Published online October 13, 2017

RESEARCH

Genomic Prediction of Autogamous and Allogamous Plants by SNPs and Haplotypes

Filipe Inacio Matias,* Giovanni Galli, Italo Stefanine Correia Granato, and Roberto Fritsche-Neto

ABSTRACT

The implementation of single-nucleotide polymorphism (SNP)-based genomic selection has demonstrated great predictive potential in plants. However, its application is sometimes limited to the biallelism of the marker. In this context, the use of haplotype blocks as multiallelic markers might improve genomic prediction. This study was performed to compare the predictive ability of Bayesian genomic prediction models using haplotypes (confidence interval and four-gamete), individual SNPs, and sets of SNPs selected according to haplotype construction. The use of haplotype matrices increased the predictive ability and selection coincidence with the phenotypic selection for the maize (*Zea mays* L.) breeding population. However, this was not observed for the rice (*Oryza sativa* L.) population, in which the use of the nonreduced SNP matrix was more efficient. Overall, the use of reduced SNP matrices did not lead to better predictive abilities. No difference was observed between the genomic prediction methods used. We found that the use of haplotypes has potential to increase predictive ability of genomic prediction in breeding populations of allogamous plants or plants with high multiallelism.

F.I. Matias, G. Galli, I.S.C. Granato, and R. Fritsche-Neto, Dep. de Genética, Escola Superior de Agricultura Luiz de Queiroz, Univ. de São Paulo, Piracicaba, São Paulo, Brasil. Received 11 Jan. 2017. Accepted 8 Aug. 2017. *Corresponding author (filipematias23@usp.br). Assigned to Associate Editor Aaron Lorenz.

Abbreviations: CI, confidence interval; cSNP, conventional single-nucleotide polymorphism matrix; *D′*, allelic association; 4G, four-gamete; GEBV, genomic breeding value; GS, genomic selection; GY, grain yield; LD, linkage disequilibrium; MAF, minor allele frequency; PH, plant height; SNP, single-nucleotide polymorphism; TS, training population; VS, validation population.

Genomic selection (GS) is based on the use of genomic information to predict the genetic value of phenotyped or nonphenotyped individuals (Heffner et al., 2009). It has been effectively used in animal (de Campos et al., 2015; Farah et al., 2016) and plant breeding (Jarquin et al., 2016; Huang et al., 2016). This strategy seeks to exploit information from markers in linkage disequilibrium (LD) with chromosomal regions that control the traits under evaluation (Heslot et al., 2015). To best explore this information, the development of genotyping technologies continues to generate significant quantities of markers. However, as marker densities increase, collinearity also increases. Additionally, evaluating a large number of individuals and complex experimental designs may lead to infeasibility of computational analysis.

To reduce the number of genomic variables maintaining information, markers can be rearranged in LD blocks, known as haplotype blocks (Cuyabano et al., 2014). Capturing epistatic interactions of nearby single-nucleotide polymorphisms (SNPs), improving estimates of alleles identical by descent, and reducing the number of tests in association studies (which leads to the reduction in type I error rate) are some advantages of using haplotypes. Besides, it infers groups of correlated genes and alleles,

Published in Crop Sci. 57:2951–2958 (2017).
doi: 10.2135/cropsci2017.01.0022

which may be useful for evolutionary studies (Lorenz et al., 2010).

Haplotypes are used in human genetics studies (Gabriel et al., 2002), in genomic association in animals (Barendse, 2011; Gregersen et al., 2012) and plants (Lorenz et al., 2010; Gawenda et al., 2015), and in genomic prediction in animals (Calus et al., 2008; Villumsen et al., 2009; Edriss et al., 2013; Cuyabano et al., 2014). However, to date, few studies on the efficiency of using haplotypes in genomic prediction in plants have been reported. Thus, the use of haplotypes combined with different statistical methods on GS predictive ability seems necessary.

We sought to compare GS models using different statistical methods (Bayesian Ridge Regression, BayesB, and Bayesian LASSO) in a maize (*Zea mays* L.) breeding population, and in a rice (*Oryza sativa* L.) population. Model genomic covariates were based on two haplotype construction methods (confidence interval [CI] and four-gamete [4G]), a reduced set of SNP markers selected by haplotype block, and the total set of SNPs.

MATERIAL AND METHODS

Dataset I

Rice phenotypic and genomic data were obtained from the public database Rice Diversity (Spindel et al., 2015a). This dataset is composed of 363 inbred lines evaluated in dry seasons from 2009 to 2012. Grain yield (GY, kg ha^{-1}) and plant height (PH, cm) were assessed in randomized complete block design with three replications and one site. Lines were genotyped with genotyping-by-sequencing technology, resulting in ~108,000 SNP markers. This number of SNPs comes from the original data file "MET_crfilt_.75_allchrom.hmp.txt.zip" (Spindel et al., 2015b). To reduce the effect of the population structure in the genomic models used, a subset of 270 individuals was selected to proceed further in the genomic analysis. This procedure was performed using principal component analysis, considering the molecular marker dataset (Supplemental Fig. S1) (Novembre and Stephens, 2008; McVean, 2009; Guo et al., 2014). For more information on the panel, see Spindel et al. (2015a).

Dataset II

A set of 452 tropical maize single-crosses was provided by Helix Sementes, Rio Claro, São Paulo, Brazil. Hybrids were obtained from a partial diallel between 128 inbred lines and evaluated for GY and PH. Field trials were performed on randomized complete block design with two replications, allocated in five sites (Ipiaçú, Minas Gerais; Patos de Minas, Minas Gerais; Sertanópolis, Paraná; Nova Mutum, Mato Grosso; and Sorriso, Mato Grosso) in the 2015 growing season. Inbred lines were genotyped with the Affymetrix Axiom Maize Genotyping Array (Unterseer et al., 2014) with ~660,000 SNP markers.

Genomic information was subject to quality control using call rate and minor allele frequency (MAF) procedures. Markers with large amounts of missing data (>95%), as well as a low level of polymorphism (MAF < 0.05), were eliminated. Similarly, individuals with >10% of missing data were also eliminated. Missing data were imputed by the Synbreed (Wimmer et al., 2012) package using Beagle 4.0 software (Browning and Browning, 2007). For the rice population, quality control was performed on the inbred lines, resulting in 39,015 markers. For the maize population, call rate was performed on inbred lines, and MAF was performed on hybrids. Hybrid genotypes were obtained by allelic combination of homozygous markers of parental lines. Given the occurrence of identical neighboring markers in all lines, only one was maintained. A total of 30,467 markers were used for the genomic analysis of the maize population. Quality control procedures were performed in R, using the snpReady package (Granato, 2017).

Phenotypic Analysis

Genotypic values were obtained by mixed model equations (restricted maximum likelihood, best linear unbiased prediction), considering the joint analysis of environments. The ASReml-R package (Gilmour et al., 2009) was used to implement the statistical models for the phenotypic data analysis. The significance of the fixed effect factors was evaluated by Wald test, and the significance of random effect factors was evaluated by the likelihood ratio test.

The model for rice phenotypic data was:

$$\mathbf{y} = \mathbf{j}\mu + \mathbf{Vt} + \mathbf{Xr} + \mathbf{Ll} + \mathbf{To} + \varepsilon$$

in which $\mathbf{y}$ is the vector of phenotypic observations of the evaluated traits; $\boldsymbol{\mu}$ is the vector of general mean, $\mathbf{t}$ is the vector of year effect; $\mathbf{r}$ is the vector of replication within year (fixed); $\mathbf{l}$ is the vector of line effect, with $\mathbf{l} \sim N(0, \sigma^2_l)$ and σ^2_l being the line variance; $\mathbf{o}$ is the vector of effect of line × year interaction, in which $\mathbf{o} \sim N(0, \sigma^2_{lxt})$ and σ^2_{lxt} is the line × year interaction variance; and ε is the vector of residual, in which $\varepsilon \sim N(0, \sigma^2_\varepsilon)$ and σ^2_ε is the residual variance. $\mathbf{j}$, $\mathbf{V}$, $\mathbf{X}$, $\mathbf{L}$, and $\mathbf{T}$ are incidence matrices that relate the effects of the vectors to $\mathbf{y}$.

The following model was used for maize phenotypic data:

$$\mathbf{y} = \mathbf{j}\mu + \mathbf{Ub} + \mathbf{Hl} + \mathbf{Sa} + \mathbf{Wx} + \varepsilon$$

in which $\mathbf{b}$ is the vector of block effect within the site, in which $\mathbf{b} \sim N(0, \sigma^2_b)$ and σ^2_b is the block variance; $\mathbf{l}$ is the vector of genotype effects, in which $\mathbf{l} \sim N(0, I\sigma^2_l)$; $\mathbf{a}$ is the vector of site effects (random) being $\mathbf{a} \sim N(0, \sigma^2_a)$ where σ^2_a is the site variance, and $\mathbf{x}$ is the vector of effects of phenotype × site interaction (random), in which $\mathbf{x} \sim N(0, \sigma^2_{lxa})$ and σ^2_{lxa} is the genotype × variance. $\mathbf{U}$, $\mathbf{H}$, $\mathbf{S}$, and $\mathbf{W}$ are the incidence matrices that relate the effect of the independent vectors to $\mathbf{y}$.

Construction of Haplotype Blocks and Reduced SNP Matrices

The set of marker data was subjected to five different ways of preparing the incidence matrix of genotypes. For the definition and identification of haplotype blocks, populations were subjected to two methods implemented in the Haploview software (Barrett et al., 2005). The first method is CI (Gabriel et al., 2002), which searched for patterns of recombination by normalized measure of allelic association (D') for each pair of SNPs. However, this measurement (D') is not precise. Instead, the authors use the CI to defining haplotype blocks regions. Absence of historical recombination is admitted when the upper limit of the CI of the disequilibrium statistics is (D') >0.98 and

the lower limit is >0.70 (95% confidence). However, when the upper limit of the CI of the SNP pairs is <0.9, there is strong evidence of historical recombination, and this point is a break region and edge of the haplotype block. Thus, a block is constructed between regions that indicate historical recombination on a very small proportion (<5%) of comparisons between pairs of informative SNPs. Polymorphisms not allocated in any block were not considered to create the incidence matrix.

The second method is an extension of the algorithm 4G (Wang et al., 2002). Blocks of haplotypes are defined as a set of SNPs with no evidence of recombination (i.e., with no observation of all four possible gametes between a pair of SNPs). This search is sequential, and SNPs are added until all four possible gametes are found. When all four gametes are present with at least 1% frequency, the occurrence of recombination is declared, and a new block is constructed.

A dimensional reduction technique of the SNP matrix was proposed based on haplotype blocks constructed by CI and 4G. For each block, the SNP with the highest variance was chosen and was considered the block's representative (tag) SNP. The reduced genomic matrix was constructed on the complete SNP matrix by replacing the SNPs that formed a haplotype block by its tag SNP (i.e., they were intercalated between the SNPs that had not been allocated to any of the blocks). When considering that all SNPs within each block are transmitted together to the next generation, defining a tag SNP with a higher variance within the block can reduce the dimension of the matrices and maintain a more explanatory position of each block. Reduced SNP matrices were identified by SNP-CI and SNP-4G. Thus, five different genomic matrices were used for predictions: conventional SNP matrix (cSNP), two forms of constructing haplotypes blocks (CI and 4G), and two forms of reduction of SNP matrix based on haplotype blocks (SNP-CI and SNP-4G).

Prediction of SNP and Haplotype Effects

The prediction of the effects of markers per se (cSNP, SNP-CI, and SNP-4G) and haplotype blocks (CI and 4G) by additive genomic prediction models BRR (de los Campos et al., 2013), BayesB (Meuwissen et al., 2001; with adaptations of Pérez and de los Campos, 2014), and BLASSO (Park and Casella, 2008) were performed following the general equation:

$$\mathbf{y} = \mathbf{j}\mu + \mathbf{Z}\beta + \mathbf{U}\mathbf{b} + \varepsilon$$

in which $\mathbf{y}$ is the vector of genotypic values of lines and hybrids, β is the vector of the effects of SNPs and haplotypes, and ε is the normally distributed error vector. $\mathbf{j}$ and $\mathbf{Z}$ are the incidence matrices for μ and β. The incidence matrix contains the values 0, 1, and 2, corresponding to the number of copies of the reference allele for the SNP and haplotype carried by a given individual (Cuyabano et al., 2015). In this case, haplotypes were linearly arranged by block, considering one haplotype per column of matrix $\mathbf{Z}$.

The genomic prediction methods BRR (de los Campos et al., 2013), BayesB (Meuwissen et al. (2001; with adaptations by Pérez and de los Campos, 2014), and BLASSO (Park and Casella, 2008) were performed assuming the hyperparameters as presented in Pérez and de los Campos (2014). Fifty thousand iterations were performed, with 5000 eliminated as burn in and a thinning interval of five. The methods were implemented using the BGLR package (Pérez and de los Campos 2014) of the R software. Genomic breeding values (GEBVs) were obtained by GEBV = $\mathbf{Z}\beta$.

Validation and Comparison of Factors

Scenarios were defined by the five-by-three factorial of five incidence matrices of markers and three GS methods. Individuals were randomly divided into two populations: a training population (TS) consisting of 80% of the individuals, and a validation population (VS) with ~20%. The GEBVs of VS were predicted based on the effect of SNPs and haplotypes obtained using TS. To estimate the predictive ability of each scenario, 100 random distributions of individuals (bootstrap sampling) were performed between the populations described above. The populations formed in each one of the samples were common to all prediction scenarios. The predictive ability was estimated as the Pearson correlation between GEBVs and VS genotypic values. The ability of GS to rank individuals under the defined scenarios was also evaluated. To this end, the selection coincidence between GEBVs and estimated genotypic values (from VS) were obtained. The general procedure is explained in Supplemental Fig. S2. We applied a selection intensity range of {10%, 11%,..., 30%}. Individuals with lower GEBVs, genotypic values were selected for PH, and those with higher GEBVs and genotypic values were selected for GY.

Finally, the mean computational time for model fitting was analyzed in each scenario, in a server of n96 2.6-GHz processors, 384.5 Gb random access memory (RAM), running Linux operating system SUSE Enterprise Server.

RESULTS

The methodologies of haplotype construction presented different results for the studied populations (Table 1). In maize, ~5253 blocks with 27,936 haplotypes were observed for the CI method. In its turn, the 4G method resulted in 7772 blocks with 51,446 haplotypes. The number of SNPs per block did not vary much between methods, ranging between 2 and 12 SNPs for the first method and between 2 and 10 SNPs for the second. In rice, for the CI method, 2189 blocks with 31,917 haplotypes were observed with 2 to 196 SNPs per block. For the 4G method, 4591 blocks with 36,625 haplotypes were formed with 2 to 104 SNPs per block. For both haplotype methods, haploblocks showed high allelic variability. Mean haplotypes per block was 5.32 and 6.62 for CI and 4G, respectively, with some blocks surpassing hundreds of haplotypes in the rice population (Table 1). On the other hand, the maize population showed less haplotype per block on average, 2.85 and 3.13 for CI and 4G, respectively, but with a larger number of haplotypes (Table 1). The 4G method generated a greater amount of total haplotypes for both populations, which consequently increased the dimensionality of the incidence matrices.

The use of haplotype blocks (CI and 4G) for predicting GY in maize hybrids resulted in predictive abilities ~0.7 for all GS methods (BRR, BayesB, and BL; Table 2), greater than that observed for single SNPs (cSNP,

Table 1. Number of single-nucleotide polymorphisms (SNPs) and haplotype blocks obtained in the methods used (four-gamete [4G] and confidence intervals [CI]) in maize and rice populations. Haplotypes with a frequency <0.05 were considered low frequency.

	Maize		Rice	
Haplotype information†	CI	4G	CI	4G
SNP	14,994	24,345	37,201	36,738
Minimum number of SNPs per block	2	2	2	2
Maximum number of SNPs per block	12	10	211	129
Blocks	5,253	7,772	2,189	4,910
Haplotypes	27,936	51,446	31,917	36,625
Haplotype per block (average)	2.85	3.13	5.32	6.62
Haplotypes in minor block	3	3	2	2
Haplotypes in major block	44	36	196	104
Mean Size (bp)	35,157	90,209	105,898	47,847

† Total number.

Table 2. Predictive ability of genomic selection (SD) for grain yield (GY) and plant height (PH) in the five constructions of incidence matrices of markers (two haplotype methods and three single single-nucleotide polymorphism methods) and three prediction models (BRR, BayesB, and BL) applied to maize and rice populations.

		Rice			Maize		
Trait	Method†	BayesB	BL	BRR	BayesB	BL	BRR
GY	4G	0.33 (0.122‡)	0.31 (0.133)	0.33 (0.123)	0.71 (0.055)	0.70 (0.105)	0.71 (0.055)
	CI	0.30 (0.124)	0.26 (0.148)	0.30 (0.124)	0.71 (0.054)	0.71 (0.057)	0.71 (0.054)
	cSNP	0.36 (0.114)	0.36 (0.116)	0.36 (0.115)	0.57 (0.072)	0.56 (0.071)	0.57 (0.072)
	SNP-4G	0.30 (0.127)	0.30 (0.128)	0.30 (0.127)	0.57 (0.072)	0.57 (0.072)	0.57 (0.072)
	SNP-CI	0.30 (0.131)	0.30 (0.131)	0.29 (0.130)	0.57 (0.071)	0.56 (0.073)	0.57 (0.072)
PH	4G	0.42 (0.093)	0.39 (0.136)	0.42 (0.095)	0.82 (0.033)	0.81 (0.081)	0.82 (0.033)
	CI	0.41 (0.097)	0.38 (0.144)	0.41 (0.097)	0.82 (0.033)	0.82 (0.033)	0.82 (0.033)
	cSNP	0.40 (0.102)	0.40 (0.102)	0.40 (0.102)	0.80 (0.036)	0.80 (0.036)	0.80 (0.036)
	SNP-4G	0.42 (0.096)	0.40 (0.096)	0.41 (0.098)	0.80 (0.036)	0.80 (0.036)	0.80 (0.036)
	SNP-CI	0.42 (0.098)	0.41 (0.095)	0.42 (0.097)	0.80 (0.036)	0.80 (0.036)	0.80 (0.036)

† Methodologies of haplotype construction: four-gamete (4G) and confidence interval (CI). cSNP represents a traditional matrix. SNP-4G and SNP-CI to SNP matrices reduced by haplotype block methodologies.

‡ Standard deviation of the distribution of the 100 correlations from the 100 random samples in the cross validation.

SNP-CI, and SNP-4G). For PH, a slight superiority of the haplotype-based prediction was observed regardless of the prediction method. For GY prediction in the rice population, cSNP had similar performance to the 4G method.

Haplotype-based predictions (4G and CI) did not show significant differences for the maize population, with GY predictive ability ~0.7 and PH ~0.8 (Table 2). For the rice dataset, only a slight difference was observed was observed in GY whereupon the 4G had 0.33 of prediction, whereas CI had around 0.30 (Table 2). For rice PH, no significant differences were observed regarding the predictive ability of the evaluated scenarios.

In general, the time to compute is proportional to the size of the genomic matrix generated. Thus, the 4G method for the maize population and the cSNP method for the rice population demanded more time for model validation, around 8000 and 3500 s, respectively (Fig. 1).

The use of haplotypes in maize provided a greater selection coincidence for GY and PH across all selection intensities (Fig. 2). However, for PH, the difference between the use of haplotypes and cSNP decreased as the selection intensity decreased. Considering GY in the rice population, the decreasing selection intensity lead to greater differences among the cSNP matrix compared with other matrices (Fig. 2). For PH in rice, different types of matrices showed the same pattern of selection coincidence across selection intensities. The genomic prediction models (BayesB, BL, and BRR) show similar patterns between them when using the same selection intensity and matrices. (Fig. 2).

DISCUSSION

Regarding the methods of haplotype construction, CI was apparently more restrictive, since it resulted in a smaller number of blocks than 4G (in both populations). On the other hand, 4G presented a greater number of haplotypes per block, which consequently increased the dimensionality of incidence matrices. This indicates that the rearrangement of haplotypes in individual columns to capture the multiallelism of each block may increase the number of covariates to be predicted in the model. Consequently, one of the advantages of multiallelism might be lost by the increase in the multicollinearity of the incidence matrix or by the estimation errors resulting from the high number of effects estimated on a few phenotypic information (Aschard et al., 2015).

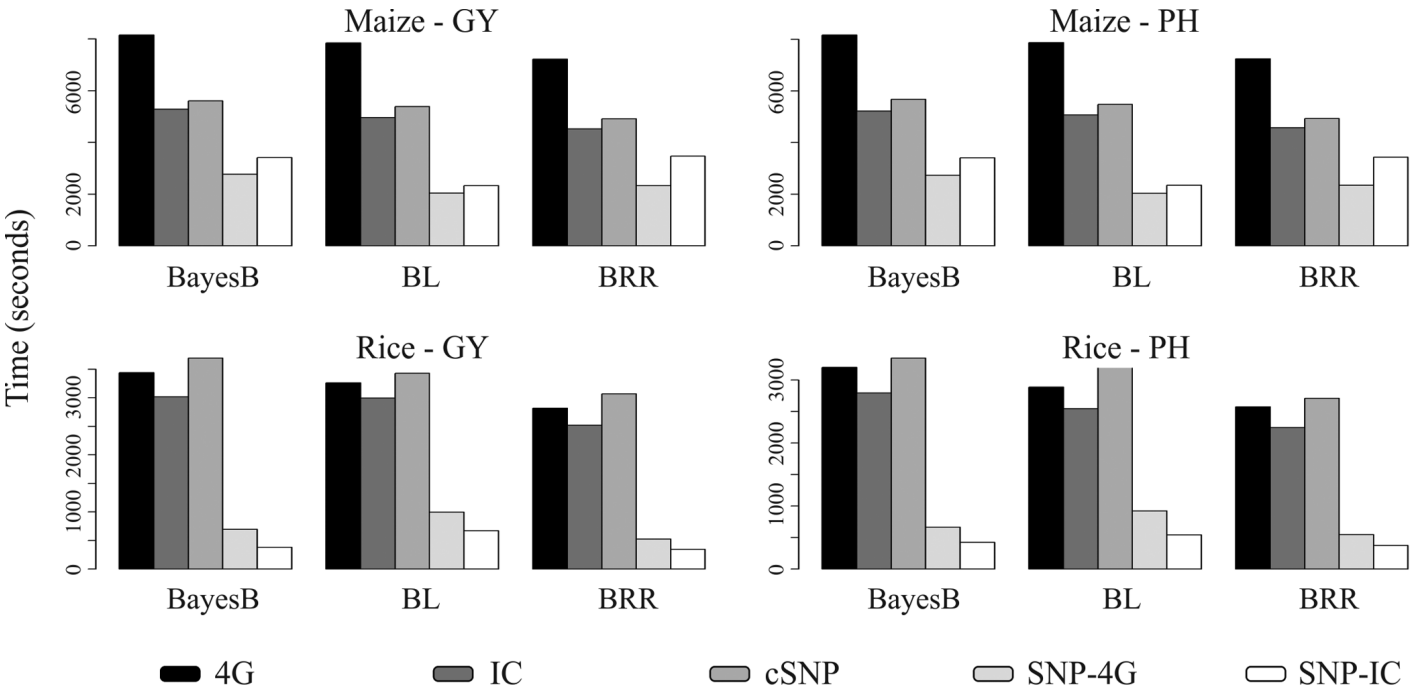

Fig. 1. Computational time in seconds for the analyses of the predictive ability of genomic selection involving five constructions of incidence matrices of markers effect (two haplotype and three single-nucleotide polymorphisms [SNPs]) and three Bayesian prediction methods (BRR, BayesB, and BL). (A) Grain yield (GY) and (B) plant height (PH) in a maize breeding population, and (C) GY and (D) PH in a rice population. Haplotype construction methodologies: four-gamete (4G) and confidence interval (CI). cSNP represents a traditional matrix. SNP-4G and SNP-CI to SNP matrices reduced the function of methodologies in haplotype blocks.

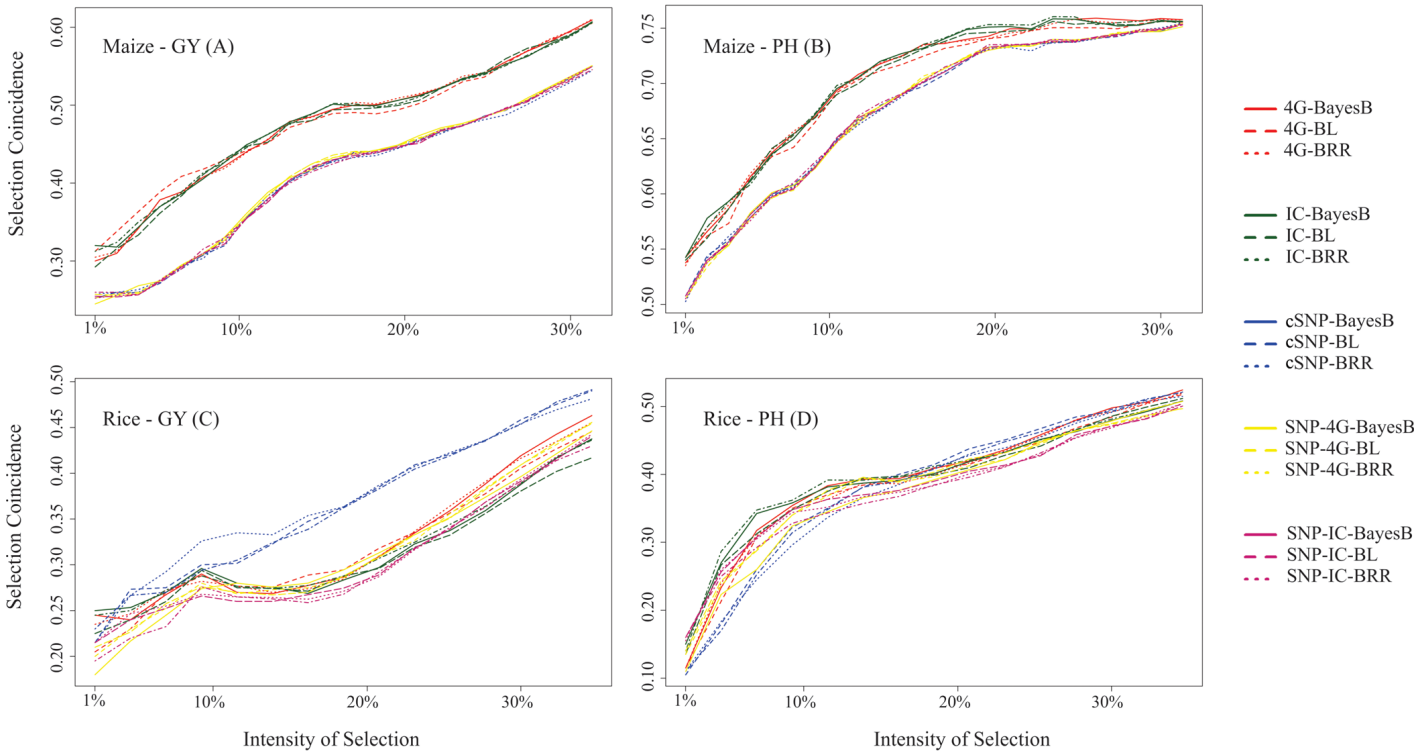

Fig. 2. Coincidence of phenotypic selection by ordering via best linear unbiased prediction and genomic selection performed by the five construction methods of incidence matrices of markers effects (two haplotypes and three single-nucleotide polymorphism methods) and three Bayesian prediction methods (BRR, BayesB, and BL). (A) Yield and (B) plant height (B) in a maize breeding population and (C) yield and (D) plant height in rice population. Haplotype construction methodologies: 4Gamete (4G) and Confidence Interval (CI). cSNP represents a traditional matrix. SNP-4G and SNP-CI to SNP matrices reduced the function of methodologies in haplotype blocks. GEBV, genomic estimated breeding value; GV, genomic variation.

Genomic selection method did not influence predictive ability. In this context, comparing the use of different Bayesian methods in cattle, Edriss et al. (2013) produced differences in predictive ability using single SNPs. However, prediction results were similar when haplotype-based schemes were used. This may be due to the fact that models based on haplotypes are less sensitive to a priori adjustments and hyperparameters (Villumsen and Janss, 2009). However, reports indicate that the use of haplotypes defined by LD can increase the predictive ability of GS models in cattle (Cuyabano et al., 2014).

In our maize population, the use of haplotypes matrices seems to be more advantageous to predict GY (Table 2). This trait had lower heritability than PH and, as reported by Combs and Bernardo (2013), has lower predictive ability accuracy than other agronomic traits regardless of population size and number of markers. Studies with empirical data indicate the superiority of models with haplotypes only for highly heritable traits (Calus et al., 2008; Cuyabano et al., 2014). However, Villumsen and Janss (2009) simulated the use of haplotypes in traits with different heritabilities and reported their advantage for low-heritability traits, which was consistent with our findings. Basically, the use of haplotypes allowed multiallelic evaluations, which consequently might better represent the variability associated with traits of low heritability, usually related to many QTLs of minor effect (Villumsen et al., 2009). In this case, haplotypes multiallelism may have promoted predictive gain for enabling more specific parameterization to the genotypes by the GS models.

The same relationship between the use of haplotypes and heritability of the trait described for maize was not observed in the rice population. Thus, cSNP led to slightly better performance than haplotype matrices regarding predictive ability for GY. This subpopulation from the diversity panel, because of its great genetic diversity, could consist of many low-frequency haplotypes, which might lead to insufficient accuracy in the estimation of their effects, reducing the efficiency of prediction. Theoretically, the LD blocks are longer in self-pollinated species (Tommasini et al., 2007; Abdurakhmonov and Abdukarimov, 2008).

The use of haplotypes for genomic prediction has been reported by several research groups (Calus et al., 2008, 2009; Edriss et al., 2013; Cuyabano et al., 2014). In general, the criteria for fixed windows of SNPs are a frequently used method for constructing haplotypes (Calus et al., 2008; Villumsen et al., 2009). However, this approach does not use some genomic properties in their definition, such as LD and historical recombination, so genetic information might be lost (Cuyabano et al., 2014). In our study, the presented methodologies for defining haplotype blocks are probably more efficient in capturing the genetic variability, because they explicitly take into account LD between markers loci (CI) or presence of historical recombination (4G) to capture genetic information. For our populations, the 4G methodology allowed more break points and block formation than CI, and the latter was more restrictive (Table 1). This was expected, given that the 4G only needs the appearance of four gametes in a certain frequency to define a block. However, if we modify the allelic frequency required in the 4G method to be more restrictive, it is possible to build longer haplotype blocks.

As pointed out by Cuyabano et al. (2014), the choice of the LD threshold to consider SNPs a block influences the number of blocks and the number of haplotypes per block. Despite the differences between the two haplotype construction methods in the number of blocks and haplotypes, we did not observe differences of predictability between those methods. Thus, haplotype information may have less of an influence on predictability.

Furthermore, since LD is influenced by the origin and/or by the genetic basis of the population or species (Flint-Garcia et al., 2003), the choice of population might influence the choice of parameters of block formation. Cavanagh et al. (2013) compared wheat populations and reported that the LD between neighboring SNPs and size of haplotypes were higher in cultivars than in landraces, showing a high level of heterogeneity in the extent of LD across the wheat genome with blocks of high-LD SNPs separated by regions with high historic recombination rates. Thus, given the differences between the two populations in the present work, it is necessary to study how the choice of these parameters might influence the block formation and the genomic prediction.

The selection of representative haploblocks had no direct influence on predictive ability, given the equality within the cSNP in the maize population and its slight inferiority in the rice population. Thus, the selection of tag SNPs within each block did not promote the increase of predictive ability, with the only advantage being the reduction of the dimensionality of the genomic matrix. The similarity of the models tested may be related to the SNP selection criteria.

Specific behavior of coincidence selection was observed regarding different selection intensities. For the maize population, haplotype matrices showed greater similarity of ranking within the phenotypic selection for both traits. For the rice population, cSNP provided greater coincidence selection for GY with practically no differences between PH prediction scenarios (Fig. 2). This result confirms the advantage of using haplotypes in breeding populations, such as the maize dataset used in this study. However, regardless of population or trait, high selection intensity reduces selection coincidence. Since selection gain is directly proportional to the selection

intensity applied, mild intensity may reduce the gain. On the other hand, it may also reduce the selection errors.

An increase in predictive ability was expected with the use of haplotypes, regardless of the population structure and genetic architecture, since they are constructed by more informative genetic approaches than single SNPs—for example, by means of LD capture. However, predictive ability on the basis of haplotype will be more informative if the mutation leading to the causal polymorphism came after the mutations creating the SNPs, placing the causal allele in greater LD with the surrounding haplotype than any single SNP.

Finally, regarding the definition of the prediction strategy, the durability of the effects over generations should also be taken into account. As reported by Villumsen et al. (2009) using simulations, models with haplotypes had less reduction in predictive ability over multiple generations. Thus, the proposition of models according to haplotypes is a powerful tool for increasing the predictive ability over several generations of a selection.

CONCLUSION

No differences were observed between the GS methods BRR, BayesB, and BL using both single SNPs and haplotypes. The use of haplotypes presented greater predictive ability and selection coincidence than single SNPs for GY in maize population. The opposite effect was observed for the rice population. Single SNP matrices reduced by haplotype information did not have predictive advantages over the use of haplotypes or all SNPs.

Conflict of Interest

The authors declare no conflict of interest.

Supplemental Material Available

Supplemental material for this article is available online.

References

Abdurakhmonov, I.Y., and A. Abdukarimov. 2008. Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int. J. Plant Genomics 2008:574927. doi:10.1155/2008/574927

Aschard, H., B.J. Vilhjálmsson, A.D. Joshi, A.L. Price, and P. Kraft. 2015. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. Am. J. Hum. Genet. 96:329–339. doi:10.1016/j.ajhg.2014.12.021

Barendse, W. 2011. Haplotype analysis improved evidence for candidate genes for intramuscular fat percentage from a genome wide association study of cattle. PLoS One 6:e29601. doi:10.1371/journal.pone.0029601

Barrett, J.C., B. Fry, J. Maller, and M.J. Daly. 2005. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265. doi:10.1093/bioinformatics/bth457

Browning, S.R., and B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097. doi:10.1086/521987

Calus, M.P.L., S.P.W. de Roos, and R.F. Veerkamp. 2009. Estimating genomic breeding values from the QTL-MAS workshop data using a single SNP and haplotype/IBD approach. BMC Proc. 3:S10. doi:10.1186/1753-6561-3-s1-s10

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. De Roos, and R.F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561. doi:10.1534/genetics.107.080838

Cavanagh, C.R., S. Chao, S. Wang, B.E. Huang, S. Stephen, S. Kiani et al. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc. Natl. Acad. Sci. U. S. A. 110:8057–8062. doi:10.1073/pnas.1217133110

Combs, E., and R. Bernardo. 2013. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6(1):1–7. doi:10.3835/plantgenome2012.11.0030

Cuyabano, B.C., G. Su, and M.S. Lund. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15:1171. doi:10.1186/1471-2164-15-1171

Cuyabano, B.C., G. Su, and M.S. Lund. 2015. Selection of haplotype variables from a high-density marker map for genomic prediction. Genet., Sel., Evol. 47:61. doi:10.1186/s12711-015-0143-3

de Campos, C.F., M.S. Lopes, F.F. de Silva, R. Veroneze, E.F. Knol, P.S. Lopes, and S.E.F. Guimarães. 2015. Genomic selection for boar taint compounds and carcass traits in a commercial pig population. Livest. Sci. 174:10–17. doi:10.1016/j.livsci.2015.01.018

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

Edriss, V., R.L. Fernando, G. Su, M.S. Lund, and B. Guldbrandtsen. 2013. The effect of using genealogy-based haplotypes for genomic prediction. Genet., Sel., Evol. 45:5. doi:10.1186/1297-9686-45-5

Farah, M.M., A.A. Swan, M.R.S. Fortes, R. Fonseca, S.S. Moore, and M.J. Kelly. 2016. Accuracy of genomic selection for age at puberty in a multi-breed population of tropically adapted beef cattle. Anim. Genet. 47:3–11. doi:10.1111/age.12362

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler. 2003. Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54:357–374. doi:10.1146/annurev.arplant.54.031902.134907

Gabriel, S.B., S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel et al. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229. doi:10.1126/science.1069424

Gawenda, I., P. Thorwarth, T. Günther, F. Ordon, and K.J. Schmid. 2015. Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. Plant Breed. 134:28–39. doi:10.1111/pbr.12237

Gilmour, A.R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide release 3.0. VSN Int., Hemel Hempstead, UK.

Granato, I.S.C. 2017. snpReady. Github. https://github.com/italo-granato/SnpReady (accessed 18 Aug. 2017).

Gregersen, V.R., L.N. Conley, K.K. Sørensen, B. Guldbrandtsen, I.H. Velander, and C. Bendixen. 2012. Genome-wide association scan and phased haplotype construction for quantitative trait loci affecting boar taint in three pig breeds. BMC Genomics 13:22. doi:10.1186/1471-2164-13-22

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

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., J.-L. Jannink, and M.E. Sorrells. 2015. Perspectives for genomic selection applications and research in plants. Crop Sci. 55:1–12. doi:10.2135/cropsci2014.03.0249

Huang, M., A. Cabrera, A. Hoffstetter, C. Griffey, D. Van Sanford, J. Costa et al. 2016. Genomic selection for wheat traits and trait stability. Theor. Appl. Genet. 129:1697–1710. doi:10.1007/s00122-016-2733-z

Jarquin, D., J. Specht, and A. Lorenz. 2016. Prospects of genomic prediction in the USDA soybean germplasm collection: Historical data creates robust models for enhancing selection of accessions. G3: Genes, Genomes, Genet. 6:2329–2341. doi:10.1534/g3.116.031443

Lorenz, A.J., M.T. Hamblin, and J.-L. Jannink. 2010. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. PLoS One 5:e14079. doi:10.1371/journal.pone.0014079

McVean, G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5:e1000686. doi:10.1371/journal.pgen.1000686

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.

Novembre, J., and M. Stephens. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 40:646–649. doi:10.1038/ng.139

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

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

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redoña et al. 2015a. 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. 11:e1004982. doi:10.1371/journal.pgen.1004982 [erratum: 11:e1005350].

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redoña et al. 2015b. 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. Dryad Digital Repository, Durham, NC. doi:10.5061/dryad.7369p

Tommasini, L., T. Schnurbusch, D. Fossati, F. Mascher, and B. Keller. 2007. Association mapping of *Stagonospora nodorum* blotch resistance in modern European winter wheat varieties. Theor. Appl. Genet. 115:697–708. doi:10.1007/s00122-007-0601-6

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

Villumsen, T.M., and L. Janss. 2009. Bayesian genomic selection: The effect of haplotype length and priors. BMC Proc. 3:S11. doi:10.1186/1753-6561-3-s1-s11

Villumsen, T.M., L. Janss, and M.S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. J. Anim. Breed. Genet. 126:3–13. doi:10.1111/j.1439-0388.2008.00747.x

Wang, N., J.M. Akey, K. Zhang, R. Chakraborty, and L. Jin. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. Am. J. Hum. Genet. 71:1227–1234. doi:10.1086/344398

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