Published online September 27, 2018

RESEARCH

Assessing Prediction Models for Different Traits in a Rice Population Derived from a Recurrent Selection Program

Odilon P. Morais Júnior,* Flávio Breseghello, João Batista Duarte, Alexandre S. G. Coelho, Tereza C. O. Borba, Jordene T. Aguiar, Péricles C. F. Neves, and Orlando P. Morais

ABSTRACT

Genomic selection (GS) is a promising approach to improve rice (*Oryza sativa* L.) populations by using genome-wide markers for selection prior to phenotyping to estimate breeding values. In this study, our objectives were to compare certain prediction models with different structures of genetic relationship and statistical approaches for relevant traits in rice and to discuss some implications for integrating GS into a recurrent selection program of irrigated rice. We assessed nine models in terms of predictive potential, using empirical data from $S_{1:3}$ progenies phenotyped for eight traits with different heritabilities and genotyped with 6174 high-quality single nucleotide polymorphism markers. For all traits, marker-based models outperformed prediction based on pedigree records alone. A similar level of accuracy was observed for many models, although the level of prediction stability and prediction bias varied widely. Random forest was slightly superior for less complex traits, although with high prediction bias, whereas the semiparametric RKHS method (reproducing kernel Hilbert spaces) was superior for many traits, showing high stability and low bias. Bayesian variable selection method Bayes Cπ showed acceptable accuracy and stability for several traits and thus could be useful for genomic prediction aiming at persisting accuracy for a long-term recurrent selection.

O.P. Morais Júnior; J.B. Duarte, and A.S.G. Coelho, Dep. of Genetics and Plant Breeding, Escola de Agronomia, Federal Univ. of Goiás, Campus Samambaia, Ave. Esperança, s/n, CEP 74690-900, Goiânia, Goiás, Brazil; F. Breseghello, T.C.O. Borba, P.C.F. Neves, and O.P. Morais, Embrapa Arroz e Feijão, Rodovia GO-462, km 12, CP 179, CEP 75375-000, Santo Antônio de Goiás, Goiás, Brazil; J.T. Aguiar, Dep. of Animal Science, Escola de Medicina Veterinária e Zootecnia, Federal Univ. of Tocantins, BR 153, km 112, CP 132, CEP 77800-000, Araguaína, Tocantins, Brazil. This article is dedicated in memoriam to O.P. Morais. Received 3 Feb. 2018. Accepted 10 Aug. 2018. *Corresponding author (odilonpmorais@gmail.com). Assigned to Associate Editor Michael Casler.

Abbreviations: ABLUP, pedigree-based best linear unbiased prediction; AGBLUP, pedigree-marker-based best linear unbiased prediction; **A**, pedigree-based relationship matrix; BCπ, Bayes Cπ; BL, Bayesian least absolute shrinkage and selection operator; BLUP, best linear unbiased predictor; BV, breeding value; DTF, days to flowering; GBLUP, marker-based best linear unbiased prediction; **G**, marker-based relationship matrix; GS, genomic selection; GY, grain yield; HBLUP, single-step best linear unbiased prediction; **H**, combined relationship matrix; IPB, incidence of panicle blast; LD, linkage disequilibrium; LWR, length/width ratio; MSE, mean squared error; PAbi, predictive ability; PAcc, predictive accuracy; PBia, prediction bias; PSta, predictive stability; PH, plant height; PLSR, partial least squares regression; QTL, quantitative trait locus; RF, random forest; RKHS, reproducing kernel Hilbert spaces; SBS, severity of brown spot; SNP, single nucleotide polymorphism; CHG, grain chalkiness; WGY, whole-grain yield.

Rice (*Oryza sativa* L.) is the world's most important food. Therefore, the current annual growth rate of rice yield of ~1% yr^{-1} is insufficient to meet the projected future demand, raising doubts about regular supply and affordable prices, especially in the face of limitation of arable land area, environmental damage, and looming climate change (Ray et al., 2013; Long et al., 2015). Rice breeding programs are expected to contribute with the challenge of increasing that rate of yield progress by

Published in Crop Sci. 58:1–13 (2018).
doi: 10.2135/cropsci2018.02.0087

developing high-yield cultivars with a number of other relevant traits, including grain quality, in addition to generating and maintaining genetic variability for greater genetic gains per selection.

Breeding methods commonly used in self-pollinated species have limited potential for improving quantitative traits and combining favorable alleles coming from many different parents (Bernardo, 2010; Ramalho et al., 2012). Recurrent selection may be a flexible tool for the continuous improvement of genetically diverse populations, increasing the frequency of favorable alleles while maintaining useful genetic variability for long-term gains (Ramalho et al., 2005; Hallauer et al., 2010; Niu et al., 2010; Morais Júnior et al., 2015, 2017a; Leite et al., 2016). Initially, the frequent crosses required for progeny recombination limited its application in rice. However, as genetic male sterility became available (Singh and Ikehashi, 1981), that limitation was overcome. Currently, the main limitation for the extensive use of recurrent selection in rice is the long cycle duration, normally 3 to 4 yr.

Genomic selection (GS; Meuwissen et al., 2001) is a breeding method that presents great potential for increasing the rate of genetic gain, especially by reducing cycle duration in recurrent selection programs. In GS, genomic estimated breeding values (GEBVs) are computed for all progenies in a breeding population, enabling early selection, thus improving breeding efficiency. Other benefits may result from a reduction in phenotyping requirements at every selection cycle (Heffner et al., 2009; Legarra et al., 2014); however, those savings must balance with the additional genotyping costs. Compared with recurrent selection programs based on phenotypic data alone, GS allows higher selection pressure through the evaluation of a large number of progenies (Bernardo, 2010). Another potential benefit of GS (over classical recurrent selection schemes) is promoting more frequent recombination, as a consequence of shortened selection-crossing cycles, generating novel and useful variation in the breeding population (Heffner et al., 2009; Müller et al., 2017).

Previous studies demonstrated a great potential of GS when applied to rice breeding (Guo et al., 2014; Grenier et al., 2015; Onogi et al., 2015; Spindel et al., 2015, 2016; Morais Júnior et al., 2017b, 2018; Monteverde et al., 2018). Nevertheless, the matter of model choice still deserves study, especially in the context of GS. Genomic selection efficiency depends on prediction accuracy, which, in turn, depends on factors like experimental quality, training population size, level of linkage disequilibrium (LD), effective population size, and marker density. It also depends on the biological nature of the trait, its genetic architecture, and its heritability (Heffner et al., 2009; Daetwyler et al., 2010). To deal with the genetic architecture of different traits in GS, several prediction models have been proposed (de los Campos et al., 2013; Gianola, 2013). Methods differ mainly in their assumptions about the number and effects of quantitative trait loci (QTLs), which influences the magnitude and persistence of prediction accuracy (Habier et al., 2011; Heslot et al., 2012).

Models must be efficient in using genetic data (molecular markers and/or pedigree information) and phenotypic data, including previously available data records from the breeding program, as much as possible (Heslot et al., 2012). Therefore, comparative assessment of the performance of prediction models, based on parametric and nonparametric methods (de los Campos et al., 2013), using information from pedigree, molecular markers, or a combination of both (Legarra et al., 2009, 2014), is essential to identify the statistical approach that will result in the highest predictive accuracy. Other studies compared prediction models in rice (Guo et al., 2014; Grenier et al., 2015; Spindel et al., 2015; Morais Júnior et al., 2017b, 2018; Monteverde et al., 2018); however, none considered simultaneously predictive accuracy and stability, prediction bias, model goodness-of-fit, and overfitting, especially in the context of GS, for several important traits in rice. Moreover, it should also be emphasized that the efficiency of prediction models may vary among populations with distinct genetic backgrounds and, mainly, between traits with different QTL effect distributions (Daetwyler et al., 2010; Windhausen et al., 2012).

In this study, we compared some prediction models with different structures of genetic relationship and statistical approaches for eight important traits with different heritabilities and genetic architectures in irrigated rice. Our objectives were to identify the most promising prediction models, to develop low-risk GS methods for use in rice breeding, and to discuss some implications for integrating GS into a recurrent selection program.

MATERIALS AND METHODS

Phenotypic Data

We used the rice population CNA12S, synthesized at Embrapa (Brazilian Agricultural Research Corporation) in 2002, by crossing 10 elite parents with six sources of resistance to rice blast disease (*Magnaporthe grisea* B.C. Couch). Details of the genealogy, developmental steps, and subpopulations composing CNA12S are described in Morais Júnior et al. (2017a). In this study, we evaluated 196 $S_{1:3}$ progenies from the third selection cycle of this population.

The phenotypic data come from a field trial conducted in 2015 in Goianira, Goiás, Brazil (16°26′12″ S, 49°23′39″ W; 729 m asl). The experimental design was an augmented lattice design with two replications for all progenies and with common check cultivars between blocks (14 × 17). Four high-yield check cultivars were used: IRGA 417 and BRS 7 Taim were present in all blocks, whereas IRGA 424 and BRS Pampa were alternated between blocks, resulting in 14 progenies and three checks per block. The plots consisted of four furrows of 5 m, spaced 0.17 m, mechanically sown with 60 seeds per meter.

The experiment was conducted in a lowland area with continuous flooding from 15 d after planting until grain maturity. Fertilizing and weed and insect pest control were done as needed for the expression of yield potential. No fungicides were applied for disease resistance expression, especially to rice blast and brown spot [*Bipolaris oryzae* (Breda de Haan) Shoemaker].

Eight traits were evaluated: (i) incidence of panicle blast (IPB, score), with visual inspection of the plants in preharvest stage, based on a diagrammatic scale (0 = no incidence, 1 = 1–5%, 3 = 6–12%, 5 = 13–25%, 7 = 26–50%, and 9 = >50% of infected panicles), according to IRRI (1996); (ii) severity of brown spot (SBS, score), with visual inspection of the plants leaves in the vegetative stage, based on a diagrammatic scale (0.5, 6, 12, 18, 24, 32, and 36% or more of the leaf area with symptoms), according to Schwanck and Del Ponte (2014); (iii) plant height (PH, cm), an average of six readings from ground level to panicle tip at prematurity; (iv) days to flowering (DTF, d), from sowing to 50% of plants at anthesis; (v) grain yield (GY, kg ha^{-1}) from a 1.36-m^2 area at 13% moisture; (vi) whole-grain yield (WGY, g), grain mass after mill; (vii) length/width ratio (LWR); and (viii) grain chalkiness (CHG, %). The last two traits were measured using a grain statistical analyzer (S21 model).

Genomic Data

Purified genomic DNA was extracted in bulk from leaf tissue of eight plants in each progeny, using an Axygen kit. Genotyping was done at Diversity Arrays Technology, Canberra, Australia, by DArTseq method in Illumina HiSeq 2500 sequencer. After filtering single nucleotide polymorphisms (SNPs) for call rate >75% and minor allele frequency >5%, 6174 markers on 174 progenies were retained for genetic analyses. According to Pérez and de los Campos (2014), missing values (~4.5%) were replaced by the expected allele computed based on estimates of allele frequencies derived from the same dataset.

Pedigree-Based Relationship Matrix

The pedigree information of the group of 196 progenies from the third selection cycle of the CNA12S population was recovered from breeding records back 10 generations. From these data, we obtained the pedigree-based relationship matrix (**A** matrix) among all the genotypes involved in the pedigree, using a modified version of the package *pedigreemm* (Bates and Vazquez, 2009) developed for R platform (R Core Team, 2018), which accounts for inbreeding (i.e., for self-pollination). The portion of this matrix corresponding to the 196 $S_{1:3}$ progenies used in this study constituted the matrix **A** considered in the models.

Marker-Based Relationship Matrix

We calculated the marker-based relationship matrix (**G** matrix) for the 174 progenies with SNP data, following the method of VanRaden (2008):

$$\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})'}{2\sum_{i=1}^{m} p_i(1-p_i)}$$

where **M** is a $n \times m$ matrix (n = number of progenies, m = number of SNPs) that specifies the SNP in each locus, encoded as 0 (homozygous for the reference allele), 1 (heterozygous), and 2 (homozygous for the alternative allele); and **P** is the matrix of observed frequencies of the alleles (p_i) expressed as $2(p_i - 0.5)$. The **M** matrix was centered (−1, 0, and 1).

Pedigree- or Marker-Based Relationship Matrix

The matrix of genetic relationship between all 196 progenies (including the 22 progenies without genotypic data) was obtained by combining the matrices **A** and **G** into the combined relationship matrix **H**, using the single-step method (HBLUP; Legarra et al., 2009) as follows:

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G}_w \end{bmatrix} = \mathbf{A}\begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}_w - \mathbf{A}_{22} \end{bmatrix}$$

where **A** is the additive relationship matrix derived from pedigree, with the subscripts 1 and 2 corresponding to nongenotyped and genotyped progenies, respectively; and $\mathbf{G}_w$ is the weighted matrix **G**, with $\mathbf{G}_w = (1-w)\mathbf{G}_n + w\mathbf{A}_{22}$, where w is the weight attributed to the portion of the genetic variance not explained by markers (Christensen and Lund, 2010), and $\mathbf{G}_n$ is the normalized **G** (we considered $w = 0.05$). Since the matrices $\mathbf{G}_n$ and $\mathbf{A}_{22}$ must be orthogonal, the matrix $\mathbf{G}_n$ was calculated via normalization based on VanRaden (2008), following Forni et al. (2011):

$$\mathbf{G}_n = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})'}{\mathrm{tr}\left[(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})'\right] \Big/ \sum_{i=1}^{N}(1+F)}$$

where **M** and **P** are as described above, tr[ ] is the trace operator of a matrix, and F is the individual inbreeding coefficient, derived from the pedigree, for the N progenies genotyped. With this procedure, matrices $\mathbf{G}_w$ and $\mathbf{A}_{22}$ become orthogonal, with similar diagonal means. The matrix **H** was standardized by dividing it by its diagonal mean. All matrix algebra was done using the R platform (R Core Team, 2018).

Phenotypic Data Analysis

A single-trait Bayesian analysis was performed to fit the phenotypic data to the following linear mixed model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{r} + \mathbf{Z}_2\mathbf{b} + \mathbf{Z}_3\mathbf{p} + \varepsilon$$

where **y** is the vector of phenotypic values, β is the vector of fixed effects (intercept, effect of check cultivars, and genotypic group—a group of check cultivars and a group of progenies), **r** is the vector of random effects of replication, **b** is the vector of random effects of blocks within replications, **p** is the vector of random effects of progenies, and ε is the vector of random residues. The terms **X**, $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ are the respective design matrices.

In the Bayesian model, we used the Gibbs sampling algorithm (Liu and Daniels, 2006), with a total chain length of 210,000 iterations, 10,000 as a burn-in period and thinning of every 20th iteration, implemented in the R platform by the package *MCMCglmm* (Hadfield, 2010). We assumed normal distributions for the parameters of fixed effects, multivariate normal distributions for the data and parameters of random effects (reduced to a normal distribution, in the case of the

single-trait model) and inverse-Wishart distributions for the variance and covariance components. The fixed effect priors were normally distributed with mean zero and variance of 10^8. For the variance and covariance components, the hyperparameters degree of belief (ν) and scale matrix (**S**) generate a uniform (noninformative) inverse-Wishart distribution by setting $\nu = -(k + 1)$ and $\mathbf{S} = \text{diag}(k) \times 0$, where k is the number of traits considered in the model (Sorensen and Gianola, 2002). Thus, we obtained posterior marginal distributions like those of the restricted maximum likelihood procedure, for variance and covariance components, and BLUP (best linear unbiased predictor) method, for the random effects.

Based on the posterior marginal distribution for each variance and covariance component, given the data, we calculated the mean and 95% highest posterior density interval for the following parameters: coefficient of genetic variation, $CV_g = \left(\sigma_g^2\right)^{0.5} / \overline{P}$; coefficient of relative variation, $CV_r = \left(\sigma_g^2 / \sigma_e^2\right)^{0.5}$; broad-sense heritability based on progeny-means, $h_{\overline{p}}^2 = \sigma_g^2 / \left[\sigma_g^2 + \left(\sigma_e^2 / k\right)\right]$; and selective accuracy, $r_{g\hat{g}} = \left(h_{\overline{p}}^2\right)^{0.5}$. In these expressions, σ_g^2 and σ_e^2 are the genetic variance between progenies and residual variance between plots, $\overline{P}$ is the grand mean of progenies, and k is the weighted numbers of replications.

Estimation of Linkage Disequilibrium and Population Structure

From the 6174 high-quality SNPs available, 5264 were mapped to one of the 12 rice chromosomes. We estimate LD as the square of the correlation of allele frequencies (r^2; Hill and Robertson, 1968) between pairs of SNPs mapped within each chromosome, using the function *pairwiseLD* from the package *synbreed* (Wimmer et al., 2012) in the R platform. The LD decay as a function of physical distance (in bp) between SNP pairs was obtained by nonlinear regression (Remington et al., 2001), considering all the chromosomes, using the R package *nls* (Grothendieck, 2013). According to Sved (1971), under equilibrium drift recombination, $E(r^2) = 1/(1 + \rho)$, where $E(r^2)$ is the expected mean value to r^2, and $\rho = 4N_e c$, with N_e being the effective population size and c the recombination rate between SNPs per generation. According to Hill and Weir (1988), assuming low mutation rate and finite sample size, its expectation is expressed as

$$E\left(r^2\right) = \left[\frac{10 + C}{(2 + C)(11 + C)}\right]\left[1 + \frac{(3 + C)\left(12 + 12C + C^2\right)}{n(2 + C)(11 + C)}\right]$$

where n is the sample size (174 progenies), and C is product between the recombination rate ($\rho = 4N_e c$) and the physical distance between SNPs in base pairs.

Population structure of CNA12S was estimated for the 196 progenies belonging to the 18 subpopulations of the population through principal component analysis (PCA), using the R package *FactoMineR* (Lê et al., 2008). In a first approach, PCA was applied on the genotypic means (BLUP values) of progenies, considering the eight evaluated traits. In a second approach, PCA was applied on the **H** matrix.

Definition of Prediction Models

Nine prediction models were tested for each trait, with different genetic relationship matrices and statistical approaches (Table 1). We used R packages for fitting the prediction models. In most cases, we considered the standard configuration defined in each package, with the following exceptions: (i) for the Bayes Cπ (BCπ) model, we defined a noninformative prior distribution, uniform in the interval (0, 1), for the parameter π (proportion of non-null markers effects), by the definition of the parameters $p0 = 2$ and $\pi 0 = 0.5$; (ii) for the partial least squares regression (PLSR) model, we previously defined the optimal number of uncorrelated components that minimized the mean squared error (MSE) of prediction by cross-validation; (iii) for the reproducing kernel Hilbert spaces (RKHS) model, we defined the Gaussian reproducing kernel $[K_{ij} = e^{-(D_{ij}/\theta)^2}]$ function (Gianola and Van Kaam, 2008) with a marker-based Euclidean distance D_{ij} between all pairs of progenies i and j, normalized to the interval [0,1], and the more provable bandwidth parameter θ, after testing 10 values in the log-likelihood profile. To fit pedigree-based best linear unbiased prediction (ABLUP), pedigree-marker-based best linear unbiased prediction (AGBLUP), HBLUP, marker-based best linear unbiased prediction (GBLUP), BCπ, and Bayesian least absolute shrinkage and selection operator (BL) models, based on a Bayesian approach, 55,000 iterations were set, with a 5000 burn-in period and thinning of every fourth iteration (Pérez and de los Campos, 2014).

Cross-Validation Procedure and Model Comparisons

Cross-validation was done using the training–testing partition procedure, according to Pérez and de los Campos (2014), with 50 random samples. In each partition, 90% of the progenies were assigned to the training set and the remaining 10% were assigned to the testing set, and those sets were used to assess each predictive model. Thus, for the group of progenies genotyped ($N = 174$ progenies), we considered training dataset (N_{trn}) equal to 157 progenies, and for the whole dataset ($N = 196$ progenies), N_{trn} was equal to 176 progenies. Therefore, we used the same training–testing partitions to assess the cross-validated-based parameters in each predictive model. The predictive accuracy (PAcc) of each model was computed by Pearson's correlation between the predicted breeding values (BVs) and genotypic values (BLUP values) from the testing dataset.

Predictive ability (PAbi) of each model was computed by Pearson's correlation between the predicted BVs and BLUP values for all progenies (i.e., without a cross-validation procedure). In this study, the general term "BV" was used to designate models that are based on genomic and/or pedigree information, rather than the term "GEBV," which refers only to models using genomic information. The predictive stability (PSta) of each model's BVs was obtained by Pearson's correlation between the BVs obtained without cross-validation and the BVs obtained from the testing dataset (i.e., based on the cross-validation procedure).

Prediction bias (PBia) was calculated as the slope coefficient for the regression of BVs on BLUP values from the testing set. Unbiased models are expected to result a slope coefficient of 1. A slope coefficient >1 indicates less shrunken BV prediction,

and a coefficient <1 indicates more shrunken BV prediction. In turn, model overfitting, a statistical phenomenon by which the regression coefficients can be misleading, was calculated as the difference between the PAbi and the PAcc for each model. The model's MSE was calculated between the BVs obtained without cross-validation and the BVs based on the cross-validation procedure. For each model, we also calculated the computation time for analysis, considering the 50 replications, performed on a PC Dell with a 3.4-GHz Intel Core i5 processor and 8 GB of RAM memory. To compare the agreement between models, regarding the ranking of the progenies based on their BVs, we calculate the Spearman's rank correlation between BVs from the testing dataset across all traits.

RESULTS

Experimental Precision and Relationship Matrices

Experimental precision was high in general, resulting in heritabilities >0.5 for yield-related traits (GY and WGY) and >0.8 for all other traits, which are less influenced by the environment (Table 2). Coefficients of relative variation (CV_r) >1 and selective accuracy ($r_{g\hat{g}}$) >0.7 indicated a favorable condition for selection, due to high levels of genetic variability.

A heat map of the relationship matrices for the 196 progenies (Fig. 1) indicated that the relationship matrix based on pedigree and marker information (matrix **H**) showed a wider range of variation of relatedness between progenies than the matrix based on pedigree relationship only (matrix **A**). This result points toward a better discriminatory ability of pedigree-marker based relationships in capturing both the allele sharing (within-family variation) or Mendelian segregation and genetic links through unknown common ancestors, which are not available in pedigree records.

Linkage Disequilibrium and Population Structure

The extent and decay of LD as a function of physical distance, considering all chromosomes, are shown in Fig. 2. The mean LD decay rate across the genome was relatively low, considering a population (CNA12S) with *indica* genetic background. The expected mean LD value (r^2 = 0.23), corresponding to a reduction of 50% of the maximum value predicted by the nonlinear regression (r^2 = 0.46), reached a physical distance of ~330 kb. The LD value of r^2 = 0.12, corresponding to one-fourth of the maximum value predicted, reached with a physical distance of ~1160 kb. There was no stabilization in the mean values of r^2 within the 6000-kb window.

Table 1. Description of each prediction model, type of genetic relationship matrix used, number of progenies considered, R package used for model fitting, and the main reference of each model.

Model†	Matrix‡	No. of progenies	R package	Reference
ABLUP	**A**	196	*BGLR*	VanRaden (2008)
GBLUP	**G**	174	*BGLR*	VanRaden (2008)
AGBLUP	**A, G**	174	*BGLR*	de los Campos et al. (2009)
HBLUP	**H**	196	*BGLR*	Legarra et al. (2009)
BCπ	**M**	174	*BGLR*	Habier et al. (2011)
BL	**M**	174	*BGLR*	Park and Casella (2008)
PLSR	**M**	174	*pls*	Wold (1966)
RF	**M**	174	*randomForest*	Breiman (2001)
RKHS	**M**	174	*rrBLUP*	Gianola and Van Kaam (2008)

† ABLUP, pedigree-based best linear unbiased prediction; GBLUP, marker-based best linear unbiased prediction; AGBLUP, pedigree-marker-based best linear unbiased prediction; HBLUP, single-step best linear unbiased prediction; BCπ, Bayes Cπ; BL, Bayesian least absolute shrinkage and selection operator; PLSR, partial least squares regression; RF, random forest; RKHS, reproducing kernel Hilbert spaces.

‡ **M**, centralized matrix; **A**, pedigree-based relationship matrix; **G**, marker-based relationship matrix; **H**, combined relationship matrix.

Table 2. Posterior mean and 95% highest posterior density interval of the marginal distribution of coefficient of genetic variation (CV_g), coefficient of relative variation (CV_r), broad-sense heritability based on progeny-means ($h^2_{\bar{p}}$), and selective accuracy ($r_{g\hat{g}}$) for the CNA12S population.

Trait†	CV_g	CV_r	$h^2_{\bar{p}}$	$r_{g\hat{g}}$
GY (kg ha^{-1})	11.5 (8.86–13.7)	0.77 (0.64–0.83)	0.54 (0.39–0.64)	0.73 (0.62–0.80)
PH (cm)	5.03 (4.49–5.69)	1.73 (1.70–1.81)	0.86 (0.82–0.89)	0.93 (0.90–0.94)
DTF (d)	4.14 (3.67–4.66)	1.76 (1.70–1.81)	0.86 (0.81–0.89)	0.92 (0.90–0.94)
IPB (score)	75.0 (66.7–83.7)	2.00 (1.91–2.01)	0.89 (0.85–0.91)	0.94 (0.92–0.95)
SBS (score)	30.3 (27.2–34.4)	1.76 (1.72–1.82)	0.85 (0.81–0.89)	0.92 (0.90–0.94)
WGY (g)	12.2 (9.88–14.5)	0.82 (0.72–0.90)	0.56 (0.44–0.67)	0.76 (0.66–0.82)
LWR	4.77 (4.28–5.30)	3.22 (3.12–3.23)	0.95 (0.94–0.96)	0.98 (0.97–0.98)
CHG (%)	51.1 (46.2–57.6)	2.15 (2.11–2.20)	0.90 (0.87–0.92)	0.95 (0.93–0.96)

† GY, grain yield; PH, plant height; DTF, days to flowering; IPB, incidence of panicle blast; SBS, severity of brown spot; WGY, whole-grain yield; LWR, length/width ratio; CHG, grain chalkiness.

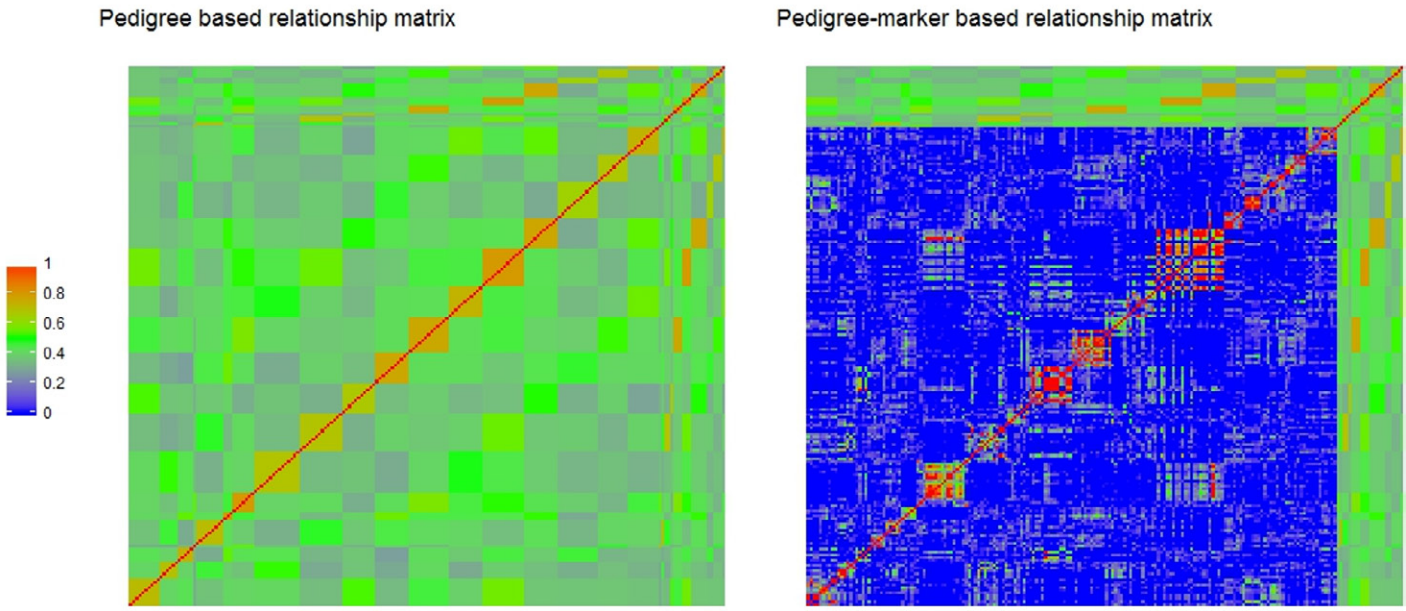

Fig. 1. Heat map of the pedigree (**A** matrix)-based and combined or pedigree-marker (**H** matrix)-based relationship matrices for the 196 progenies, displaying the familial relatedness between the progenies.

No evident population structure was detected by the first two principal components in PCA, based on BLUP values for the traits evaluated (Fig. 3, Panel A). While suggesting certain clustering of progenies, PCA on matrix **H** did not indicate prominent population structure (Fig. 3, Panel B). The dispersion between BVs (predicted by HBLUP model) and BLUP values for each trait did not reveal clear clusters (Fig. 4). Those results suggest an absence of heterogeneity in SNP marker effects in the set of progenies; therefore, stratified sampling of the population for genomic prediction was not justified.

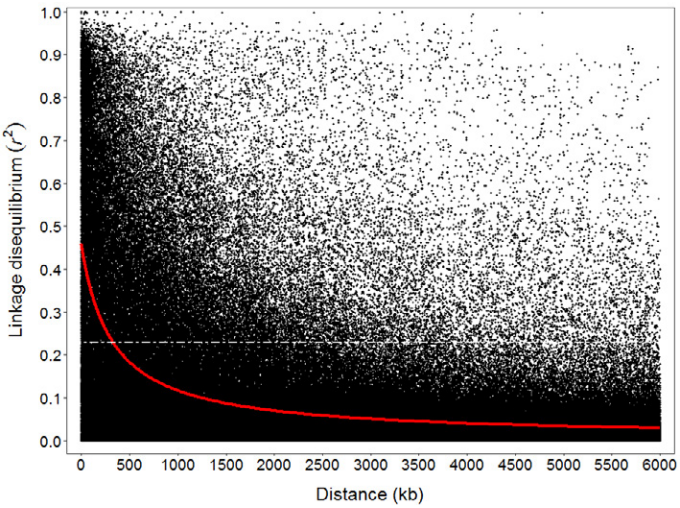

Fig. 2. Linkage disequilibrium (LD, r^2) vs. physical map distance, between pairs of single nucleotide polymorphism markers per chromosome, with measures of all chromosomes in a panel of 174 progenies, considering a 6000-kb window. The black dots correspond to the observed LD values, and the red line corresponds to the nonlinear trend of expected LD decay; the white dashed line intersects the value of r^2 corresponding to the half (r^2 = 0.23) of the maximum estimated (r^2 = 0.46) in the trend line, with R^2 = 0.255.

Genetic Complexity of Traits

The parameter π from the BCπ model gives an estimate of the proportion of loci with nonzero effects on the trait. This estimate can be considered an indication of the complexity of the genetic architecture of the trait, since complex traits are normally influenced by many genes. The highest π values were found for PH (0.48), LWR (0.48), GY (0.46), IPB (0.43), and WGY (0.43), indicating that almost half of the markers analyzed had some effect on those traits. The traits DTF and CHG showed moderate complexity, with π values equal to 0.35 and 0.34, respectively. The trait SBS showed the lowest genetic complexity, with only ∼3% of the loci having significance.

Comparison of Prediction Models

The heritability of a trait corresponds to the upper limit of the phenotypic variance explained by a linear predictor based on DNA markers (Wray et al., 2013). Therefore, for all prediction models and traits (data not shown), the PAcc was linearly correlated (e.g., accuracy = 0.49, for the GBLUP model) with the broad-sense heritability as expected.

Predictive accuracies of all traits with the models under study are presented in Table 3. The model ABLUP, which uses only pedigree relationship information, showed lower predictive accuracies in general. Among the models using genomic relationship data, random forest (RF) presented the highest PAcc for SBS, LWR, and CHG, whereas HBLUP was the most accurate for PH and DTF. All of the models using genomic relationship matrices were similarly accurate for predicting GY, WGY, and IPB.

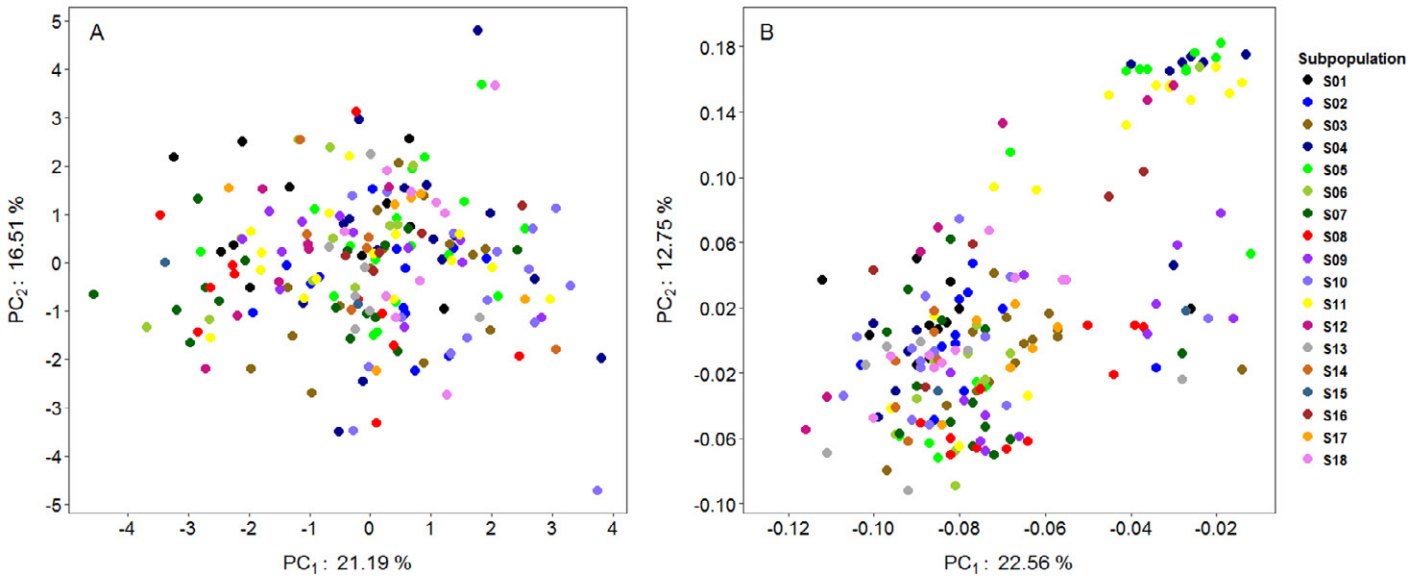

Fig. 3. Dispersion of the 196 progenies (with the identification of the 18 constituent subpopulations: S01–S18), in relation to the first (PC_1) and second (PC_2) principal components, estimated based on (A) genotypic means or (B) pedigree-marker-based relationship matrix.

The PSta measures the dependency of BVs on phenotypes. The PSta of BVs from each model is shown in Table 4. There was large variation in model stability between traits. The highest prediction stabilities were found for the following combinations of models and traits: RKHS for GY, PH, and IPB; PLSR for DTF and IBP; BL for WGY and LWR; and RF for CHG.

The PAbi of each model's BVs (Table 5) indicates the model's quality of fit. Considering the average across all traits, RF was the model with the highest PAbi, but ABLUP and PGBLUP were very similar to RF in this criterion. The PLSR model showed high PAbi for WGY, LWR, CHG, and PH but fell to lower values for other traits.

As a measure of the PBia of each model's BVs, we also calculated the slope coefficient of BVs on genotypic values (Table 6). In addition to large differences between the models for the same trait, there were slopes significantly different from 1, indicating certain bias in the prediction. The AGBLUP, BL, and RF models presented the most bias for six traits, whereas BCπ and RKHS presented biased predictions for only two traits. The PLSR model tended to show less shrunken prediction, whereas the RF model was prone to more shrunken BV prediction.

We calculated overfitting (lower values indicate less overfitting), MSE, and computation time across traits for each prediction model (Table 7). The model ABLUP, which uses only pedigree information (matrix **A**), is clearly overfitting. This result suggests that pedigree-based model is more tailored to fit the quirks and random noise in a specific sample, rather than reflecting the overall population. On the other hand, HBLUP, PLSR, and RKHS models have lower overfitting; therefore, they should result in better performance of this population after selection.

Although BL, HBLUP, and BCπ showed lower MSE values, PLSR and RF performed poorly in this criterion (Table 7). As the MSE measures the deviation between the BVs without cross-validation and the BVs based on cross-validation, a high MSE suggests a worse fitting when predicted through cross-validation. Therefore, the higher MSE values found for PLSR and RF models might explain the greater PBia with those models (Table 6). The mean computation time for the univariate analyses done with 50 replications (Table 7) was short for the RKHS model (<1 min); intermediate for GBLUP-based models in Bayesian framework, PLSR and RF models (∼20–50 min); and long for BCπ and BL models (4–5 h).

The Spearman's rank correlations between BVs for all pairs of models were high, despite the differences between models for several criteria. The lowest values of rank correlations were obtained between RF model and GBLUP-based models. However, even those were ≥0.9.

DISCUSSION

Factors Related to the Predictive Accuracy

Phenotypic data quality is an important factor in GS studies, because of the strong influence on the PAcc. According to Resende and Duarte (2007), the higher the trait heritability, the greater the efficacy in inference about genetic values and, consequently, the greater the accuracy. The significant linear correlation (data not shown) detected in our study between PAcc and trait heritability demonstrates the dependence of genomic predictions on phenotypic data quality.

Linkage disequilibrium is the base of GS, since its decay in the genome is inversely proportional to the magnitude of the accuracy. In our study, the mean LD (r^2 = 0.23) for a distance of ∼330 kb was higher than observed in other

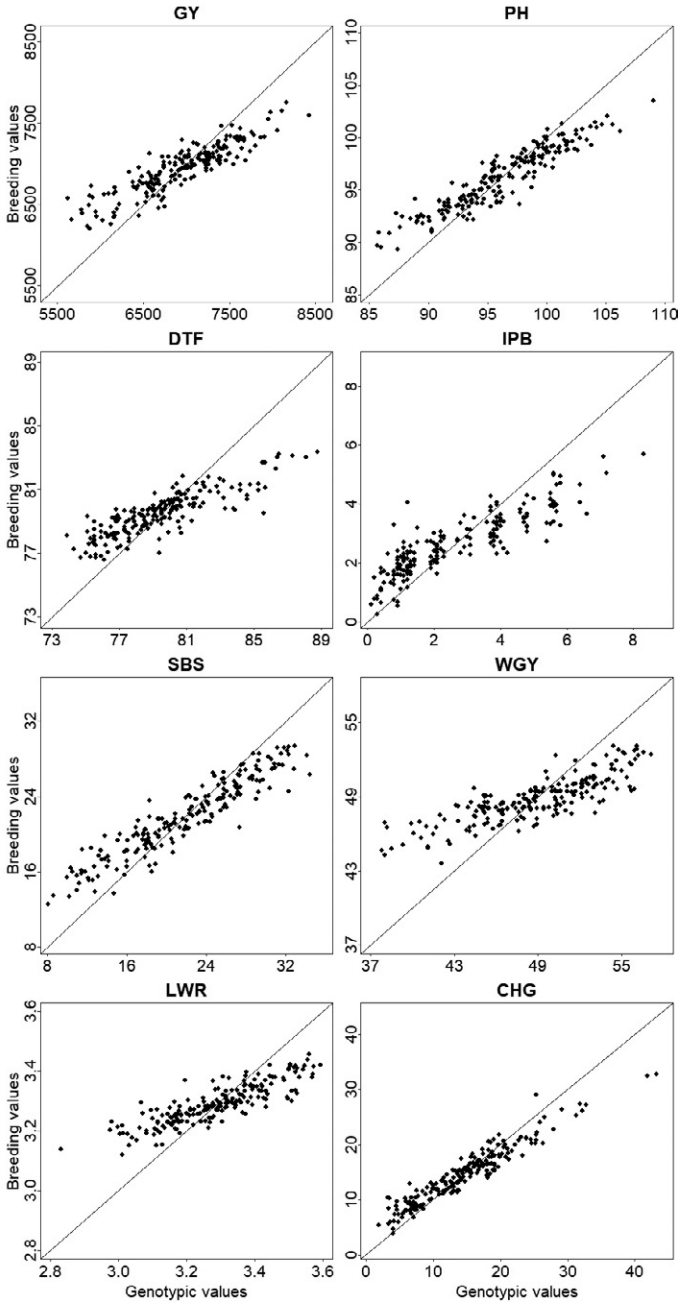

Fig. 4. Dispersion between the breeding values predicted by the single-step best linear unbiased prediction (HBLUP) model and the genotypic values (best linear unbiased prediction [BLUP] values) for grain yield (GY, kg ha^{-1}), plant height (PH, cm), days to flowering (DTF, d), incidence of panicle blast (IPB, score), severity of brown spot (SBS, score), whole-grain yield (WGY, g), length/width ratio (LWR), and grain chalkiness (CHG, %).

studies in cultivated rice, especially for *indica* germplasm, which shows mean LD decay at distances between 50 and 200 kb (Huang et al., 2010; Xu et al., 2012). On the other hand, in a *japonica* genetic background, LD decay is reported at distances between 200 and 650 kb (Xu et al., 2012; Grenier et al., 2015).

There is a strong relationship between LD extent and marker density for genomic prediction (Zhong et al., 2009; Daetwyler et al., 2010). A low rate of LD decay implies that a lower marker density is needed to detect effects of causal polymorphisms (Zhong et al., 2009). For an elite population of irrigated rice, Spindel et al. (2015) detected that about one SNP per 52 kb (∼6000–7000 SNPs) was sufficient for GS. Grenier et al. (2015) reported that about one SNP per 45 kb (∼8000–9000 SNPs) allowed sufficient genome coverage for GS in upland rice. The high LD detected in the CNA12S population suggests that approximately one SNP per 60 kb should be sufficient for GS with high prediction accuracy.

According to Daetwyler et al. (2010), the training population in GS should be sufficiently large to ensure that all relevant haplotypes are effectively represented by their phenotypes. Notwithstanding, haplotype frequencies are dependent on allele frequencies, LD decay, and genetic distance between markers such that when LD is high, fewer individuals are required to represent the number of possible haplotypes. The high LD found in our study, which was expected in a population under recurrent selection, allows the use of training sets of this size (∼176 individuals), thus contributing to the efficiency of use of resources in the plant breeding program.

Population structure occurs when the population contains subgroups with pronounced differences in allelic frequencies (Price et al., 2010). The absence of population structure is a favorable situation for obtaining reliable estimates of accuracy (Guo et al., 2014). The recombination of the CNA12S population by circulating diallel scheme in consecutive cycles probably contributed to allele sharing and preventing population structure. This result is encouraging towards the application of GS in rice recurrent selection.

This research is based on a dataset from only one environment, possibly contributing to an upward bias in estimates of genetic parameters (e.g., heritability estimates and predictive accuracies), due to the likely presence of genotype × environment interaction. Nevertheless, the dataset used showed great experimental precision, with a large amount of genetic variation and significant realized heritabilities for several traits (Morais Júnior et al., 2017b). Thus, any upward bias associated with genotype × environment interaction was sufficiently small that it did not negate the potential for genetic improvement in this population based on selection within one environment.

Assessment of Trait Genetic Complexity

The genetic complexity of traits is an important aspect to be considered when applying GS in plant populations. The genetic architecture of a trait in a given population may be estimated by the parameter π, obtained by the BCπ model (Habier et al., 2011). However, this inference should be made with caution (Gianola, 2013), due to the superparametrization of the model that occurs when the number of markers is much larger than the number of progenies in

the training set. The quantitative traits GY, PH, LWR, and WGY presented high π values (π > 0.4), justifying the use of infinitesimal models, like GBLUP, for their prediction.

Considering that the synthesis of the CNA12S population was meant to create a base population with stable resistance to rice blast (*Magnaporthe oryzae* B.C. Couch), the PAcc obtained for IPB (∼0.5) evidences high potential of GS applied in a population under recurrent selection to enhance resistance to this disease. The high genetic complexity of IPB, indicated by π = 0.43, agrees with the current knowledge of a large number of genes with small and large effects related to durability and broad-spectrum resistance to *M. oryzae* (Chen et al., 2015). In a genome-wide association study with 420 rice lines, Kang et al. (2016) identified 116 genes, underlying 97 loci associated with resistance, distributed on the 12 chromosomes. Therefore, high levels of resistance in synthetic populations might be reached by the accumulation of alleles via GS.

Table 3. Predictive accuracy of model's breeding value for grain yield (GY), plant height (PH), days to flowering (DTF), incidence of panicle blast (IPB), severity of brown spot (SBS), whole-grain yield (WGY), length/width ratio (LWR), and grain chalkiness (CHG).

	Model†								
Trait‡	ABLUP	GBLUP	AGBLUP	HBLUP	BCπ	BL	PLSR	RF	RKHS
GY	0.24b	0.44a	0.45a	0.44a	0.43a	0.43a	0.43a	0.42a	0.44a
PH	0.36c	0.52b	0.54ab	0.60a	0.52b	0.52b	0.53b	0.48b	0.51b
DTF	0.39a	0.26b	0.28b	0.40a	0.26b	0.26b	0.23b	0.30b	0.25b
IPB	0.14b	0.50a	0.50a	0.47a	0.50a	0.50a	0.48a	0.51a	0.50a
SBS	0.39c	0.62b	0.62b	0.62b	0.69a	0.62b	0.59b	0.74a	0.61b
WGY	0.16b	0.34a	0.32a	0.37a	0.33a	0.32a	0.36a	0.34a	0.31a
LWR	0.36c	0.45b	0.46b	0.46b	0.46b	0.44b	0.49b	0.52a	0.43b
CHG	0.33d	0.68b	0.69b	0.61c	0.69b	0.67bc	0.67bc	0.76a	0.67bc

† ABLUP, pedigree-based best linear unbiased prediction; GBLUP, marker-based best linear unbiased prediction; AGBLUP, pedigree-marker-based best linear unbiased prediction; HBLUP, single-step best linear unbiased prediction; BCπ, Bayes Cπ; BL, Bayesian least absolute shrinkage and selection operator; PLSR, partial least squares regression; RF, random forest; RKHS, reproducing kernel Hilbert spaces.

‡ For each trait, values with the same letter do not differ by Tukey's test (α = 0.05) with false discovery rate correction, and the greatest value is underlined.

Table 4. Predictive stability of model's breeding value for grain yield (GY), plant height (PH), days to flowering (DTF), incidence of panicle blast (IPB), severity of brown spot (SBS), whole-grain yield (WGY), length/width ratio (LWR), and grain chalkiness (CHG).

	Model†								
Trait‡	ABLUP	GBLUP	AGBLUP	HBLUP	BCπ	BL	PLSR	RF	RKHS
GY	0.47f	0.76c	0.66d	0.78c	0.75c	0.77c	0.88b	0.57e	0.93a
PH	0.56e	0.80b	0.71c	0.82b	0.80b	0.82b	0.66cd	0.63d	0.89a
DTF	0.60d	0.74c	0.54d	0.79bc	0.76bc	0.81b	0.94a	0.45e	0.75c
IPB	0.41e	0.84b	0.73c	0.82b	0.84b	0.86b	0.96a	0.66d	0.93a
SBS	0.57f	0.81cd	0.75e	0.78de	0.91a	0.83bc	0.92a	0.86b	0.93a
WGY	0.42c	0.70a	0.56b	0.74a	0.69a	0.75a	0.43cd	0.47c	0.60b
LWR	0.55d	0.71ab	0.63c	0.71ab	0.71a	0.74a	0.54d	0.66bc	0.71ab
CHG	0.56d	0.81b	0.80b	0.81b	0.81b	0.81b	0.69c	0.86a	0.69c

† ABLUP, pedigree-based best linear unbiased prediction; GBLUP, marker-based best linear unbiased prediction; AGBLUP, pedigree-marker-based best linear unbiased prediction; HBLUP, single-step best linear unbiased prediction; BCπ, Bayes Cπ; BL, Bayesian least absolute shrinkage and selection operator; PLSR, partial least squares regression; RF, random forest; RKHS, reproducing kernel Hilbert spaces.

‡ For each trait, values with the same letter do not differ by Tukey's test (α = 0.05) with false discovery rate correction, and the greatest value is underlined.

Table 5. Predictive ability of model's breeding value for grain yield (GY), plant height (PH), days to flowering (DTF), incidence of panicle blast (IPB), severity of brown spot (SBS), whole-grain yield (WGY), length/width ratio (LWR), and grain chalkiness (CHG). For each trait, the greatest value is underlined.

	Model†								
Trait	ABLUP	GBLUP	AGBLUP	HBLUP	BCπ	BL	PLSR	RF	RKHS
GY	0.96	0.90	0.96	0.90	0.90	0.88	0.77	0.98	0.70
PH	0.97	0.93	0.97	0.93	0.92	0.91	0.96	0.98	0.84
DTF	0.97	0.83	0.95	0.86	0.81	0.75	0.53	0.98	0.81
IPB	0.96	0.88	0.96	0.88	0.88	0.86	0.68	0.98	0.74
SBS	0.98	0.96	0.98	0.93	0.87	0.94	0.83	0.98	0.83
WGY	0.96	0.89	0.96	0.84	0.89	0.84	0.99	0.98	0.92
LWR	0.97	0.93	0.98	0.87	0.93	0.91	0.99	0.98	0.93
CHG	0.97	0.97	0.98	0.96	0.97	0.96	0.99	0.99	0.99

† ABLUP, pedigree-based best linear unbiased prediction; GBLUP, marker-based best linear unbiased prediction; AGBLUP, pedigree-marker-based best linear unbiased prediction; HBLUP, single-step best linear unbiased prediction; BCπ, Bayes Cπ; BL, Bayesian least absolute shrinkage and selection operator; PLSR, partial least squares regression; RF, random forest; RKHS, reproducing kernel Hilbert spaces.

As LWR and DTF showed high heritability, high accuracies were expected by using additive models. However, even RF and RKHS models, which also capitalize epistatic effects, were not able to obtain accuracies for those traits at levels equivalent to other traits. These results may be explained by two factors: (i) genetic control with a large contribution of nonadditive effects, especially dominance deviation; or (ii) multicollinearity between SNP markers, due to the high LD, which makes it difficult to identify QTL effects by regression models. Multicollinearity between markers was detected for DTF by Grenier et al. (2015). Therefore, additional studies are needed to evaluate the relevance of the relationship structures for additive–dominant or additive–dominant–epistatic models for application in rice.

The small π value (0.03) for SBS is an evidence of a low genetic complexity of rice resistance to *B. oryzae*. This was also reflected in higher accuracies by BCπ and RF models, which promote variable selection to account for the genetic variance relative to specific regions of the genome (Habier et al., 2011). Sato et al. (2015) identified five QTLs responsible for explaining 9.7 to 19.2% of the phenotypic variation in resistance to brown spot. Therefore, high genetic gains for SBS in the CNA12S population could be obtained through GS or via specific marker-assisted selection.

Comparison of Prediction Models

One of the primary steps in the application of GS in a breeding program is the choice of the best model for each trait of interest. In this case, the optimal model should have the highest PAcc and PSta, associated with low PBia and low overfitting on the training dataset (Habier et al., 2011; Patry and Ducrocq, 2011). In addition, the prediction model should be based as much as possible on marker-QTL LD, rather than on genetic relationship (Habier et al., 2013), be easy to implement for a wide range of traits, and be computationally efficient (Heslot et al., 2012). Accuracy values obtained through cross-validation varies between traits and populations, due to different factors such as heritability, genetic variance and genetic architecture of traits, LD decay, marker density, and population size (Zhong et al., 2009; Daetwyler et al., 2010). The accuracies detected in the CNA12S population

Table 6. Model bias (slope coefficient [*b*]) and SE (in parentheses) of each model for grain yield (GY), plant height (PH), days to flowering (DTF), incidence of panicle blast (IPB), severity of brown spot (SBS), whole-grain yield (WGY), length/width ratio (LWR), and grain chalkiness (CHG).

	Model†								
Trait	ABLUP	GBLUP	AGBLUP	HBLUP	BCπ	BL	PLSR	RF	RKHS
GY	–	–	0.81* (0.10)	–	–	0.74* (0.12)	1.28** (0.06)	0.78* (0.11)	–
PH	0.78* (0.11)	–	0.83** (0.06)	0.75** (0.06)	–	0.85* (0.07)	1.27** (0.04)	0.67** (0.09)	–
DTF	0.58** (0.13)	1.23* (0.14)	–	–	–	–	1.54** (0.11)	–	1.29* (0.17)
IPB	–	–	0.83* (0.10)	–	–	0.73** (0.11)	–	0.69** (0.11)	–
SBS	0.70** (0.10)	0.80** (0.06)	0.68** (0.06)	–	0.87** (0.04)	0.80** (0.06)	–	0.88** (0.04)	–
WGY	–	–	–	–	–	–	1.44** (0.06)	–	–
LWR	0.36** (0.17)	–	0.79* (0.11)	1.18* (0.10)	–	0.66* (0.14)	1.26** (0.06)	0.64** (0.09)	–
CHG	0.62* (0.17)	0.71** (0.06)	0.65** (0.06)	–	0.70** (0.06)	0.67** (0.06)	–	0.49** (0.07)	0.84** (0.05)

* Significant at the 0.05 probability level by Student *t* test.

** Significant at the 0.01 probability level by Student *t* test.

† ABLUP, pedigree-based best linear unbiased prediction; GBLUP, marker-based best linear unbiased prediction; AGBLUP, pedigree-marker-based best linear unbiased prediction; HBLUP, single-step best linear unbiased prediction; BCπ, Bayes Cπ; BL, Bayesian least absolute shrinkage and selection operator; PLSR, partial least squares regression; RF, random forest; RKHS, reproducing kernel Hilbert spaces.

Table 7. Estimates of predictive ability (PAbi), predictive stability (PSta), predictive accuracy (PAcc), overfitting, mean squared error (MSE), and computation time for each prediction model, averaged across the eight traits assessed.

	Model†								
Parameter‡	ABLUP	GBLUP	AGBLUP	HBLUP	BCπ	BL	PLSR	RF	RKHS
No. of traits with bias	5	3	6	3	2	6	5	6	2
PAbi	0.97	0.91	0.97	0.90	0.90	0.88	0.84	0.98	0.85
PSta	0.52	0.77	0.67	0.78	0.78	0.80	0.75	0.64	0.81
PAcc	0.29	0.48	0.48	0.46	0.48	0.47	0.47	0.51	0.46
Overfitting	0.67	0.43	0.48	0.39	0.41	0.41	0.37	0.47	0.38
MSE	0.44	0.32	0.54	0.28	0.28	0.25	0.70	0.64	0.36
Computation time (min)	25.1	19.3	36.0	24.5	238.8	309.0	48.7	30.5	0.7

† ABLUP, pedigree-based best linear unbiased prediction; GBLUP, marker-based best linear unbiased prediction; AGBLUP, pedigree-marker-based best linear unbiased prediction; HBLUP, single-step best linear unbiased prediction; BCπ, Bayes Cπ; BL, Bayesian least absolute shrinkage and selection operator; PLSR, partial least squares regression; RF, random forest; RKHS, reproducing kernel Hilbert spaces.

‡ Overfitting is the difference between the predictive ability (non-cross-validated correlation) and the predictive accuracy (cross-validated correlation). The MSE is calculated between the non-cross-validated breeding values and cross-validated breeding values of progenies. The computation time of each model considered 50 replications.

for traits related to GY, agronomic aspects, resistance to diseases, and grain quality are similar or superior to those reported in other studies for rice (Grenier et al., 2015; Spindel et al., 2015).

For all traits evaluated, prediction models based exclusively on genomic data (GBLUP, BCπ, BL, PLSR, RF, and RKHS) or combining genomic and pedigree data (AGBLUP and HBLUP), presented higher accuracy than that based only on pedigree data (ABLUP). This result is consistent with several studies that have reported superiority of marker-based models over pedigree (de los Campos et al., 2009; Spindel et al., 2015), although differences between approaches are dependent on the population, environment, and trait under selection. In this context, there are benefits in using marker-based relationship matrices, such as avoiding selection of closely related progenies, providing better accuracies when unrelated progenies are involved, and correcting for pedigree errors (Daetwyler et al., 2010). This expresses the importance of markers in the capitalization of allele sharing (within family variation) or Mendelian segregation effects, useful for increasing PAcc (VanRaden, 2008).

For the traits PH and DTF, there were tendencies of higher accuracies by using HBLUP model, possibly by accommodating genotyped and nongenotyped progenies, therefore using maximum size. This result agrees with other studies, in which there have been verified the increase in PAcc as population size increases, mainly for values up to 1000 individuals (Grattapaglia and Resende, 2010). However, especially in self-pollinated species, satisfactory PAcc values have been detected, even with low to moderate population size (50–200), especially under high marker density, associated with high minor allele frequency (Xavier et al., 2016).

Taking all the criteria for model selection used in this study, none of the models was clearly superior to the others. The nonlinear semiparametric RKHS tends to outperform other models in terms of PSta and PBia for several traits, besides allowing high computational efficiency. It should be noted that RKHS based on a Gaussian kernel could potentially capture epistatic effects between markers (Gianola et al., 2006). However, in this case, the predictions obtained would not be exclusively BVs, since they may not be fully transferred to the progeny.

The HBLUP model shows great potential for GS (Legarra et al., 2014); however, it was superior only for PH and DTF in our study. This model uses all the information related to pedigree-based relationship and marker-based kinship, using all individuals (genotyped and nongenotyped) in the population. Studies have observed an increase in PAcc and reduction of PBia when considering polygenic effects in prediction models, especially for populations with reduced LD and limited marker density (Calus and Veerkamp, 2007; Gao et al., 2012); however, its superiority has been more pronounced for populations with high N_{trn} (Onogi et al., 2015), differently from the case in our study.

Random forest is a widely used machine-learning algorithm that has been applied to classification and regression problems. For GS, this model has presented better performance when applied to traits controlled by a small number of QTLs, with moderate to high heritability, on populations with lower N_{trn} (Heslot et al., 2012; Onogi et al., 2015; Spindel et al., 2015). In fact, RF performed well in terms of PAcc in our study, for traits with high heritability, like SBS, LWR, and CHG. This model is also computationally efficient, less prone to overfitting, and less sensitive to noise. However, the RF model showed high PBia and low PSta.

Bayesian methods that perform variable selection (BCπ and BL models), although presenting PAcc and satisfactory PSta, presented the highest computational time. Additionally, BL presented high PBia for some traits. The relevance of computational time as a criterion for model choice should be evaluated in each case. Some breeding programs may count on high-performance computing or have dedicated machines to run models overnight, not impairing its practical application in the breeding program. The BCπ model, which allows for a large number of variant effects to be set to zero, assuming a common variance for all remaining polymorphisms (Habier et al., 2011), may be recommended for traits controlled by a low number of QTLs (e.g., SBS). In turn, the requirement of a prior for the parameter π is circumvented in the BL model, which assumes all QTL effects independently and identically distributed as a double exponential but possessing thicker tails than the Gaussian prior (Legarra et al., 2011).

Prediction accuracy rapidly erodes over recombination cycles, especially under directional selection (Müller et al., 2017). The persistence of accuracy is a very important factor to consider in GS, since it reduces the need for frequent retraining of the prediction equation. The accuracy of Bayesian models tends to persist longer, because those methods are able to consider the co-segregation between markers and QTLs within the family, thus capturing Mendelian sampling (Daetwyler et al., 2010). This is especially relevant under small effective population size and when dealing with oligogenic traits with large QTL effects (Habier et al., 2013). Bayesian methods are therefore advantageous, even if they show equal or less accuracy in the training population, since greater persistence over generations is expected when applied in selection populations (Habier et al., 2013; Müller et al., 2017).

Conflict of Interest

The authors declare that there is no conflict of interest.

Acknowledgments

The authors dedicate this article to the memory of Dr. Orlando Peixoto de Morais, recognized mentor of the rice breeding team at Embrapa. We thank the entire rice breeding staff at Embrapa, especially the research assistants and field workers, and the Capes/Brazilian Ministry of Education for the doctoral scholarship granted to the first author.

References

Bates, D., and A. Vazquez. 2009. Pedigreemm: Pedigree-based mixed-effects models. R Found. Stat. Comput., Vienna. http://CRAN.R-project.org/package=pedigreemm (accessed 19 May 2018).

Bernardo, R. 2010. Genomewide selection with minimal crossing in self-pollinated crops. Crop Sci. 50:624–627. doi:10.2135/cropsci2009.05.0250

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32. doi:10.1023/A:1010933404324

Calus, M.P.L., and R.F. Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J. Anim. Breed. Genet. 124:362–368. doi:10.1111/j.1439-0388.2007.00691.x

Chen, J., P. Peng, J. Tian, Y. He, L. Zhang, Z. Liu, et al. 2015. *Pike*, a rice blast resistance allele consisting of two adjacent NBS–LRR genes, was identified as a novel allele at the *Pik* Locus. Mol. Breed. 35:117. doi:10.1007/s11032-015-0305-6

Christensen, O.F., and M.S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet., Sel., Evol. 42:2. doi:10.1186/1297-9686-42-2

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031. doi:10.1534/genetics.110.116855

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 193:327–345. doi:10.1534/genetics.112.143313

de los Campos, G., H. Naya, D. Gianola, C. Jose, A. Legarra, E. Manfredi, et al. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385. doi:10.1534/genetics.109.101501

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet., Sel., Evol. 43:1. doi:10.1186/1297-9686-43-1

Gao, H., O.F. Christensen, P. Madsen, U.S. Nielsen, Y. Zhang, M.S. Lund, and G. Su. 2012. Comparison of genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genet., Sel., Evol. 44:1–8. doi:10.1186/1297-9686-44-8

Gianola, D. 2013. Priors in whole-genome regression: The Bayesian alphabet returns. Genetics 194:573–596. doi:10.1534/genetics.113.151753

Gianola, D., R.L. Fernando, and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776. doi:10.1534/genetics.105.049510

Gianola, D., and J.B.C.H.M. Van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303. doi:10.1534/genetics.107.084285

Grattapaglia, D., and M.D.V. Resende. 2010. Genomic selection in forest tree breeding. Tree Genet. Genomes 7:241–255. doi:10.1007/s11295-010-0328-4

Grenier, C., T.V. Cao, Y. Ospina, C. Quintero, M.H. Châtel, J. Tohme, et al. 2015. Predictive ability of genomic selection in a rice synthetic population developed for recurrent selection breeding. PLoS One 10:e0136594. doi:10.1371/journal.pone.0136594 [erratum: 11(5):e0154976].

Grothendieck, G. 2013. nls2: Non-linear regression with brute force. R package version 0.2. R Found. Stat. Comput., Vienna. http://CRAN.R-project.org/package=nls2 (accessed 19 May 2018).

Guo, Z., D.M. Tucker, C.J. Basten, H. Gandhi, E. Ersoz, B. Guo, et al. 2014. The impact of population structure on genomic prediction in stratified populations. Theor. Appl. Genet. 127:749–762. doi:10.1007/s00122-013-2255-x

Habier, D., R.L. Fernando, and D.J. Garrick. 2013. Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics 194:597–607. doi:10.1534/genetics.113.152207

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186. doi:10.1186/1471-2105-12-186

Hadfield, J.D. 2010. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. J. Stat. Softw. 33:(2). doi:10.18637/jss.v033.i02

Hallauer, A.R., M.J. Carena, and J.B. Miranda Filho. 2010. Quantitative genetics in maize breeding. Springer, New York.

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12. doi:10.2135/cropsci2008.08.0512

Heslot, N., H.P. Yang, M.E. Sorrells, and J.-L. Jannink. 2012. Genomic selection in plant breeding: A comparison of models. Crop Sci. 52:146–160. doi:10.2135/cropsci2011.06.0297

Hill, W.G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231. doi:10.1007/BF01245622

Hill, W.G., and B.S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. 33:54–78. doi:10.1016/0040-5809(88)90004-4

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, et al. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961–967. doi:10.1038/ng.695

IRRI. 1996. Standard evaluation system for rice. 4th ed. IRRI, Los Baños, the Philippines.

Kang, H., Y. Wang, S. Peng, Y. Zhang, Y. Xiao, D. Wang, et al. 2016. Dissection of the genetic architecture of rice resistance to the blast fungus *Magnaporthe oryzae*. Mol. Plant Pathol. 17:959–972. doi:10.1111/mpp.12340

Lê, S., J. Josse, and F. Husson. 2008. FactoMineR: An R package for multivariate analysis. J. Stat. Softw. 25:1–18. doi:10.18637/jss.v025.i01

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–4663. doi:10.3168/jds.2009-2061

Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014. Single step, a general approach for genomic selection. Livest. Sci. 166:54–65. doi:10.1016/j.livsci.2014.04.029

Legarra, A., C. Robert-Granie, P. Croiseau, F. Guillaume, and S. Fritz. 2011. Improved LASSO for genomic selection. Genet. Res. 93:77–87. doi:10.1017/S0016672310000534

Leite, M.E., J.A. Dias, D.A. Souza, F.C. Alves, L.R. Pinheiro, and J.B. dos Santos. 2016. Increasing the resistance of common bean to white mold through recurrent selection. Sci. Agric. 73:71–78. doi:10.1590/0103-9016-2015-0084

Liu, X.F., and M.J. Daniels. 2006. A new algorithm for simulating a correlation matrix based on parameter expansion and re-parameterization. J. Comput. Graph Stat. 16:897–914. doi:10.1198/106186006X160681

Long, S.P., A. Marshall-Colon, and X.G. Zhu. 2015. Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. Cell 161:56–66. doi:10.1016/j.cell.2015.03.019

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

Monteverde, E., J.E. Rosas, P. Blanco, F.P. Vida, V. Bonnecarrère, G. Quero, et al. 2018. Multienvironment models increase prediction accuracy of complex traits in advanced breeding lines of rice. Crop Sci. 58:1519–1530. doi:10.2135/cropsci2017.09.0564

Morais Júnior, O.P., F. Breseghello, J.B. Duarte, O.P. Morais, P.H.N. Rangel, and A.S.G. Coelho. 2017a. Effectiveness of recurrent selection in irrigated rice breeding. Crop Sci. 57:3043–3058. doi:10.2135/cropsci2017.05.0276

Morais Júnior, O.P., J.B. Duarte, F. Breseghello, A.S.G. Coelho, T.C.O. Borba, J.T. Aguiar, et al. 2017b. Relevance of additive and nonadditive genetic relatedness for genomic prediction in rice population under recurrent selection breeding. Genet. Mol. Res. 16:gmr16039849. doi:10.4238/gmr16039849

Morais Júnior, O.P., J.B. Duarte, F. Breseghello, A.S.G. Coelho, O.P. Morais, and A.M. Magalhães Júnior. 2018. Single-step reaction norm models for genomic prediction in multienvironment recurrent selection trials. Crop Sci. 58:592–607. doi:10.2135/cropsci2017.06.0366

Morais Júnior, O.P., P.G.S. Melo, O.P. Morais, A.P. Castro, F. Breseghello, M.M. Utumi, et al. 2015. Genetic progress after cycles of upland rice recurrent selection. Sci. Agric. 72:297–305. doi:10.1590/0103-9016-2014-0137

Müller, D., P. Schopp, and A.E. Melchinger. 2017. Persistency of prediction accuracy and genetic gain in synthetic populations under recurrent genomic selection. G3 (Bethesda) 7:801–811. doi:10.1534/g3.116.036582

Niu, N., V.N. Arief, I.H. Delacy, D. Lush, J. Sheppard, G. Zhang, and M.J. Dieters. 2010. Genetic gain in yield and protein over two cycles of a wheat recurrent selection program. Breed. Sci. 60:181–186. doi:10.1270/jsbbs.60.181

Onogi, A., O. Ideta, and Y. Inoshita. 2015. Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.). Theor. Appl. Genet. 128:41–53. doi:10.1007/s00122-014-2411-y

Park, T., and G. Casella. 2008. The Bayesian lasso. J. Am. Stat. Assoc. 103:681–686. doi:10.1198/016214508000000337

Patry, C., and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic pre-selection in dairy cattle. J. Dairy Sci. 94:1011–1020. doi:10.3168/jds.2010-3804

Pérez, P., and G. de los Campos. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495. doi:10.1534/genetics.114.164442

Price, A.L., N.A. Zaitlen, D. Reich, and N. Patterson. 2010. New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11:459–463. doi:10.1038/nrg2813

R Core Team. 2018. R: A language and environment for statistical computing. R Found. Stat. Comput., Vienna, Austria.

Ramalho, M.A.P., Â.F.B. Abreu, and J.B. Santos. 2005. Genetic progress after four cycles of recurrent selection for yield and grain traits in common bean. Euphytica 144:23–29. doi:10.1007/s10681-005-5694-x

Ramalho, M.A.P., Â.F.B. Abreu, J.B. Santos, and J.A.R. Nunes. 2012. Aplicações da genética quantitativa no melhoramento de plantas autógamas. Fed. Univ. of Lavras, Lavras, Minas Gerais, Brazil.

Ray, D.K., N.D. Mueller, P.C. West, and J.A. Foley. 2013. Yield trends are insufficient to double global crop production by 2050. PLoS One 8:e66428. doi:10.1371/journal.pone.0066428

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley et al. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98:11479–11484. doi:10.1073/pnas.201394398

Resende, M.D.V., and J.B. Duarte. 2007. Precisão e controle de qualidade em experimentos de avaliação de cultivares. Pesqui. Agropecu. Trop. 37:182–194.

Sato, H., K. Matsumoto, C. Ota, T. Yamakawa, J. Kihara, R. Mizobuchi. 2015. Confirming a major QTL and finding additional loci responsible for field resistance to brown spot (*Bipolaris oryzae*) in rice. Breed. Sci. 65:170–175. doi:10.1270/jsbbs.65.170

Schwanck, A.A., and E.M. Del Ponte. 2014. Accuracy and reliability of severity estimates using linear or logarithmic disease diagram sets in true colour or black and white: A study case for rice brown spot. J. Phytopathol. 162:670–682. doi:10.1111/jph.12246

Singh, R.J., and H.I. Ikehashi. 1981. Monogenic male-sterility in rice: Introduction, identification and inheritance. Crop Sci. 21:286–289. doi:10.2135/cropsci1981.0011183X002100020020x

Sorensen, D., and D. Gianola. 2002. Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer, New York. doi:10.1007/b98952

Spindel, J.E., H. Begum, D. Akdemir, B. Collard, E. Redoña, J.-L. Jannink, and S. McCouch. 2016. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity 116:395–408. doi:10.1038/hdy.2015.113

Spindel, J.E., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redoña, et al. 2015. Genomic selection and association mapping in rice (*Oryza sativa* L.): 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. 11:e1004982. doi:10.1371/journal.pgen.1004982 [erratum: 11(6):e1005350].

Sved, J.A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2:125–141. doi:10.1016/0040-5809(71)90011-6

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980

Wimmer, V., T. Albrecht, H.J. Auinger, and C.C. Schön. 2012. synbreed: A framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086–2087. doi:10.1093/bioinformatics/bts335

Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J.-L. Jannink, M.E. Sorrells, et al. 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 (Bethesda) 2:1427–1436. doi:10.1534/g3.112.003699

Wold, H. 1966. Estimation of principal components and related models by iterative least squares. In: P.R. Krishnaiah, editor, Multivariate analysis. Acad. Press, New York. p. 391–420.

Wray, N.R., J. Yang, B.J. Hayes, A.L. Price, M.E. Goddard, and P.M. Visscher. 2013. Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. 14:507–515. doi:10.1038/nrg3457

Xavier, A., W.M. Muir, and K.M. Rainey. 2016. Assessing predictive properties of genome-wide selection in soybeans. G3 (Bethesda) 6:2611–2616. doi:10.1534/g3.116.032268

Xu, X., X. Liu, S. Ge, J.D. Jensen, F.Y. Hu, X. Li, et al. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. 30:105–111. doi:10.1038/nbt.2050

Zhong, S.Q., J.C.M. Dekkers, R.L. Fernando, and J.-L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182:355–364. doi:10.1534/genetics.108.098277