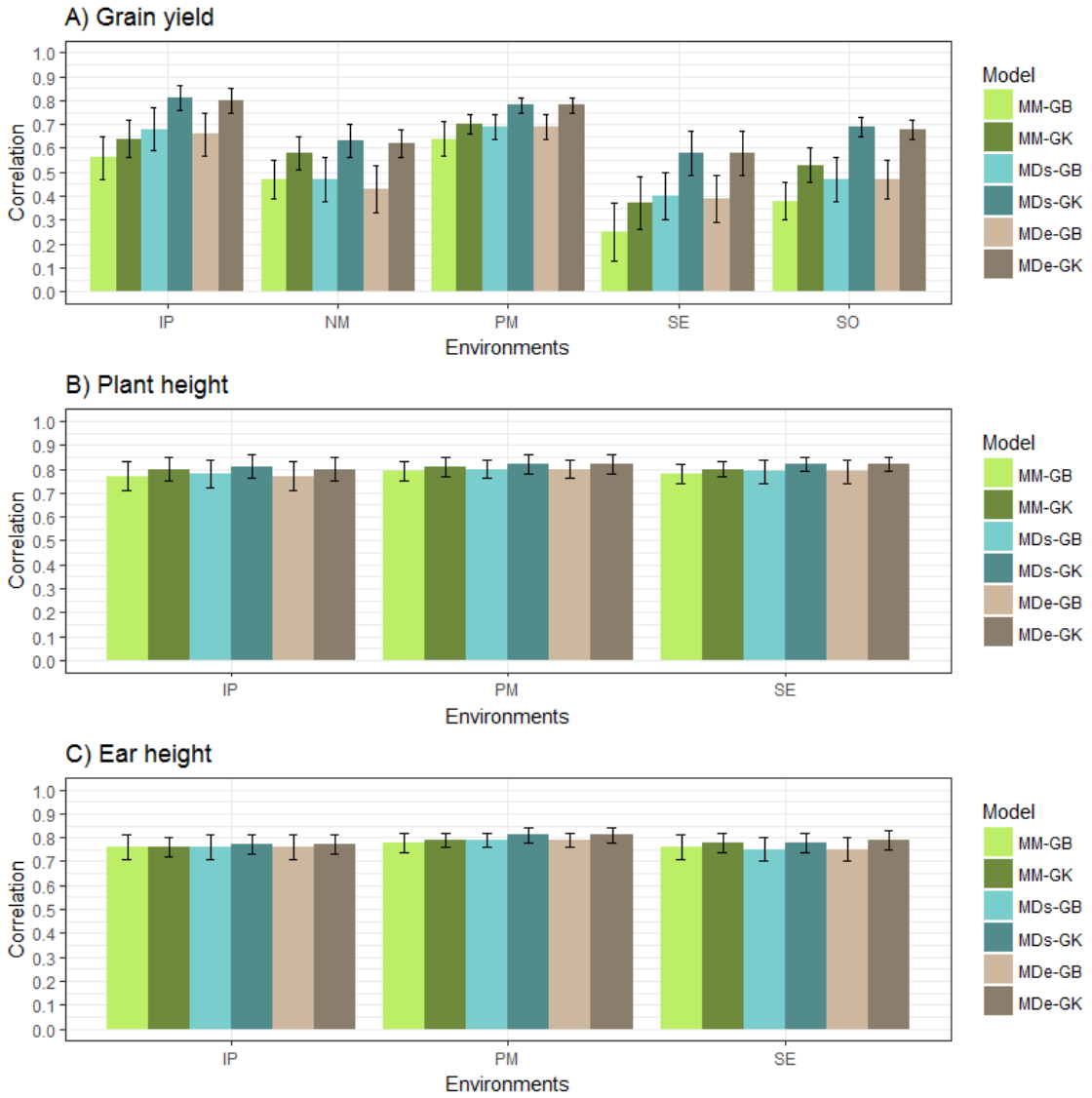


**Figure 3**. HEL data set. Mean correlation between observed and predictive values (average of 50 random cross-validations partitions, CV2) for multi-environment, main genotypic effect model GBLUP kernel (MM-GB), multi-environment, main genotypic effect Gaussian kernel (MM-GK), multi-environment, single variance G×E deviation model GBLUP kernel (MDs-GB), multi-environment, single variance G×E deviation model Gaussian kernel (MDs-GK), multi-environment, environment-specific variance G×E deviation model GBLUP kernel (MDe-GB), multi-environment, environment-specific variance G×E deviation model Gaussian kernel (MDe-GK) for: (A) five environments (horizontal axis) for grain yield, (B) three environments for plant height and (C) three environments for ear height. Environments are: Ipiaçú (IP), Nova Mutum (NM), Pato de Minas (PM), Sertanópolis (SE) and Sorriso (SO). Error bars show standard deviations.

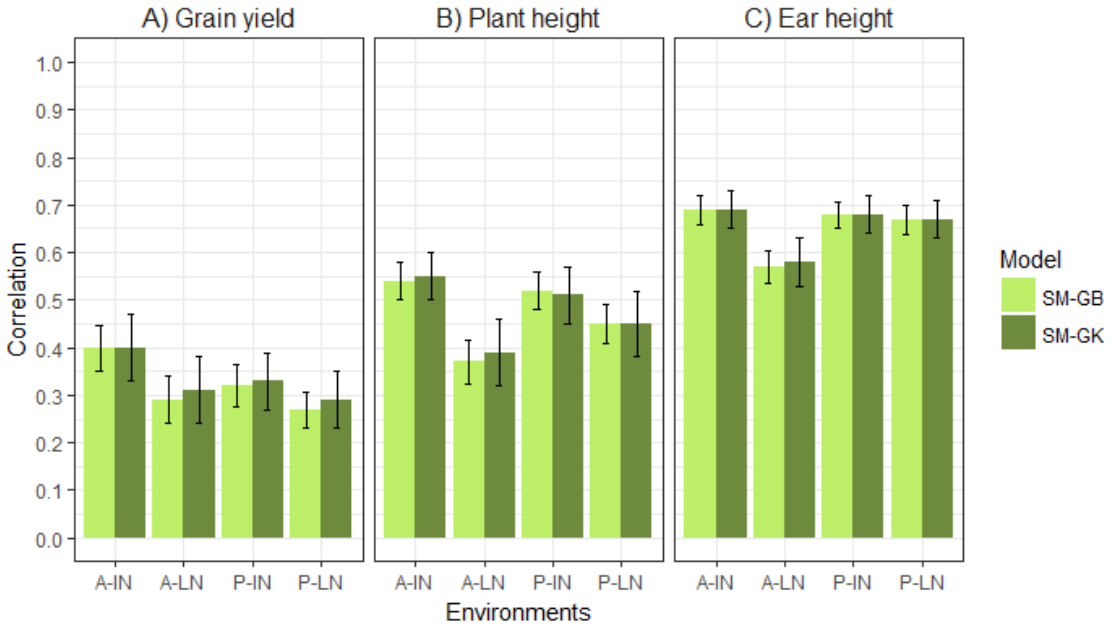


**Figure 4**. USP data set. Mean correlation between phenotypes and predictions (average of 50 random partitions) for single-environment, main genotypic effects model model with GBLUP kernel method (SM-GB) and single-environment, main genotypic effects model with Gaussian kernel method (SM-GK) in four environments (horizontal axis) for (A) grain yield, (B) plant height and (C) ear height. Environments are: Anhumas ideal Nitrogen (A-IN), Anhumas low Nitrogen (A-LN), Piracicaba ideal Nitrogen (P-IN) and Piracicaba low Nitrogen (P-LN). Error bars show standard deviations.

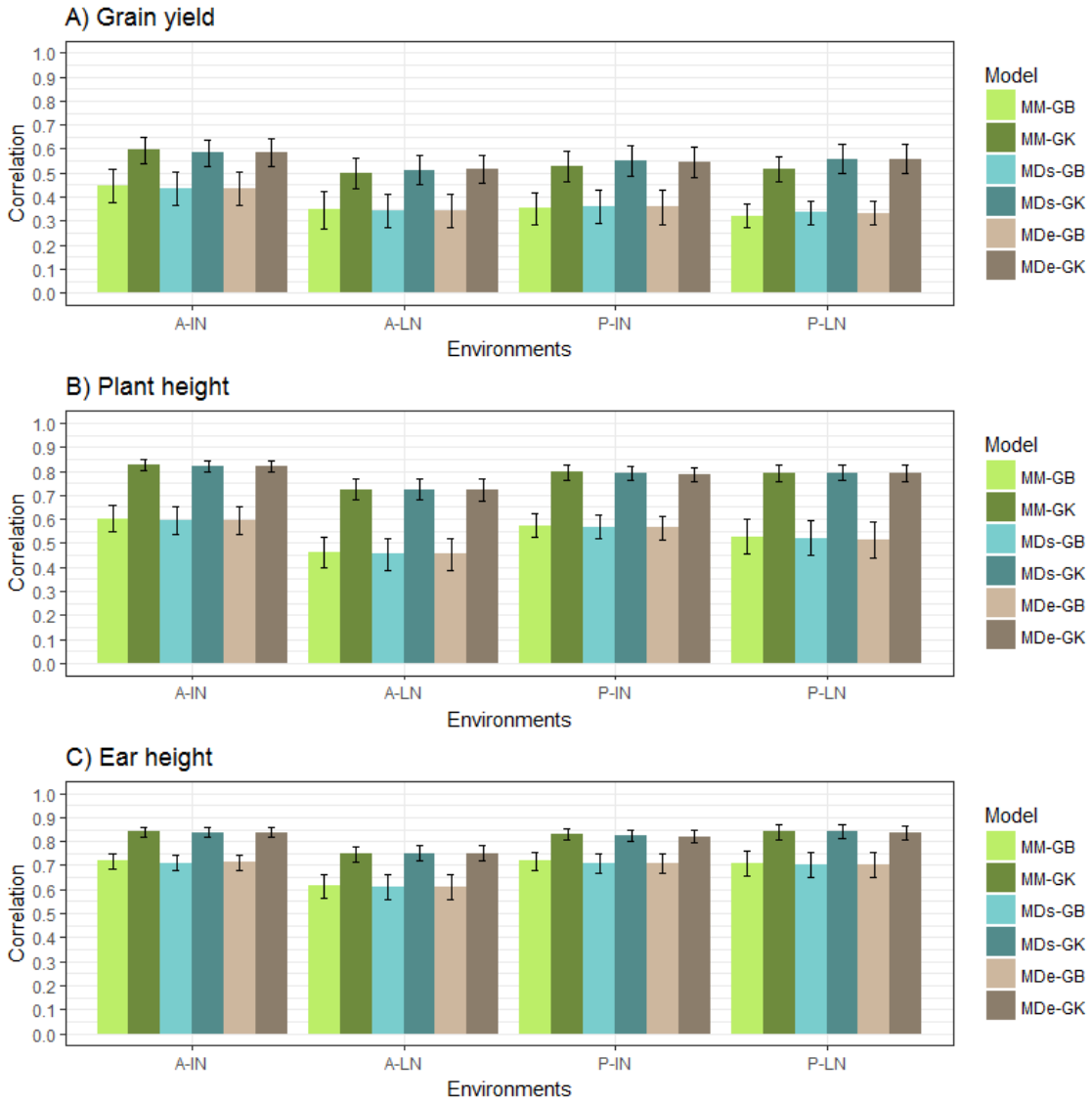


**Figure 5**. USP data set. Mean correlation between observed and predictive values (average of 50 random cross-validations partitions, CV2) for multi-environment, main genotypic effect model GBLUP kernel (MM-GB), multi-environment, main genotypic effect Gaussian kernel (MM-GK), multi-environment, single variance G×E deviation model GBLUP kernel (MDs-GB), multi-environment, single variance G×E deviation model Gaussian kernel (MDs-GK), multi-environment, environment-specific variance G×E deviation model GBLUP kernel (MDe-GB), multi-environment, environment-specific variance G×E deviation model Gaussian kernel (MDe-GK) in four environments (horizontal axis) for: (A) grain yield, (B) plant height and (C) ear height. Environments: Anhumas ideal Nitrogen (A-IN), Anhumas low Nitrogen (A-LN), Piracicaba ideal Nitrogen (P-IN) and Piracicaba low Nitrogen (P-LN). Error bars show standard deviations.

# APPENDICES

## Appendix C

In Box C1 we provide simplified scripts that can be used to obtain the GBLUP and Gaussian kernel matrix, to be used in the multi-environment models used in this study. The example uses the following R-object:

- X (n×p) a genomic marker matrix. The line IDs can be retrieved using either rownames(X) or SNP IDs in colnames(X).

```
X<-scale(X,center=TRUE,scale=TRUE)

# G-BLUP kernel (GB)
GB<-tcrossprod(X)/ncol(X)

# Gaussian Kernel (GK)
dist<-as.matrix(dist(X))^2
GK<-exp(-dist/median(dist))
```

**Box C1.** An example of how to obtain GBLUP (GB) and Gaussian Kernel (GK) matrix

In the examples used to fit MM, MDs and MDe models, it is necessary an R-object as matrix (pheno_geno) compound by three vectors of Location_id, Germplasm_id and Y, where:

- pheno_geno$Location_id vector is compound by environment information
- pheno_geno$Germplasm_id vector is compound by germplasm names
- pheno_geno$Y is a numeric and standardized vector with yield records in all environments

In Box C2 we give simplified scripts that can be used to fit models MM, MDs or MDe.

```
# Loading objects needed for MM, MDs or MDe models
 library(BGLR)
 env<-c(1,2,3,4)        ### Choose enviroment
K<-GK                   #### Choose kernel GK or GB
model<- "MDs"           ### Choose model "MM"or"MDs"or"MDe"
CV<- 2                  ### Choose CV=1 (CV1), CV=2 (CV2)
nEnv<-length(env)
pheno_geno1<-pheno_geno[pheno_geno$Location==env[1],]
i<-2
while ( i <=nEnv){
      pheno_geno1<-rbind(pheno_geno1,pheno_geno[pheno_geno$Location==env[i],])
      i<-i+1   }
envID<-as.factor(pheno_geno1[,1])
IDs<-as.character(unique(pheno_geno1$Germplasm_id))
IDy<-pheno_geno1$Germplasm_id
Y <- as.numeric(pheno_geno1[,3])
Y <- scale(Y,center=TRUE,scale=TRUE)
names(Y)<-IDy
nEnv <- length(env)
K<-K[IDs,IDs]
phenol_geno1$Germplasm_id<-factor(x=phenol_geno1$ Germplasm_id,levels=rownames(K),ordered=TRUE)
Zg <- model.matrix(~factor(pheno_geno1$Germplasm_id)-1)  # genotype design matrix
Ze <- model.matrix(~factor(pheno_geno1$Location)-1)  # environment design matrix
```

```
K1 <- Zg%*%K%*%t(Zg)
# Cross Validation scheme fit
set.seed(12345)
nFolds <- 5
partitions <- 10
if (CV == 1) {
  mfold <- matrix(NA, ncol = partitions, nrow = length(IDs))
  sets <- matrix(NA, nrow = nrow(pheno_geno1), ncol = partitions)
  for (s in 1:partitions) {
    mfold[, s] <- sample(1:nFolds, size = length(IDs), replace = TRUE)
    for (i in 1:nrow(pheno_geno1)) {
      sets[i, s] <- mfold[which(IDs == pheno_geno1$Germplasm_id[i]), s]    }}}
if (CV == 2) {
  sets <- matrix(NA, nrow = nrow(pheno_geno1), ncol = partitions)
  for (s in 1:partitions) {
    for (i in IDs) {
      tmp = which(IDy == i)
      ni = length(tmp)
      tmpFold <- sample(1:nFolds, size = ni, replace = ni > nFolds)
      sets[tmp, s] <- tmpFold   }}}
trn_tst<-matrix(nrow=nrow(pheno_geno1),ncol=nFolds,NA)
NA.y <- matrix(Y, ncol = nFolds, nrow = length(Y))
# Fitting  MM model
if (model=="MM") {
            ETA<-list(ENV=list(X=Ze,model="FIXED"),
                      Grm=list(K=K1,model="RKHS"))}
# Fitting MDs model
if(model=="MDs"){
          ZEZE<-tcrossprod(Ze)
          K2<-K1*ZEZE
          ETA<-list(ENV=list(X=Ze,model="FIXED"),
                    Grm=list(K=K1,model="RKHS"),
                    EGrm=list(K=K2,model="RKHS"))}
# Fitting MDe model
if (model=="MDe"){
             ETA<-list(ENV=list(X=Ze,model="FIXED"),
                       Grm=list(K=K1,model="RKHS"))
                       for(k in 1:nEnv){
                ZEE<-matrix(0,nrow=nrow(Ze),ncol=ncol(Ze))
                ZEE[,k]<-Ze[,k]
                ZEEZ<-(ZEE%*%t(Ze))
                K3<-K1*ZEEZ
                ETA[[k+2]]<-list(K=K3,model='RKHS')      }}
  COR<-matrix(0,nrow=nEnv,ncol=(nFolds*partitions))
for (s in 1:partitions) {
              NA.y <- matrix(Y, ncol = (nFolds), nrow = length(Y))
                 for(i in 1:ncol(trn_tst)){
                      trn_tst[,i]<-ifelse(sets[,s]==i,'tst','trn')
                      NA.y[,i][(trn_tst[,i]=="tst")]<-NA
                       A <-  as.matrix(NA.y[,i])
              fm <- BGLR(y=A,ETA=ETA,nIter=10000,burnIn=1000,thin=2)
                     for(g in 1:length(env)){
                          tst1 <- which(is.na(fm$y[envID==env[g]]))
               COR[g,((s-1)*nFolds+i)] <- cor(Y[envID==env[g]][tst1],fm$yHat[envID==env[g]][tst1])  }}}
```

**Box C2.** An example of how to fit MM model with cross-validation 2 design (CV2).

# Appendix D

**Table D1**. HEL data set. Mean correlation (for 50 random partitions, CV1) for models multi-environment, main genotypic effect model GBLUP kernel (MM-GB), multi-environment, main genotypic effect Gaussian kernel (MM-GK), multi-environment, single variance G×E deviation model GBLUP kernel (MDs-GB), multi-environment, single variance G×E deviation model Gaussian kernel (MDs-GK), multi-environment, environment-specific variance G×E deviation model GBLUP kernel (MDe-GB), multi-environment, environment-specific variance G×E deviation model Gaussian kernel (MDe-GK) for three traits, grain yield, plant height, and ear height (standard deviation in parentheses)

| Model / Environment* | Multi-environment, main genotypic effect model (MM) | | % Change | Multi-environment, main, single variance G×E deviation model (MDs) | | % Change | Multi-environment, environment-specific variance G×E deviation model (MDe) | | % Change |
|---|---|---|---|---|---|---|---|---|---|
| | GB | GK | | GB | GK | | GB | GK | |
| Grain yield | | | | | | | | | |
| IP | 0.48 (0.12) | 0.60 (0.08) | 25 | 0.65 (0.09) | 0.75 (0.05) | 15 | 0.64 (0.09) | 0.75 (0.05) | 17 |
| NM | 0.36 (0.10) | 0.48 (0.09) | 33 | 0.41 (0.08) | 0.55 (0.07) | 34 | 0.41 (0.08) | 0.54 (0.07) | 32 |
| PM | 0.60 (0.06) | 0.70 (0.05) | 17 | 0.68 (0.04) | 0.75 (0.04) | 10 | 0.68 (0.04) | 0.75 (0.04) | 10 |
| SE | 0.17 (0.13) | 0.28 (0.11) | 65 | 0.32 (0.11) | 0.52 (0.08) | 63 | 0.31 (0.11) | 0.51 (0.09) | 65 |
| SO | 0.27 (0.11) | 0.42 (0.10) | 56 | 0.39 (0.08) | 0.56 (0.07) | 44 | 0.38 (0.08) | 0.56 (0.07) | 47 |
| Plant height | | | | | | | | | |
| IP | 0.72 (0.07) | 0.74 (0.06) | 3 | 0.73 (0.06) | 0.75 (0.06) | 3 | 0.73 (0.06) | 0.75 (0.06) | 3 |
| PM | 0.75 (0.06) | 0.78 (0.05) | 4 | 0.76 (0.06) | 0.79 (0.05) | 4 | 0.76 (0.06) | 0.79 (0.05) | 4 |
| SE | 0.74 (0.06) | 0.76 (0.05) | 3 | 0.75 (0.06) | 0.77 (0.05) | 3 | 0.75 (0.06) | 0.77 (0.05) | 3 |
| Ear height | | | | | | | | | |
| IP | 0.71 (0.06) | 0.72 (0.06) | 1 | 0.72 (0.06) | 0.73 (0.06) | 1 | 0.72 (0.06) | 0.72 (0.06) | 0 |
| PM | 0.74 (0.06) | 0.76 (0.06) | 3 | 0.76 (0.06) | 0.78 (0.05) | 3 | 0.76 (0.06) | 0.78 (0.05) | 3 |
| SE | 0.72 (0.06) | 0.73 (0.05) | 1 | 0.72 (0.06) | 0.73 (0.05) | 1 | 0.72 (0.06) | 0.73 (0.05) | 1 |

*Environments: Ipiaçú (IP), Nova Mutum (NM), Pato de Minas (PM), Sertanópolis (SE) and Sorriso (SO).

**Table D2**. USP data set. Mean correlation (for 50 random partitions, CV1) for models multi-environment, main genotypic effect model GBLUP kernel (MM-GB), multi-environment, main genotypic effect Gaussian kernel (MM-GK), multi-environment, single variance G×E deviation model GBLUP kernel (MDs-GB), multi-environment, single variance G×E deviation model Gaussian kernel (MDs-GK), multi-environment, environment-specific variance G×E deviation model GBLUP kernel (MDe-GB), multi-environment, environment-specific variance G×E deviation model Gaussian kernel (MDe-GK) for three traits, grain yield, plant height, and ear height (standard deviation in parentheses)

| Model / Environment* | Multi-environment, main genotypic effect model (MM) | | % Change | Multi-environment, main, single variance G×E deviation model (MDs) | | % Change | Multi-environment, environment-specific variance G×E deviation model (MDe) | | % Change |
|---|---|---|---|---|---|---|---|---|---|
| | GB | GK | | GB | GK | | GB | GK | |
| Grain yield | | | | | | | | | |
| P-LN | 0.26 (0.06) | 0.27 (0.06) | 4 | 0.29 (0.07) | 0.29 (0.07) | 0 | 0.29 (0.07) | 0.28 (0.07) | -3 |
| P-IN | 0.31 (0.07) | 0.30 (0.06) | -3 | 0.34 (0.07) | 0.33 (0.06) | -3 | 0.34 (0.07) | 0.33 (0.06) | -3 |
| A-LN | 0.30 (0.08) | 0.28 (0.07) | -7 | 0.31 (0.07) | 0.27 (0.07) | -13 | 0.31 (0.07) | 0.28 (0.07) | -10 |
| A-IN | 0.40 (0.06) | 0.39 (0.06) | -3 | 0.42 (0.05) | 0.41 (0.05) | -2 | 0.42 (0.05) | 0.41 (0.05) | -2 |
| Plant height | | | | | | | | | |
| P-LN | 0.47 (0.06) | 0.45 (0.06) | -4 | 0.47 (0.06) | 0.44 (0.06) | -6 | 0.47 (0.06) | 0.44 (0.06) | -6 |
| P-IN | 0.52 (0.06) | 0.48 (0.06) | -8 | 0.52 (0.06) | 0.49 (0.06) | -6 | 0.52 (0.06) | 0.49 (0.06) | -6 |
| A-LN | 0.40 (0.07) | 0.41 (0.07) | 3 | 0.40 (0.07) | 0.40 (0.06) | 0 | 0.40 (0.07) | 0.40 (0.06) | 0 |
| A-IN | 0.55 (0.05) | 0.51 (0.05) | -7 | 0.55 (0.05) | 0.52 (0.05) | -5 | 0.55 (0.05) | 0.52 (0.05) | -5 |
| Ear height | | | | | | | | | |
| P-LN | 0.68 (0.04) | 0.67 (0.04) | -1 | 0.68 (0.04) | 0.67 (0.04) | -1 | 0.68 (0.04) | 0.66 (0.04) | -3 |
| P-IN | 0.69 (0.04) | 0.68 (0.04) | -1 | 0.69 (0.04) | 0.68 (0.04) | -1 | 0.69 (0.04) | 0.68 (0.04) | -1 |
| A-LN | 0.59 (0.06) | 0.59 (0.05) | 0 | 0.59 (0.06) | 0.58 (0.05) | -2 | 0.59 (0.06) | 0.58 (0.05) | -2 |
| A-IN | 0.69 (0.03) | 0.68 (0.03) | -1 | 0.69 (0.03) | 0.68 (0.03) | -1 | 0.69 (0.03) | 0.68 (0.03) | -1 |

*Environments: Anhumas ideal N (A-IN), Anhumas low N (A-LN), Piracicaba ideal N (P-IN) and Piracicaba low N (P-LN).

# 4. GENERAL CONCLUSION

From with a high-density panel, it is possible to select the most informative markers to improve accuracy and build a low-cost array to implement genomic selection in breeding programs.

The best strategy to obtain markers subsets is the re-estimation of the marker effect, which increases the accuracy and reduces the bias.

The use of Gaussian kernel and the including G×E effect, there is an increase in the accuracy of the genomic prediction models.