Euphytica (2019) 215:18
https://doi.org/10.1007/s10681-019-2339-z

REVIEW

Increasing accuracy and reducing costs of genomic prediction by marker selection

Massaine Bandeira e Sousa · Giovanni Galli · Danilo Hottis Lyra · Ítalo Stefanini Correia Granato · Filipe Inácio Matias · Filipe Couto Alves · Roberto Fritsche-Neto

Received: 17 February 2018 / Accepted: 3 January 2019 / Published online: 15 January 2019

Abstract Genotyping costs can be reduced without decreasing the genomic selection accuracy through methodologies of markers subsets assortment. Thus, we compared two strategies to obtain markers subsets. The former uses the primary and the latter the re-estimated markers effects. Moreover, we analyzed each subset via prediction accuracy, bias, and relative efficiency by main genotypic effect model (MGE) fitted, using genomic best linear unbiased predictor linear kernel (GB), and Gaussian nonlinear kernel (GK). All scenarios (subset of markers × kernels models) were applied to a public dataset of rice diversity panel (RICE) and two hybrids maize datasets (HEL and USP). The highest prediction accuracies were obtained by MGE-GB and MGE-GK for grain yield and plant height when we decrease the number of markers. Overall, marker subsets via re-estimated effects method showed a higher relative efficiency of genomic selection. Based on a high-density panel, we can conclude that it is possible to select the most informative markers in order to improve accuracy and build a low-cost SNP chip to implement genomic selection in breeding programs. In addition, we recommend REE (re-estimated effect) strategies to find markers subsets in training population, increasing accuracy of genomic selection.

Keywords SNP array subset · Relative efficiency · Reliability · Model-kernel

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). Genomic estimated breeding values (GEBV) are calculated based on the effects of these marker. Many statistical methods have been used as alternative to estimate the effect of markers on the training population. Ridge regression best linear unbiased prediction (RR-BLUP) (Meuwissen et al. 2001) was the first to be adopted. This method assumes that all effects of the markers are normally distributed and have equal variance (Meuwissen et al. 2001). Based on the same assumption, the genomic best linear unbiased prediction (GBLUP) model is the most commonly used method. This model uses genetic marker information to compute associations between individuals based on

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10681-019-2339-z) contains supplementary material, which is available to authorized users.

M. B. e Sousa (✉) · G. Galli · D. H. Lyra · Í. S. C. Granato · F. I. Matias · F. C. Alves · R. Fritsche-Neto
Department of Genetics, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, São Paulo, Brazil
e-mail: massainebandeira@hotmail.com

the Genomic Relationship Matrix (GRM, G) (Habier et al. 2007). Although GBLUP has shown satisfactory results, some authors suggest modifications in order to improve prediction accuracy. Besides the linear GBLUP kernel, Gianola et al. (2006) proposed semi-parametric methods to model the relationship between phenotype and markers in a GP context. One of them is the reproducing kernel Hilbert space (RKHS), which is 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 the use of RKHS methods to find genetic relationship between individuals has the potential to improve prediction accuracy, mainly when genotype is added to the models given the environment interaction effect (Cuevas et al. 2016a, b; e Souza et al. 2017).

Currently, high-density SNP chips provide the best genome coverage; however, implementing genomic selection commonly demands huge population genotyping for both training individuals and selection candidates, which may increase the total program cost. On the other hand, genotyping by sequencing (GBS) is an affordable alternative, despite the relatively high rates of missing data shown by the method. The quality of genotypes tends to be lower in this method, since it depends on the in-depth reading of the genome sequence (Gorjanc et al. 2015b). Another approach is linked to the development of low density and cost-fixed arrays, which are highly efficient in smaller breeding programs (Spindel et al. 2015). Advantages of fixed arrays include robust allele calling, cost-effectiveness per data point and speed of genotyping turn-around (Thomson 2014). Marker density has been investigated in some studies, although they have shown divergent results about the use of sets of data of lower or higher density SNP (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). Higher genotyping density does not always improve accuracy and markers subsets sometimes outperform the entire dataset (Zhang et al. 2010; Ma et al. 2016).

Accordingly, propositions point out different methodologies to find markers subsets, such as random (markers 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). Overall, the use of selected markers subsets based on their effects, or on positions, has provided an efficient strategy for accuracy improvement (Vazquez et al. 2010; Resende et al. 2012; Zhang et al. 2015). Tayeh et al. (2015) decreased the number of markers from 9824 to 2945 by retaining a single marker per unique map position. These authors did not find reduced prediction accuracy in any of the evaluated traits. Based on Ma et al. (2016), the marker preselection based on haplotype block analysis is an interesting option to reduce costs with the implementation of genomic selection.

Accordingly, genotyping training-individuals with a high-density panel and evenly spaced selection-candidates, is another approach for the cost-effective implementation of genomic selection in plant breeding programs. However, it is necessary to have a specific SNP set for each targeted trait, because the selected SNPs can vary depending on the trait. In addition, the efficiency of a trait-specific low-density SNP chip depends on the linkage disequilibrium between SNPs presenting significant estimated effects and the true causative loci affecting the trait of interest (Wu et al. 2016). Thus, we compared two strategies in order to find markers subsets. The first is based on the original effect of the marker, which was obtained in the first estimate by using the entire set of data (original effect—ORI strategy). The second is based on the re-estimated effects of markers that are re-estimated based on dataset reductions (re-estimated effect—REE strategy). Thus, the aims of our study were (1) to compare two strategies in order to find markers subsets based on their effects—original and re-estimated effects; (2) to compare prediction accuracy, bias and the relative efficiency of a main genotypic effect (MGE) model adjusted both through the linear kernel genomic best linear unbiased predictor, GBLUP (GB) and the nonlinear kernel Gaussian kernel (GK) by using markers subsets. The training dataset was used in traits with complex genetic architecture in three different datasets.

Materials and methods

Phenotypic data

Rice and two maize datasets were taken into consideration in this study by using GY (grain yield, ton ha^{-1}) and PH (plant height, cm): (1) RICE dataset is

available at the Rice Diversity platform (http://www.ricediversity.org). We used 270 elite breeding lines (F6–F7) from the International Rice Research Institute (IRRI) irrigated rice breeding program. These lines were evaluated in a single location in Los Baños, Philippines, for 3 years (2009 to 2011), during the dry season (Spindel et al. 2015). (2) HEL dataset was provided by Helix Sementes® Company, São Paulo, Brazil. HEL consisted of 452 maize hybrids obtained through the crossing of 111 inbred lines in partial diallel. The experimental trial was carried out at five sites located in Southern, Southeastern and Midwestern Brazil, during the first 2014/15 growing season. The study followed a completely randomized block design with two replications per genotype and environment. (3) USP dataset regarded maize hybrids and was provided by University of São Paulo (USP). It consisted of 739 maize hybrids obtained through the crossing of 49 inbred lines in partial diallel. The hybrids were evaluated at Piracicaba and Anhumas, São Paulo State, Brazil, in 2016. The evaluation was based on an augmented block design, using two commercial hybrids as checks. Hybrids were evaluated under ideal nitrogen (N) level (100 kg N ha^{-1}) in both sites. Each plot had 7 m, on average, with 0.50 m spacing between rows and 0.33 m between plants—there was phenotypic imbalance in both maize datasets.

Genotypic data

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

Statistical models

Estimating BLUPs

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{r} + \mathbf{T}\boldsymbol{g} + \mathbf{B}\boldsymbol{s} + \mathbf{H}\boldsymbol{x} + \boldsymbol{\varepsilon} \quad (1)$$

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

Random effect significance was estimated through the Likelihood Ratio Test (LRT) (Gilmour et al. 2009). Variance components estimated for the effect of each model were used to calculate entry-mean based heritability (H^2) by using the mean of each environment (by taking into consideration each year, or site, as environment).

Obtaining markers subsets

We used the RR-BLUP (random regression—best linear unbiased predictor) method (Meuwissen et al. 2001) to find 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 the matrix of markers, $\boldsymbol{\beta}$ is a vector of the effects of the marker, and $\boldsymbol{\varepsilon}$ is the residual effect. The following distribution was assumed: $\boldsymbol{\beta} \sim N\left(0, \boldsymbol{I}\sigma_\beta^2\right)$ and $\varepsilon \sim N\left(0, \boldsymbol{I}\sigma_\varepsilon^2\right)$.

We applied two strategies to find markers subsets: by reducing the number of markers based on (1) their original effect (ORI strategy) and (2) re-estimation the effects of the markers (REE strategy). We found a subset of 10,000 markers for each strategy, and subsequently evaluated five subsets (5000, 2500, 1000, 500, and 100 SNPs) for each strategy. For ORI strategy, the vector of the effect of the markers was first estimated by using all the markers and ranked in decreasing order based on their absolute values to select the 10,000 SNPs recording the highest absolute value. The 5000 SNPs presenting the highest absolute values were selected. This step was repeated until a subset with 100 SNPs was found. The subset containing 10,000 SNPs was subjected to the REE strategy, similar to the ORI strategy. However, the effects of the markers were re-estimated and ranked in decreasing order based on their absolute values. We selected the 5000 SNPs recording the highest absolute values. The procedure was repeated until a subset with 100 SNPs was found.

Genomic prediction

The five markers subsets of each method described above were used in genomic prediction based on the main genotypic effect model (MGE). This model fits data of each subset separately and considers the main effect of the genotypes. The model is written in matrix notation as follows:

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

where $\boldsymbol{y} = (y_1, \ldots, y_k)'$ is the response vector, and y_i represents the observations in the ith genotype $(i = 1, \ldots, k)$ in each subset of markers; $\boldsymbol{\mu}$ is the general mean; $\boldsymbol{Z}_u$ is the incidence matrix that connects the random genetic effects to the phenotypes; $\boldsymbol{u}$ is the random genetic effects, and $\boldsymbol{\varepsilon}$ are residual random effects. The MGE model (3) assumes that the distribution of the $\boldsymbol{u}$ vector is multivariate normal with mean zero and covariance matrix $\boldsymbol{K}$, with $\boldsymbol{u} \sim N(0, \sigma_u^2\boldsymbol{K})$, where σ_u^2 is the genetic variance component of $\boldsymbol{u}$, and $\boldsymbol{K}$ is a symmetric positive semi-definite matrix that denotes the variance–covariance of the genetic values constructed from the genomic molecular markers by subsets. The error $\boldsymbol{\varepsilon}$ was independent of from each other and normally distributed $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(0, \sigma_\varepsilon^2\boldsymbol{I})$. 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 effect that approximates the errors.

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

Main genotypic effect model with GBLUP (MGE-GB) in the MGE model (3), 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 the molecular markers of the individuals in the 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 individuals ith and i'th$(i = 1, \ldots, n_j)$, given by the markers; $h > 0$ is the bandwidth parameter controlling the decay rate of $\boldsymbol{K}$ values (Pérez-Rodríguez et al. 2012; Cuevas et al. 2016a, b). 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, were used as scaling factor (Crossa et al. 2010).

Variance components, prediction accuracy, and bias parameters

The MGE model adjusted through the GB and GK methods was used for the entire RICE, HEL, and USP datasets in all the traits. Phenotypic data were centered and standardized before the model was adjusted, i.e., each phenotypic data point was centered by subtracting the overall mean and, subsequently, standardized by dividing it by the sample standard deviation). These analyses were performed to derive variance component and genomic heritability estimates. The summation of variance components in a specific model would approximate to 1, since the collected data was standardized.

Prediction accuracy (PA) was performed in separate for each trait and subset of markers (all markers and the five markers subsets) based on MGE-GB and MGE-GK models. PA was assessed using thirty random partitions (repetitions)—80% of the hybrids/lines composed the training set (TRN) and the remaining 20% composed the testing set (TST). All the parameters in the models, including variance components

resulting from residual effects and genetic effects, were re-estimated based on TRN data in each of the TRN–TST partitions in validation procedure.

Models were adjusted to the set of TRN data for each TRN–TST partition and prediction accuracy was assessed by computing the Pearson's correlation between predictions and BLUPs in the set of TST data. The same TRN–TST partitions were used to evaluate the prediction accuracy of each model; thus, thirty correlations of each model and trait were computed.

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

Relative efficiency of genomic prediction

The relative efficiency of the genomic prediction of all proposed scenarios was compared to the traditional phenotypic selection. Two parameters were used in this calculation: (1) the relative efficiency per cycle (RE_c), which was calculated through $r/\sqrt{H^2}$, where r is model accuracy and H is the phenotypic heritability of the trait (Hoffstetter et al. 2016b); (2) selection coincidence was estimated based on the top 10% and 20% individuals identified through TRN–TST and on individuals recording the best phenotypic performance of each trait.

Softwares

We used the rrBLUP-R package (Endelman 2011) to run the RR-BLUP model. The BGLR package (de los Campos and Perez-Rodriguez 2016) was used to carry out the MGE models based on 30,000 MCMC iterations, 5000 for burn-in and 5 for thinning.

Results

Phenotypic analysis

There was genetic variability in plant height (PH) and grain yield (GY) in the RICE, HEL and USP dataset according to the likelihood ratio test based on joint analysis (Table S1). Similarly, the random effects of Genotype × Year (RICE) and Genotype × Site (HEL and USP) interactions were significant, except for PH in the set of USP data. Such significant effect suggests the differential genotype performance between sites, or years, in 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) in the USP dataset (Additional Table S1), which reflected good accuracy in the phenotypic evaluation.

Variance components of markers subsets

Variance components (residual and genetic) and genomic heritability were obtained through the MGE model based on GBLUP and GK kernels (MGE-GB and MGE-GK) (Fig. 1).

RICE dataset

The estimated residual variance components were smaller than those found through the ORI strategy by using markers subsets found through REE strategy in the PH and GY trait (Fig. 1a, c). Based on results of MGE-GK and MGE-GB models, the use of small markers subsets induced great reduction in estimated residual variance. The variance component of genetic effects of each subset increased when the MGE-GK model was used, rather than the MGE-GB model (Fig. 1b, d). Furthermore, genetic effects from on the ORI strategy were slightly higher than the ones from the REE strategy.

HEL and USP datasets

PH and GY results of residual variance showed that MGE-GK tends to better adjust to the data than MGE-

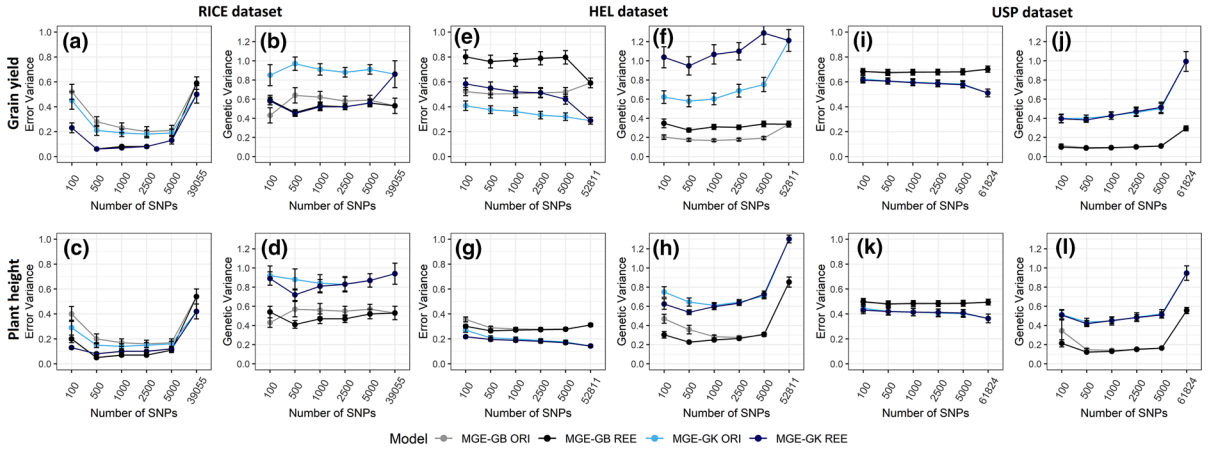

Fig. 1 Changes in proportions of estimated genetic and residual variances in MGE model, with the increase of SNPs number of SNPs (horizontal axis). Markers subsets obtained by primary effect (ORI) and re-estimated effect (REE) strategies for grain yield (**a**, **b**) and plant height (**c**, **d**) in RICE dataset, grain yield (**e**, **f**) and plant height (**g**, **h**) in HEL dataset and grain yield (**i**, **j**) and plant height (**k**, **l**) in USP dataset

GB. Overall, markers subset obtained through ORI strategy and adjusted to the MGE-GK model showed slight increase in residual variance in comparison to the use of all markers available (ranged from 0.29 to 0.41) for GY trait (Fig. 1e, g). The genetic variance of GY in MGE-GB (ranged from 0.17 to 0.34—ORI, and from 0.28 to 0.35—REE) was always lower than in MGE-GK (ranged from 1.21 to 0.60—ORI, and from 1.29 to 0.95—REE) (Fig. 1f). To PH trait, genetic variance was higher, approximately 1.30 in MGE-GK and 0.85 in MGE-GB when all markers were used (Fig. 1h). However, the genetic variances were always higher (higher than 0.53) than the ones found through MGE-GB (range from 0.23 to 0.36) when the markers subsets were measured through MGE-GK. Differences between prediction scenarios result from the kernel used to obtain GRM in the USP dataset. For GY trait, MGE-GB provided higher values of residual variance than MGE-GK (Fig. 1i). Genetic variance was lower in MGE-GB than through MGE-GK (Fig. 1j). When all markers were used, values were higher than using markers subsets. For PH residual and genetic variance followed the same trend of GY (Fig. 1k, l).

Prediction accuracy and bias of GEBV

RICE dataset

Overall, based on the results, prediction accuracy (PA) based on markers subsets was better than results recorded when all markers were used, mainly in PH (Fig. 2). As for GY, PA values using all markers was 0.30 through MGE-GB and 0.29 through MGE-GK. However, when the markers subsets were used in the predictions, PA values were similar, mainly with subsets presenting 2500 and 5000 SNPs. There was 7% increase when 5000 markers were used in the MGE-GK models. This subset recorded the best PA and was found through the application of the REE strategy. Overall, we verified that REE was better than ORI by comparing the two approaches. PA measured through MGE-GB ranged from 0.18 (100 markers) to 0.28 (5000 markers) when it was measured by means of ORI strategy and from 0.19 (100 markers) to 0.31 (5000 markers) in REE strategy-based measurements. Overall, GK kernel was slightly better than GB kernel in the MGE model. PA was moderate when all markers were used to calculate PH, whereas it was 0.38 in MGE-GB model and 0.41 in MGE-GK (Fig. 2). There was 24% and 19% PA increase when a subset of markers was used in the MGE-GB and MGE-GK models, respectively. The subset of markers recording the best PA (0.47 through MGE-GB and 0.49 through

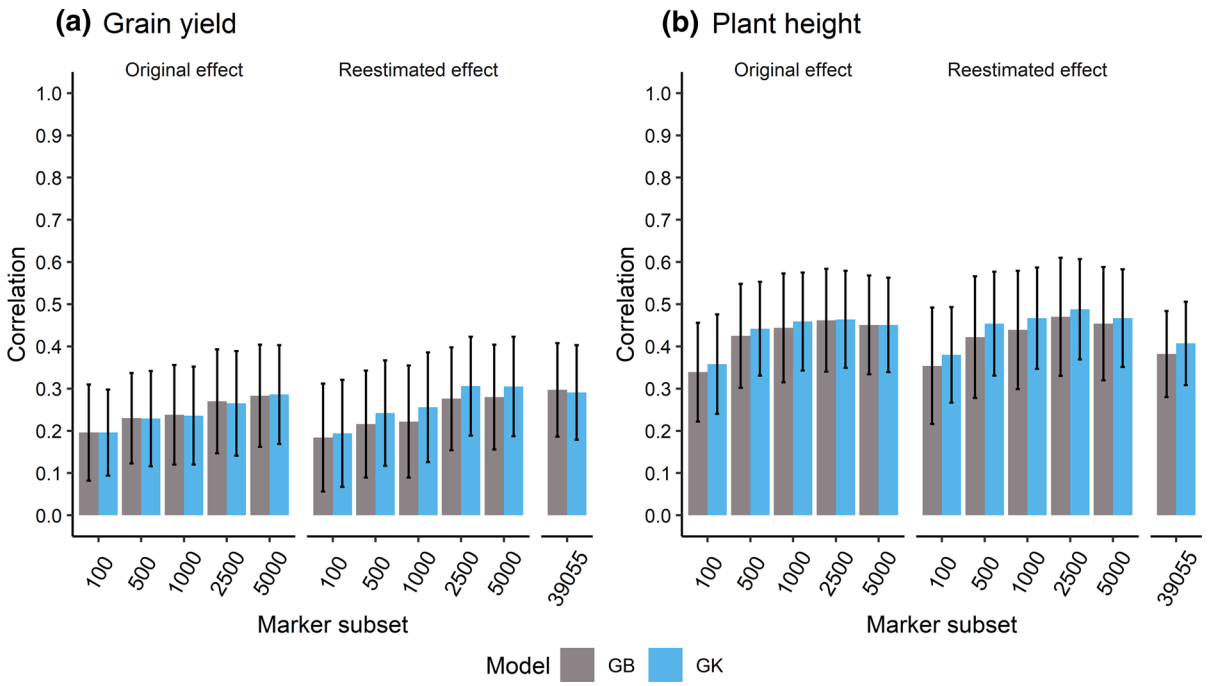

Fig. 2 RICE dataset. Mean correlation between observed and predictive values (mean of 30 random TRN–TST repetitions) estimated through MGE-GB and MGE-GK models by taking into consideration markers subsets (horizontal axis) found through original effect (ORI) and re-estimated effect (REE) strategies applied to grain yield and plant height

MGE-GK) was the one presenting 2500 markers found through the REE strategy.

Phenotypic coefficient of regression (slope) in GEBVs was calculated to measure the bias of each subset of strategies and models (Table S2). When $\beta = 1$, it is possible assuming no bias in the prediction. When ORI or REE approaches and MGE-GB or MGE-GK models were compared for GY, slopes were close to 0.50—this number indicates consistent bias in predictions when markers subsets are used. β values derived from all scenarios, since PH was equal to 0.70, on average. Overall, the fewer the markers, the bigger the bias. The use of subsets with fewer markers resulted in slightly higher reliability in both traits (Table S2).

HEL and USP datasets

GY and PH results indicated that the use of a subset of markers showed similar accuracy to calculations using all markers (Fig. 3). PA estimated with all markers was intermediate in MGE-GB (0.55) and MGE-GK (0.67) for GY and higher in MGE-GB (0.78) and MGE-GK (0.81) for PH. PA was 0.55 in MGE-GB and 0.65 in MGE-GK when 5000 markers were used; this number represents improvement by 15%. REE was similar to ORI when the two strategies were compared. Coefficient of regression (slope) in all scenarios, mainly in PH, was close to one, thus indicating no consistent bias in the prediction. Overall, model reliability in GY showed that MGE-GK (0.79 to 0.86) recorded higher values than MGE-GB (0.66 to 0.809) (Table S2). On the other hand, for PH, all scenarios presented similar reliability (> 0.80).

For USP dataset, all scenarios in both traits presented similar accuracies (Fig. 4). The slope of GY and PH traits, in both models and subsets, was close to one, thus indicating no consistent bias in the prediction (Table S2).

Relative efficiency of genomic prediction

RICE dataset

We observed similar coincidence of selection (CS) when markers subsets were used by taking into consideration 10% intensity of selection (IS) (Table S2 and Fig. 5a, b). CS values in GY obtained

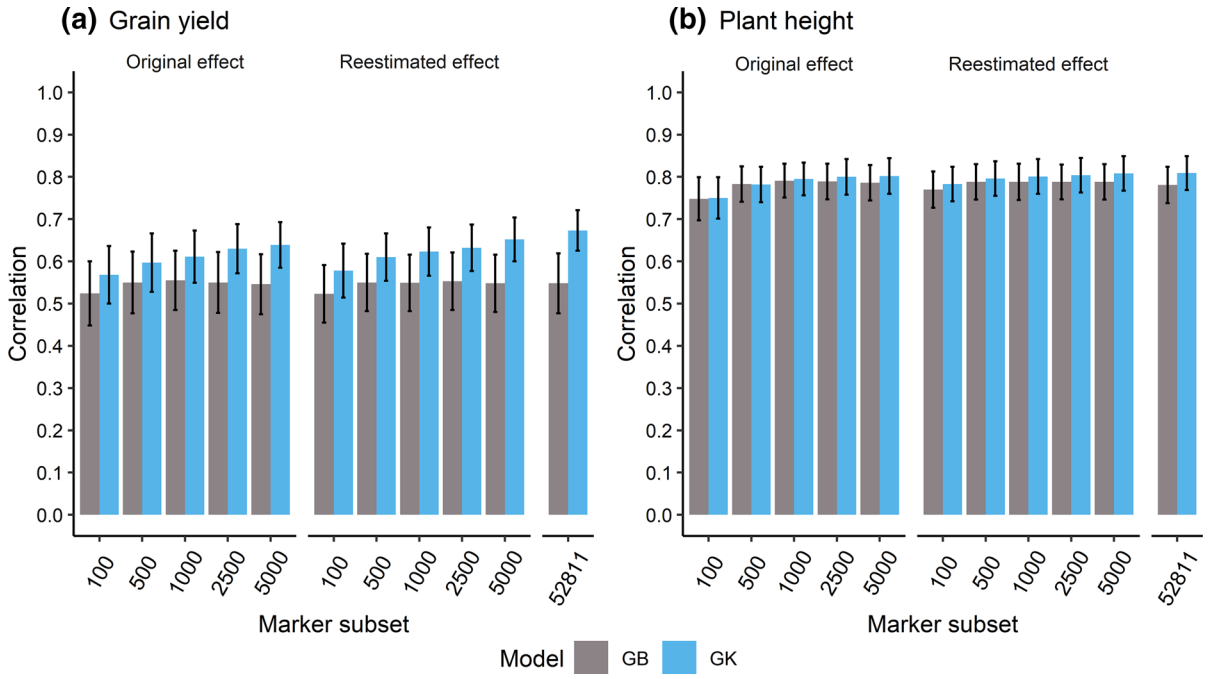

Fig. 3 HEL dataset. Mean correlation between observed and predictive values (mean of 30 random TRN–TST repetitions) estimated through MGE-GB and MGE-GK models by taking into consideration markers subsets (horizontal axis) found through original effect (ORI) and re-estimated effect (REE) strategies applied to grain yield and plant height

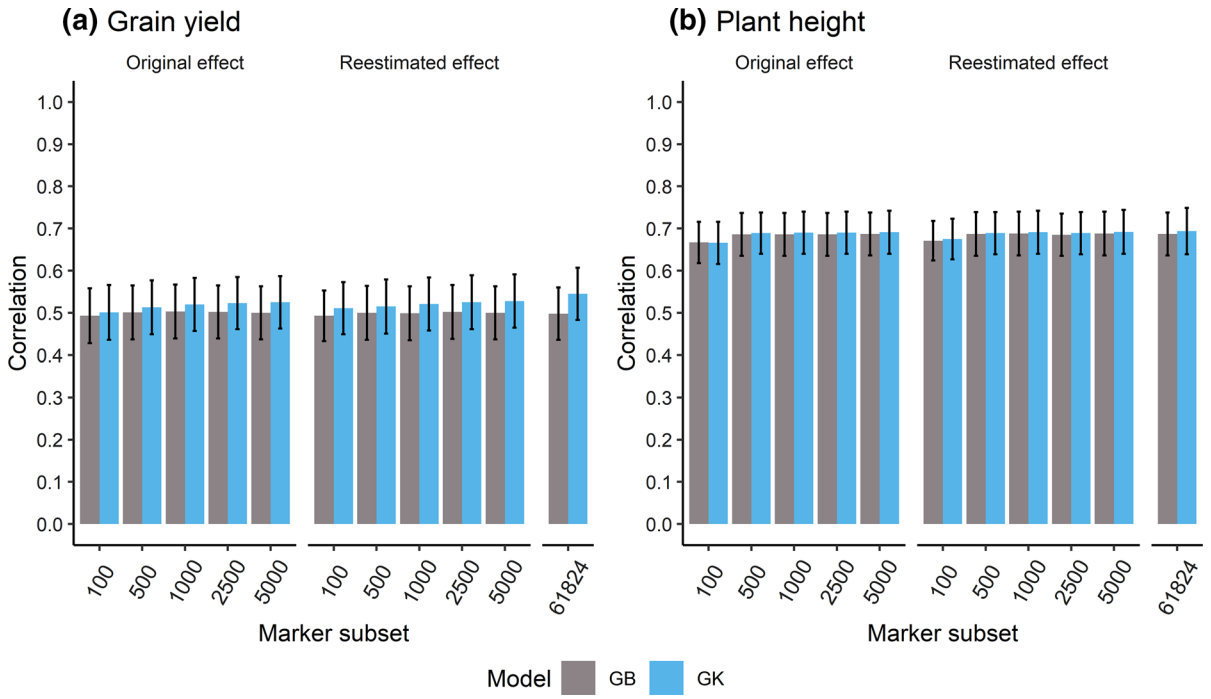

Fig. 4 USP dataset. Mean correlation between observed and predictive values (mean of 30 random TRN–TST repetitions) estimated through MGE-GB and MGE-GK models by taking into consideration markers subsets (horizontal axis) found through original effect (ORI) and re-estimated effect (REE) strategies applied to grain yield and plant height

through the MGE-GB model ranged from 14% (100 markers) to 26% (all markers) and from 15% (100 markers) to 29% (all markers) in the MGE-GK. The REE strategy selects markers that provide greater coincidence between genomic and phenotypic selection. CS under 20% IS showed the same trend of that under 10% IS. However, values were slightly higher when the MGE-GB model was used. MGE-GK using 5000 (ORI), 5000 and 2500 (REE) markers recorded values equal to, or higher than, 0.37 (all markers) for the relative efficiency (REc) per cycle to GY, thus showing genomic prediction efficiency over the phenotypic selection. Overall, PH using subsets showed better performance in Cs10, Cs20 and Rec than predictions using all markers.

HEL and USP datasets

There was similar result in Cs10 and Cs20 when markers subsets were used (Table S2 and Fig. 5c, d). PH values were close to 40% in Cs10 and values based on 20% IS were slightly higher than those selected

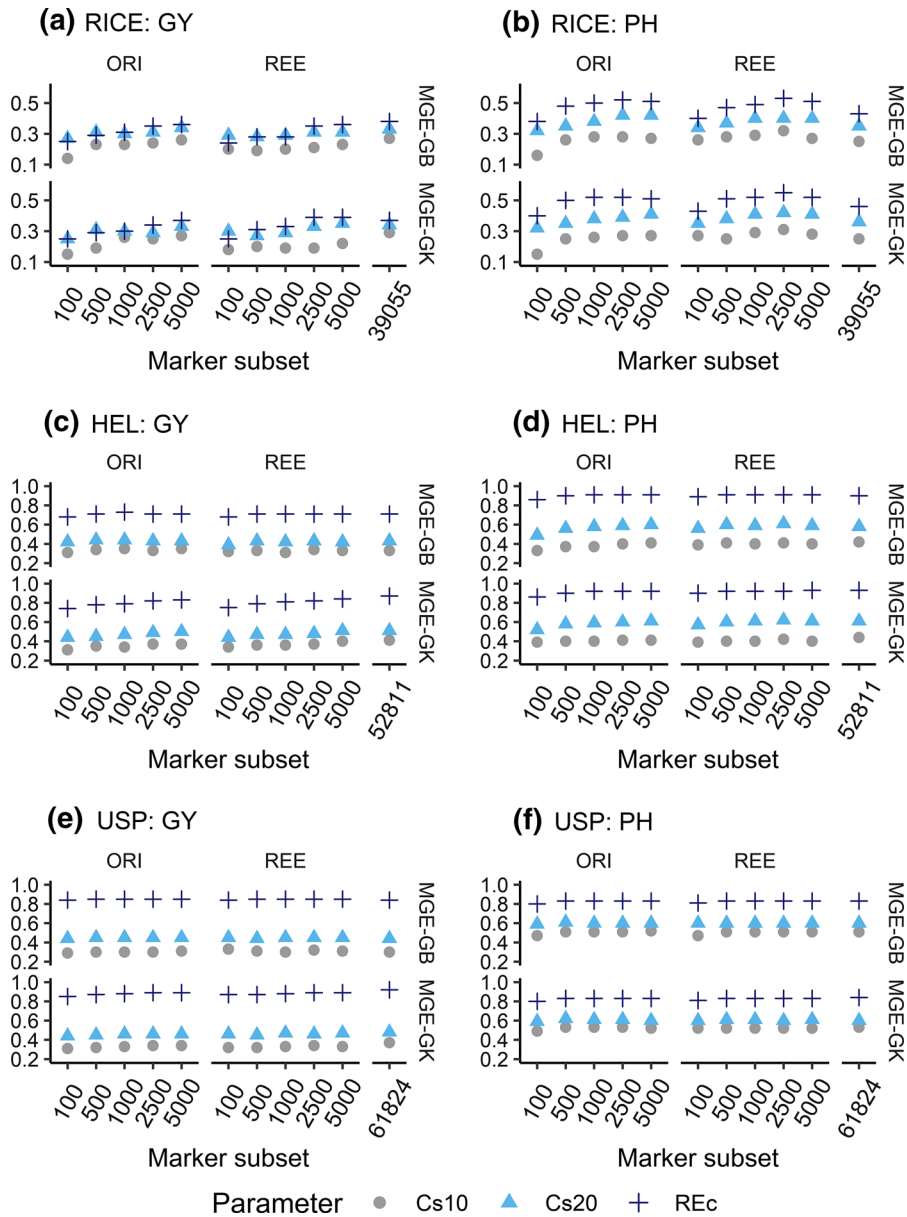

Fig. 5 Coincidence of selection at 10% (Cs10) and 20% (Cs20) selection intensity and relative efficiency (RE_c) applied to a genomic prediction cycle through MGE-GB and MGE-GK models, by taking into consideration a subset of markers found through original effect (ORI) and re-estimated effect (REE) strategies applied to plant height and grain yield in RICE, HEL and USP datasets

with values close to 59%. Values were lower than 1 in PH and GY relative efficiency (REc) per cycle. Such outcome shows that, in these cases, genomic prediction efficiency did not exceed the phenotypic selection used (Table S2 and Fig. 5c, d).

The results of the USP dataset showed the same trend presented by the HEL dataset (Table S2 and Fig. 5c, d). MGE-GK generated higher CS values than MGE-GB. There was higher coincidence via MGE-GK under 20% IS. On the other hand, PH results evidenced that selection coincidences were similar among models, methods and selection intensities. PH (0.80 to 0.84) and GY (0.84 to 0.92) values were lower than 1 when the relative efficiency (REc) per cycle was taken into account, thus showing that, in these cases, genomic prediction efficiency did not exceed the phenotypic selection (Table S2 and Fig. 5c, d).

Discussion

Low-density marker panels are desirable to implement genomic selection (GS) in breeding programs in order to reduce genotyping costs. Several studies have documented the benefits of using markers subsets to improve prediction accuracy in genomic selection analyses (Moser et al. 2010; Resende et al. 2012; Szyda et al. 2013; Spindel et al. 2015; Tayeh et al. 2015; Hoffstetter et al. 2016a; Li et al. 2018). Two strategies based on the effect of the markers were used to find markers subsets in training individuals in our study. Similarly, Resende et al. (2012) also used the effect of re-estimating markers and found better results of rust disease resistance and wood density traits in Loblolly pine than the ones found through Bayesian methods Prediction accuracy, using each subset and all markers, was compared through a main genetic effect (MGE) model using GB (GBLUP) and GK (Gaussian) kernels. Strategies to select subsets of SNPs recorded similar accuracy values and bias than using the whole SNP panel. According to Porto-Neto et al. (2015), SNP selection would take into account different marker allele frequencies between individuals; the missing heritability can be largely recovered and lead to improved accuracy. Overall, many sequences of variants associated with complex traits have small effects and low repeatability (Bian and Holland 2017). Strategies adopted to reduce the number of markers allow increasing accuracy duo eliminating redundant markers with small effect. This procedure allows reducing the multicollinearity issue between markers—it happens because markers in close positions are expected to be highly correlated (Neves et al. 2012).

GS applicability might be influenced by many factors, including heritability, genetic architecture, marker density and population structure (Guo et al. 2014; Crossa et al. 2017). Complex traits presenting low heritability and small marker effects are suitable for GS, whereas less complex traits, showing high heritability, can be predicted through few markers presenting relatively strong effects. Distant genetic relationship between reference and validation population resulted in poor genomic prediction performance in complex traits (Wang et al. 2017). Besides, genomic prediction accuracy is lower in complex traits presenting large numbers of markers that are not found in linkage disequilibrium (LD) through quantitative trait loci (QTL) (Daetwyler et al. 2010). The efficiency of trait-specific subset of markers depends critically on the LD between the SNPs with substantial estimated effects and the true causative loci that affect the trait of interest (Wu et al. 2016). We evaluated a high-density panel in data of three populations presenting different population structure in complex traits (grain yield and plant height). Accuracy calculated based on a high-density panel was higher in moderate complex traits (plant height) and in more closely related individuals (set of maize data).

After selecting the markers subsets, it was observed small accuracy loss—approximately 4–16% (HEL dataset) and 1–8% (USP dataset), when high-density panel prediction (all markers available) was compared to a subset formed by 100 SNPs presenting strong effect. However, the loss in accuracy was higher for RICE, approximately 11–38%. These results may also due to the expanded opportunity to borrow information from closely related genotypes and from better population structure in HEL and USP (hybrids obtained through crossings using 111 and 49 parental inbreed lines, respectively) than from less related genotypes included in the RICE dataset (from a diversity panel) (Figure S1). We observed a wide distribution of a subset with 2500 markers obtained through the ORI and REE strategies applied to all maize and rice chromosomes. Only around 4% (RICE dataset) and 16% (USP dataset) of markers were

coincident between the two evaluated traits (Table S4).

There were fewer markers close to the QTL when lower marker densities were used, fact that explains the effect of QTL. However, markers in linkage disequilibrium with QTL are not likely to explain all the existing genetic variations; the remaining genetic variation will be included in the polygenic variance (Solberg et al. 2008). According to Solberg et al. (2008), the polygenic variance is expected to be more important for prediction accuracy. Several studies showed that the selection of markers based on high density genotyping can be an effective alternative to genomic prediction (Spindel et al. 2015, 2016; Tayeh et al. 2015; Ma et al. 2016; Li et al. 2018). Another way would be to weight the genomic relationship matrix with SNPs previously selected from genome-wide association studies (GWAS) (Su et al. 2014; Spindel et al. 2016). Zhang et al. (2015) showed that G-matrix weighted according to individual SNPs that present strong and robust association signals can effectively improve genomic predictions. VanRaden et al. (2017) observed accuracy increase due to marker addition. These markers were selected based on the highest significance test, the largest absolute effect or on the largest genetic variance, which contributed to the locus, in high-density SNP chips for animal breeding. However, we used only markers with higher effects to build GRM and our results show that this approach is efficient and easy to be implemented in breeding programs. The main genetic effect model uses the GRM for genomic predictions, with the advantage of being able to capture LD between markers and causal genes, Mendelian segregation, genetic links through unknown common ancestors, and a low computational requirement. The selection of markers with higher effects for trait-specific, probably allowed associations between selected SNP markers and true QTL regions. It was also observed that the value of genetic variance per SNP was similar using subset of markers to all markers (except to USP dataset), which would be partly due to the additional variance explained by the correlated effect of SNPs around those used to built GRM. We built GRM by taking into consideration the GBLUP and Gaussian kernel and concluded that by using Gaussian kernel we could achieve improved prediction accuracy in comparison to the linear kernel, mainly in grain yield. Gaussian kernel can capture non-additive effects, whereas GBLUP only captures the additive effects (Gianola et al. 2014; Cuevas et al. 2016b).

Yu et al. (2014) recently developed accurate fixed arrays with 6 K SNPs through high density genotyping for a rice breeding program. They selected representative SNPs found through genetic diversity studies and markers located inside genes of important traits evenly distributed in 12 chromosomes. Furthermore, Spindel et al. (2016) used genomic prediction models to incorporate de novo GWAS. They used same set of rice data adopted in our study and found that approximately 5000 SNPs were effective to predict and suggest the possible use of smaller fixed SNP arrays. Besides, machine learning methods have been applied to genomic selection to identify a subset of SNPs presenting direct links to candidate genes that affect the complex traits. Li et al. (2018) found that the 3000 top SNPs identified through machine learning methods had similar accuracy values to those calculated based on the whole SNP panel. Based on our results, eliminating markers presenting small effect is a good strategy that has the potential to build a trait-specific low-density SNP chip and to reduce genotyping costs in breeding programs.

Similar methods based on selected markers evenly distributed across the genome were reported in the literature (Moser et al. 2010; Spindel et al. 2015; Tayeh et al. 2015). Spindel et al. (2015) observed that subsets evenly distributed throughout the genome were the most important contributors to PA and they had less accuracy variance than the randomly chosen ones. Moser et al. (2010) showed that a chip with 3000 evenly spaced markers could provide approximately 90% accuracy achieved through high-density. The selection of equally-spaced markers across traits overcame the problem by using a subset of SNPs specific to the trait of interest (Ogawa et al. 2014). In our study, with regards to GY and PH traits, 2500 SNPs evenly distributed along the genome were able to satisfy the need of accurate GS prediction in the investigated RICE and maize (HEL and USP) populations.

Regarding the bias parameters, we found slight differences between models and strategies (Table S2). According to Porto-Neto et al. (2015), biases affect the range of estimated breeding values, whereas accuracy affects the ranking of individuals. The aforementioned authors worked with data of animals and found lower bias when they used SNPs selected from related

individuals. We found lower GEBVs bias in the set of maize data (more related individuals) in comparison to the RICE dataset. Strategies and kernels showed similar results. Ma et al. (2015) analyzed the same strategies to reduce biases in predicted genetic trends and they concluded that this parameter should be taken into consideration to validate genomic predictions. The accuracy of the best genomic-selection methods was comparable to the accuracy of phenotypic selection, even at the herein investigated relatively low density of the markers. We found similar coincidence by using markers subsets and all markers by comparing the coincidence of the genome to the phenotypic selection and relative efficiency per cycle (ERc) (Table S3). Although the coincidence did not reach 100%, genomic selection remains a promising alternative, mainly because it has the advantage of reducing the time of a reproductive cycle (Bassi et al. 2015; Bhat et al. 2016; Crossa et al. 2017). Overall, our results showed that 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 based on a high-density panel. The best strategy adopted for the markers subsets is the effect of the original marker, which slightly increases the accuracy and reduces biases.

This study analyzed two strategies to identify low-density SNPs for genomic prediction in a training population. However, we only evaluated two strategies to select subsets of SNPs for genomic prediction in a single trait, rather than multiple traits. Studies available in the literature show the application of machine learning to the genomic prediction of multiple traits (He et al. 2016). Accurate fixed arrays can be developed based on a high-density genotyping to select representative SNPs found through single traits and markers located inside the genes of important traits. Nevertheless, several factors should be taken into consideration in practical application in order to show how different breeding populations require different numbers of markers to perform genomic selection. The accuracy of genomic selection will rapidly decay as the result of decreasing family relationship after several selection generations (Wang et al. 2017). A genomic selection model should, therefore, be updated after the phenotypes of next generation individuals are available. Therefore, higher marker density is expected to capture more LD in the training population for multi-generation selection.

Conclusion

The main reasons contributing to the slow implementation of genomic selection in breeding programs include the non-existence of commercially available large-sized SNP panels. It happens due to the lack of reference quality genome sequences and high costs associated with the need of a large numbers of genotype individuals in reference populations for the genomic prediction of target populations. Therefore, 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 based on a high-density panel. Both strategies adopted to find markers subsets allowed increasing accuracy with lower biases. However, it is necessary developing more studies to prove the efficiency of ORI and REE strategies to find markers subsets and to be used in genomic prediction models.

Acknowledgements We thank Helix Sementes® (São Paulo, Brazil) for the dataset, and Allogamous Plant Breeding Laboratory for the technical and scientific support. Funding was provided by National Council for Scientific and Technological Development (CNPq).

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. https://doi.org/10.1016/j.plantsci.2015.08.021

Bhat JA, Ali S, Salgotra RK et al (2016) Genomic selection in the era of next generation sequencing for complex traits in plant breeding. Front Genet 7:1–11. https://doi.org/10.3389/fgene.2016.00221

Bian Y, Holland JB (2017) Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. Heredity (Edinb) 118:585–593. https://doi.org/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. https://doi.org/10.1016/j.ajhg.2009.01.005

Crossa J, de los Campos G, Perez-Rodriguez P et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. https://doi.org/10.1534/genetics.110.118521

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

Crossa J, Pérez-Rodríguez P, Cuevas J et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975

Cuevas J, Crossa J, Montesinos-Lopez O et al (2016a) Bayesian genomic prediction with genotype × environment interaction kernel models. G3 (Bethesda). https://doi.org/10.1534/g3.116.035584

Cuevas J, Crossa J, Soberanis V et al (2016b) Genomic prediction of genotype × environment interaction kernel regression models. Plant Genome. https://doi.org/10.3835/plantgenome2016.03.0024

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031. https://doi.org/10.1534/Genetics.110.116855

de los Campos G, Perez-Rodriguez P (2016) BGLR: Bayesian generalized linear regression. R package version 1.0.5. http://CRAN.R-project.org/package=BGLR. Accessed 10 Aug 2016

e Souza MB, Cuevas J, de Couto EGO et al (2017) Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. G3 Genes Genomes Genet. https://doi.org/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. https://doi.org/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. https://doi.org/10.1534/genetics.105.049510

Gianola D, Weigel KA, Krämer N et al (2014) Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS ONE. https://doi.org/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. https://doi.org/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. https://doi.org/10.1186/s12711-015-0102-z

Guo Z, Tucker DM, Basten CJ et al (2014) The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet 127:749–762. https://doi.org/10.1007/s00122-013-2255-x

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

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

He S, Schulthess AW, Mirdita V et al (2016) Genomic selection in a commercial winter wheat population. Theor Appl Genet 129:641–651. https://doi.org/10.1007/s00122-015-2655-1

Heslot N, Yang H-P, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146. https://doi.org/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) 6:2919–2928. https://doi.org/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. https://doi.org/10.2135/cropsci2015.06.0332

Li B, Zhang N, Wang YG et al (2018) Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet 9:1–20. https://doi.org/10.3389/fgene.2018.00237

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. https://doi.org/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. https://doi.org/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

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. https://doi.org/10.1186/1297-9686-42-37

Neves HHR, Carvalheiro R, Queiroz SA (2012) A comparison of statistical methods for genomic selection in a mice population. BMC Genet 13:1. https://doi.org/10.1186/1471-2156-13-100

Ogawa S, Matsuda H, Taniguchi Y et al (2014) Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese Black beef cattle. BMC Genet 15:15. https://doi.org/10.1186/1471-2156-15-15

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 Genet 2:1595–1605. https://doi.org/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) 2:1595–1605. https://doi.org/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. https://doi.org/10.1186/s12711-015-0162-0

Resende MFR, Munoz P, Resende MDV et al (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190:1503–1510. https://doi.org/10.1534/genetics.111.137026

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and

densities. J Anim Sci 86:2447–2454. https://doi.org/10.2527/jas.2007-0010

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. PLoS Genet. https://doi.org/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. https://doi.org/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. https://doi.org/10.3168/jds.2014-8210

Szyda J, Zukowski K, Kamiński S, Zarnecki A (2013) Testing different single nucleotide polymorphism selection strategies for prediction of genomic breeding values in dairy cattle based on low density panels. Czech J Anim Sci 58:136–145

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. https://doi.org/10.3389/fpls.2015.00941

Thomson MJ (2014) High-throughput SNP genotyping to accelerate crop improvement. Plant Breed Biotechnol 2:195–212. https://doi.org/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 Genom 15:823. https://doi.org/10.1186/1471-2164-15-823

VanRaden PM (2007) Genomic measures of relationship and inbreeding. Interbull Annu Meet Proc 37:33–36. https://doi.org/10.1007/s13398-014-0173-7.2

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980

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. https://doi.org/10.1186/s12711-017-0307-4

Vazquez AI, Rosa GJM, Weigel KA 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. https://doi.org/10.3168/jds.2010-3335

Wang Q, Yu Y, Yuan J et al (2017) Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp *Litopenaeus vannamei*. BMC Genet 18:1–9. https://doi.org/10.1186/s12863-017-0507-5

Weigel KA, 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. https://doi.org/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, pp 1–43

Wu XL, Xu J, Feng G et al (2016) Optimal design of low-density SNP arrays for genomic prediction: algorithm and applications. PLoS ONE 11(9):e0161719

Yu H, Xie W, Li J et al (2014) A whole-genome SNP array (RICE6 K) for genomic breeding in rice. Plant Biotechnol J 12:28–37. https://doi.org/10.1111/pbi.12113

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. https://doi.org/10.1371/journal.pone.0012648

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 Genet 5:615–627. https://doi.org/10.1534/g3.114.016261