

## ORIGINAL ARTICLE

# Genomic prediction in CIMMYT maize and wheat breeding programs

J Crossa<sup>1</sup>, P Pérez<sup>2</sup>, J Hickey<sup>1,3</sup>, J Burgueño<sup>1</sup>, L Ornella<sup>4</sup>, J Cerón-Rojas<sup>1</sup>, X Zhang<sup>1</sup>, S Dreisigacker<sup>1</sup>, R Babu<sup>1</sup>, Y Li<sup>1</sup>, D Bonnett<sup>1</sup> and K Mathews<sup>1</sup>

Genomic selection (GS) has been implemented in animal and plant species, and is regarded as a useful tool for accelerating genetic gains. Varying levels of genomic prediction accuracy have been obtained in plants, depending on the prediction problem assessed and on several other factors, such as trait heritability, the relationship between the individuals to be predicted and those used to train the models for prediction, number of markers, sample size and genotype  $\times$  environment interaction (GE). The main objective of this article is to describe the results of genomic prediction in International Maize and Wheat Improvement Center's (CIMMYT's) maize and wheat breeding programs, from the initial assessment of the predictive ability of different models using pedigree and marker information to the present, when methods for implementing GS in practical global maize and wheat breeding programs are being studied and investigated. Results show that pedigree (population structure) accounts for a sizeable proportion of the prediction accuracy when a global population is the prediction problem to be assessed. However, when the prediction uses unrelated populations to train the prediction equations, prediction accuracy becomes negligible. When genomic prediction includes modeling GE, an increase in prediction accuracy can be achieved by borrowing information from correlated environments. Several questions on how to incorporate GS into CIMMYT's maize and wheat programs remain unanswered and subject to further investigation, for example, prediction within and between related bi-parental crosses. Further research on the quantification of breeding value components for GS in plant breeding populations is required.

*Heredity* (2014) **112**, 48–60; doi:10.1038/hdy.2013.16; published online 10 April 2013

**Keywords:** Bayesian LASSO; International Maize and Wheat Improvement Center; genomic selection; genotype  $\times$  environment interaction; reproducing kernel Hilbert space regression

## INTRODUCTION

In plants, quantitative trait locus (QTL) mapping was initiated more than 20 years ago (Bernardo, 2008) and has been used very successfully in the genetic analysis of several non-complex traits and, in particular, of abiotic and biotic stress tolerance traits where much of the genetic variance is conditioned by one or a few loci. This has led to the wide-scale deployment in farmers' fields of alleles that confer resistance to submergence and salinity in rice and fungal leaf diseases in wheat and maize, and to identifying and mobilizing alleles with large effects. However, exploiting QTL for more complex traits to predict candidates for selection in practical breeding programs has had limited impact because, among other factors, of the small genetic variance accounted for by QTL owing to the model being used, and because many detected QTLs are specific to a particular genetic background.

With high-density single-nucleotide polymorphism (SNP) markers, it is possible to use multi-locus linkage disequilibrium (LD) between QTLs and markers across the whole genome to predict genetic values that can be used in predictive genomic selection (GS) methodologies. As first proposed by Meuwissen *et al.* (2001), breeding values can be predicted as the sum of all marker effects by regressing phenotypic

values on all available markers. In animals, the first to implement GS was the US dairy industry (vanRaden, 2007; VanRaden *et al.*, 2008); later Hayes *et al.* (2009) and Goddard and Hayes (2009) successfully implemented GS in other animal species. Daetwyler *et al.* (2008) showed how to use genomic prediction for analyzing the genetic risk of human diseases, and Bernardo and Yu (2007), Lorenzana and Bernardo (2009) and Heffner *et al.* (2011) showed promising results of GS in plants.

In plants, several authors have used breeding data with individuals genotyped with intermediate- to high-density markers to show that traits such as grain yield, biomass yield, disease resistance and flowering evaluated under different environmental conditions can be predicted with varying levels of accuracy depending on, among other factors, the heritability of the trait, the size of the training population, the number of markers, the relationship between the training and testing sets, and genotype  $\times$  environment interaction (GE) (de los Campos *et al.*, 2009; Crossa *et al.*, 2010, 2011; Pérez *et al.*, 2010, 2012; Burgueño *et al.*, 2012; Gonzalez-Camacho *et al.*, 2012; Hickey *et al.*, 2012; Riedelsheimer *et al.*, 2012; Windhausen *et al.*, 2012; Zhao *et al.*, 2012).

The main results of most prediction studies carried out on maize and wheat trials at International Maize and Wheat Improvement

<sup>1</sup>Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Mexico, Mexico; <sup>2</sup>Colegio de Postgraduados, Montecillo, Mexico; <sup>3</sup>School of Environmental and Rural Science, University of New England, Armidale, New South Wales, Australia and <sup>4</sup>French-Argentine International Center for Information and Systems Sciences (CIFASIS), Rosario, Argentina

Correspondence: Dr J Crossa, Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, Mexico DF 06600, Mexico  
Received 16 August 2012; revised 20 December 2012; accepted 21 January 2013; published online 10 April 2013

Center (CIMMYT) show that when the aim is to predict the genetic values of individuals in a population (with unknown subpopulations), the use of intermediate- to high-density markers, along with pedigree information, produces diverse prediction accuracy of the genetic values of unobserved individuals, with pedigree accounting for a large proportion of predictive ability. Pedigree (when available) is the baseline model that considers family links (or breeders' structures) and, when used together with marker information, increases the prediction accuracy for all traits most of the time.

General findings of CIMMYT studies show that the accuracy of genotype imputation from low- to high-marker density is high for large training sets and for individuals who have close relatives in the training set even if the training set is relatively small; this offers a valuable economic option for imputing high-density markers to large sample populations from small sample populations that were genotyped with high-density markers. However, prediction accuracy between genetically unrelated populations (or families) is low, when using any marker density; likewise, when heritability is low, genomic information does not help prediction accuracy.

In terms of the algorithms used to make predictions in CIMMYT trials, no one prediction model fits all situations (trait–environment combinations), yet empirical evidence shows that models that include nonlinear functions of the markers, when used on complex traits (that is, grain yield), tend to increase prediction ability as compared with linear models. For less complex traits that are predominantly additive (that is, male and female flowering in maize and resistance to rust diseases in wheat), linear models gave the best prediction accuracy.

In a recent article using data from a highly structured CIMMYT maize population, Windhausen *et al.* (2012) showed that when predicting between unrelated subpopulations, accuracy was close to zero, and that almost all cross-validation accuracy achieved by marker-based prediction was essentially the result of assigning lines to subpopulations with differing trait means; identical accuracies were obtained when prediction was simply based on breeder knowledge of structural groupings, rather than on marker-based prediction. Windhausen *et al.* (2012) discussed several possible applications of GS to maize breeding programs, that is, prediction within progenies of the same cross and prediction of progenies of different crosses. Prediction within the same cross implies training and validating within, for example, F<sub>2</sub> individuals or doubled haploids or advanced inbred lines (that is, F<sub>6</sub>) derived from one cross for which phenotyping in multi-environment trials is more expensive than genotyping. Another prediction problem that usually arises in practical breeding is predicting individuals within and between different crosses, that is, within and between F<sub>2</sub> individuals (or their doubled haploids). It is expected that factors such as the degree of relatedness of individuals between crosses, and whether the LD between markers and QTL is stable across different crosses, will affect the prediction accuracy between crosses.

A similar structural paradigm should be considered for genomic prediction in multi-environment plant breeding trials. GE produces environments that are structured into related or unrelated subsets that may increase or decrease prediction ability. In plants, prediction accuracy has been estimated by evaluating training and validation sets in single environments or in a subset of similar environments (within optimum or drought environments). Burgueño *et al.* (2012) studied the effect of modeling GE (or not) on genomic prediction ability when using marker and pedigree information. They found that multi-environment models benefit from borrowing information from correlated environments and from using information regarding pedigree and genetic markers. Thus, when the environmental structure (GE) was modeled, this increased the prediction ability of

unobserved individuals by about 20% with respect to the single-environment prediction model. Windhausen *et al.* (2012) indicated that the impact on prediction accuracy of environmental structure in combination with population structure should be considered for further research.

Two basic applications of GS are being studied in CIMMYT's global maize and wheat breeding programs. One application consists of predicting the genotypic values of individuals for potential release as cultivars. In this application, both additive and non-additive effects are important for determining the final commercial value of the maize or wheat lines to be released in the near future. The other application is concerned with predicting the breeding value of candidates in rapid-cycle populations (that is, focusing on the precise prediction of additive effects in the breeding program's early generations).

The main purpose of this article is to describe and summarize the results of genomic prediction in CIMMYT's maize and wheat breeding programs starting from the initial assessment of the predictive ability of different models using pedigree and marker information to the present, when methods for implementing GS in practical global maize and wheat breeding programs are being studied and investigated. The article is organized as follows: first, we briefly summarize the genomic prediction results obtained using several CIMMYT maize and wheat breeding data sets and different models for predicting genetic values of lines in a breeding population in a single environment and/or across environments (drought or optimum). We studied the effect of modeling GE (structuring environments) on the prediction ability of a wheat multi-environment trial using pedigree and marker data. Second, we show the results of predicting between related and/or unrelated populations; predicting between unrelated populations in CIMMYT breeding populations has been unsuccessful. Third, we present prediction ability within two biparental F<sub>2</sub> crosses in maize when a small number of lines and markers are considered.

## MATERIALS AND METHODS

### Models for genomic prediction

The standard linear genetic model considers that the phenotypic response of the *i*th individual ( $y_i$ ) is explained by a factor common to all individuals, a genetic factor specific to that individual ( $g_i$ ) and the residual comprising all other non-genetic factors ( $\varepsilon_i$ ), including environmental effects (temporal or spatial) and effects described by the design of the experiment, among others. Thus, the linear genetic model for  $n$  genotypes ( $i = 1, \dots, n$ ) is represented by  $y_i = g_i + \varepsilon_i$ . In this standard linear genetic model, the genetic factor  $g_i$  can be described by adding the molecular marker effects or by using pedigree, or both. When  $g_i$  is defined as a parametric linear regression on marker covariates  $x_{ij}$  (for example, 0, 1 and 2 for additive effects), then  $g_i = \sum_{j=1}^p x_{ij}\beta_j$ , such that

$$y_i = \sum_{j=1}^p x_{ij}\beta_j + \varepsilon_i \quad (1)$$

where  $\beta_j$  is the effect of the *j*th marker covariate on  $y_i$  (Meuwissen *et al.*, 2001) ( $j = 1, 2, \dots, p$ , markers). Assuming that the distribution of the residuals is normal with mean zero and variance  $\sigma_\varepsilon^2$ , then the likelihood of model (1) is

$$p(\mathbf{y}|\mathbf{g}, \sigma_\varepsilon^2) = \prod_{i=1}^n N\left(y_i \mid \sum_{j=1}^p x_{ij}\beta_j, \sigma_\varepsilon^2\right)$$

where  $N\left(y_i \mid \sum_{j=1}^p x_{ij}\beta_j, \sigma_\varepsilon^2\right)$  denotes a normal density for random variable  $y_i$  centered at  $\sum_{j=1}^p x_{ij}\beta_j$  and with variance  $\sigma_\varepsilon^2$ .

The large variation resulting from hundreds of thousands of markers can be controlled by various shrinkage methods that can be formulated in frequentist or Bayesian models. Depending on how priors on the marker effects are assigned, different Bayesian linear regression models can also be derived. We

considered models for GS that differ in the information used (pedigree, molecular markers or both) and in the way markers are incorporated into the model. We used parametric linear regression models, for example, the Bayesian LASSO (BL) and the GBLUP, as well as semiparametric nonlinear models, that is, Reproducing Kernel Hilbert space (RKHS) and the Neural Network methods with a Radial Basis Function (RBFNN) and Bayesian regularization. The pedigree model, BL and GBLUP are explained in detail in several articles, for example, de los Campos *et al.* (2009), Crossa *et al.* (2010; 2011) and Pérez *et al.* (2010). The GBLUP models using pedigree and genomic information while modeling GE using factor analytic (FA) model information are described in Burguño *et al.* (2012). Detailed descriptions of the nonlinear models RKHS and RBFNN are given in de los Campos *et al.* (2009), Gonzalez-Camacho *et al.* (2012) and Pérez *et al.* (2012). A brief description of these and other models is given in the Appendix.

### Models accounting for epistasis

There is concrete proof that the agglomeration of multiple gene  $\times$  gene interactions (epistasis) having small effects and acting in small epistatic networks is important for explaining the heritability of complex traits in genome-wide association studies (McKinney and Pajewski, 2012). Epistasis networks and methods with connections based on gene  $\times$  gene interactions can also be integrated for GS using statistical genetic models that consider these network complexities. Evidence from studies conducted at CIMMYT for complex traits, such as grain yield, shows that models that allow for nonlinear components consistently predicted the unobserved individuals better than linear models (Crossa *et al.*, 2010; 2011; Gonzalez-Camacho *et al.*, 2012; Pérez *et al.*, 2012). In principle, epistasis can be incorporated when the number of markers is large, but the large number of epistatic effects (two-locus interaction) makes their use unfeasible. The RKHS regression approach and the RBFNN implicitly map marker effects into a high-dimensional feature space that utilizes non-additive genetic effects. Recently, Wang *et al.* (2012) developed an approach for incorporating the large number of markers used in GS into the adaptive mixed LASSO (Wang *et al.*, 2011), which accounts for two-locus epistasis effects for predicting genetic values.

In a recent study, González-Camacho *et al.* (2012) analyzed real and simulated data by Zhang and Xu (2005). Simulated data have a sample size of 600 individuals and a genome with only a single chromosome (1800 cm long) and 121 evenly spaced markers with a 15 cm per marker interval. The authors simulated 9 main QTL effects and 13 interactions between different QTL effects; all QTL effects overlapped with markers. Each QTL made a contribution to phenotypic variance that varied from 0.5 to 20%. Results show that the prediction ability of RKHS and RBFNN was superior to that of the BL, indicating that RKHS and RBFNN are able to capture patterns (for example, gene  $\times$  gene effects) that cannot be captured by a linear model. When marker main effects and two-marker interactions were fitted, the performance of the linear model increased sharply and that of the semiparametric methods decreased, indicating that, for nonlinear models, information on gene  $\times$  gene interaction incorporated into the input space becomes redundant in the feature space, whereas for the linear models, information on gene  $\times$  gene interaction incorporated in the input space is useful for predicting the feature space. The linear model was able to detect this via estimates of regression coefficients that weigh the contribution of each marker to the estimated conditional expectation. On the other hand, in both RBFNN and RKHS, each marker gets a similar weight in the basis function or kernel, and the effect of adding non-signal covariates reduces method performance.

### Assessing the models' predictive ability

The predictive ability of the models was computed as a Pearson correlation between predicted and observed values. Commonly, phenotypic and genotypic data comprise inbred lines that are either breeders' candidates for selection or segregating populations at different levels of inbreeding. Phenotypic data may include candidates for selection from different years tested in different environments.

Two types of partition designs were used for the data sets (training and testing sets). The first partition design considered a cross-validation that divides the data into a training sample and a validation sample or testing set.

The models are fitted using the training sample, and the fitted models are used to predict outcomes in the validation sample. Here, the data set is divided into  $k$  groups; this is done by assigning observations  $\{i = 1, \dots, n\}$  to  $k$  disjoint sets  $\{S_1, \dots, S_k\}$ . Each of these sets can then be used to measure predictive ability. For example, using the first set, the data can be divided so that the training set contains all the observations in  $\{S_2, \dots, S_k\}$ , and the testing set those in  $S_1$ . Subsequently, models are fitted using data in  $\{S_2, \dots, S_k\}$ , and the fitted models are used to obtain predictions for observations in  $\{S_1\}$ , that is,  $\{y_i; i \in S_1\}$ . Repeating this exercise for the second, third, ...,  $k$ th sets yields a whole set of CV predictions  $\{\hat{y}_i\}_{i=1}^n$  that can be compared with actual observations  $\{y_i\}_{i=1}^n$  to assess predictive ability. In general, we used a 10-fold cross-validation scheme.

In addition to using a  $k$ -fold cross-validation partition layout, we used a second partition design for splitting the data into training and testing sets that consists of 50 random partitions, with each partition randomly assigning 90% (or less) of the lines to the training set and the remaining 10% (or more) of the lines to the testing set. This partition scheme is similar to that used by Gonzalez-Camacho *et al.* (2012) and Gianola *et al.* (2011), and helps obtain a distribution of the correlations in the testing set for each tested model.

## RESULTS

### Predicting genetic values of lines in a breeding population

The main objective of these initial studies was to compare the prediction ability of various parametric and semiparametric models using intermediate- to high-marker density for assessing the genomic prediction of genetic values in real plant breeding populations evaluated under different environmental conditions. The maize data used consist of diverse inbred lines (that is, tropical, subtropical, tolerant to acid soils, and so on) from CIMMYT's Global Maize Program that were evaluated for several traits in different environments and genotyped with low and high marker density. The wheat lines in one data set are selection candidates evaluated for grain yield and other traits in several site-year combinations (Pérez *et al.*, 2012), whereas the other data set comprises a set of historical wheat lines from CIMMYT's Global Breeding Program evaluated in several mega environments (Crossa *et al.*, 2010; Burguño *et al.*, 2012).

*Maize.* A summary of the results is presented in Table 1 for the mean correlation of 10-fold cross-validation (Crossa *et al.*, 2010) and 50 random partitions (Gonzalez-Camacho *et al.*, 2012), each with 90–100% for the training and testing sets using two-marker densities in a maize population of 284 inbred lines evaluated for various trait-environment combinations and genotyped with 1148 and 55 000 SNPs. Broad-sense heritabilities of the trait-environment combinations were relatively high for male and female flowering traits, and intermediate for grain yield; these estimates were consistently higher under well-watered (WW) conditions than under severe drought stress (SS). Results indicate that increasing the marker density increased the prediction ability of the models for most trait-environment combinations. It is clear that prediction for all traits under both marker densities is always more problematic under SS than under WW conditions. This effect is clearer for a complex trait such as GY under stress conditions (GY-SS or GY-LOW), compared with GY under optimum conditions (GY-WW or GY-HIGH). This result confirms the importance of improving field evaluations under stress conditions (that is, drought and/or heat) by using appropriate experimental designs and statistical models that account for uneven soil characteristics and other environmental trends occurring during the crop cycle.

For more simple traits, such as male and female flowering, the effect of stress conditions decreased prediction ability, but this ability was higher when using higher marker density than with intermediate marker density. Another trend that emerged from these results is that

**Table 1** Mean correlations between predicted and observed values of the three models, BL, RKHS regression and RBFNN for four traits (female flowering, FFL; male flowering, MFL; grain yield, GY; anthesis-silking interval, ASI) measured in the following environments: WW, SS, HIGH and LOW in 284 maize inbred lines genotyped with 55 000 and 1148 SNPs

Trait-environment (heritability)	BL	RKHS	RBFNN
<i>55 000 SNPs<sup>a</sup></i>			
FFL-WW (0.89)	0.814	<u>0.836</u>	0.834
FFL-SS (0.81)	0.754	<u>0.763</u>	0.757
MFL-WW (0.88)	0.817	<u>0.841</u>	0.832
MFL-SS (0.83)	0.776	<u>0.782</u>	0.780
ASI-WW (0.79)	0.582	<u>0.586</u>	<u>0.594</u>
ASI-SS (0.77)	0.612	<u>0.621</u>	0.605
GY-WW (0.49)	<u>0.557</u>	0.548	0.529
GY-SS (0.38)	0.326	<u>0.330</u>	0.288
GY-HI (0.50)	0.633	<u>0.663</u>	0.653
GY-LOW (0.40)	<u>0.410</u>	0.402	0.393
<i>1148 SNPs<sup>b</sup></i>			
FFL-WW	<u>0.781</u>	0.588	—
FFL-SS	<u>0.774</u>	0.648	—
MFL-WW	<u>0.790</u>	0.607	—
MFL-SS	<u>0.778</u>	0.674	—
ASI-WW	0.513	<u>0.547</u>	—
ASI-SS	0.517	<u>0.572</u>	—
GY-WW	<u>0.525</u>	0.514	—
GY-SS	0.415	<u>0.453</u>	—
GY-HI	—	—	—
GY-LOW	—	—	—

Abbreviation: BL, Bayesian LASSO; HI, optimum; LOW, stress; RBFNN, radial basis function neural network; RKHS, reproducing kernel Hilbert space; SNP, single-nucleotide polymorphism; SS, severe drought stress; WW, well-watered.  
The best prediction model for each trait-environment combination in the data set is underlined.  
<sup>a</sup>Extracted from Gonzalez-Camacho *et al.* (2012).  
<sup>b</sup>Extracted from Crossa *et al.* (2010).

**Table 2** Mean correlations between observed and predicted values for GY using GBLUP; P; Bayesian LASSO with markers, and with markers + pedigree (MBL, PMBL, respectively); RKHS regression with markers, and with marker + pedigree (MRKHS and PMRKHS, respectively); MRBFNN, and MBRNN models in two data sets, one with 306 wheat lines genotyped with 1717 DArTs markers and evaluated in seven environments, and the other with 599 wheat lines genotyped with 1279 DArTs and evaluated in four environments

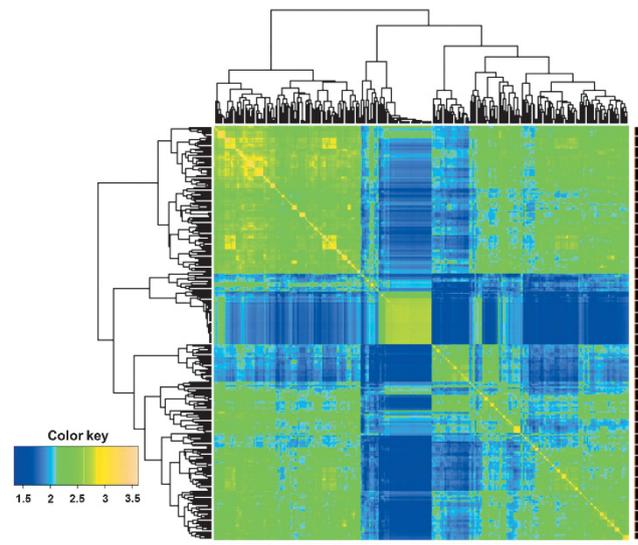
Trait	Environment	Model				
		GBLUP	MBL	MRKHS	MRBFNN	MBRNN
<i>306 lines and 1717 DArT markers<sup>a</sup></i>						
GY	1	0.43	0.48	<u>0.51</u>	<u>0.51</u>	0.50
	2	0.41	0.48	<u>0.50</u>	<u>0.43</u>	0.43
	3	0.29	0.20	0.37	<u>0.42</u>	0.32
	4	0.46	0.45	0.53	<u>0.55</u>	0.49
	5	0.56	0.59	0.64	<u>0.66</u>	0.63
	6	0.67	0.70	<u>0.73</u>	<u>0.71</u>	0.69
	7	0.50	0.46	0.53	<u>0.54</u>	0.50
<i>599 lines and 1279 DArT markers<sup>b</sup></i>						
Trait	Environment	P	MBL	MRKHS	PMBL	PMRKHS
GY	1	0.45	0.52	0.60	0.54	<u>0.61</u>
	2	0.42	0.49	0.49	<u>0.50</u>	0.49
	3	0.42	0.40	<u>0.45</u>	<u>0.45</u>	0.40
	4	0.45	0.46	<u>0.52</u>	0.50	0.46

Abbreviations: BRNN, Bayesian regularized neural networks; GBLUP, genomic best linear unbiased predictor; GY, grain yield; MBRNN, Bayesian regularized neural networks with markers; MRBFNN, radial basis function neural networks with markers; P, pedigree; RBFNN, radial basis function neural networks; RKHS, reproducing kernel Hilbert space.  
The best prediction model for each environment in the data set is underlined.  
<sup>a</sup>Extracted from Pérez *et al.* (2012).  
<sup>b</sup>Extracted from Crossa *et al.* (2010).

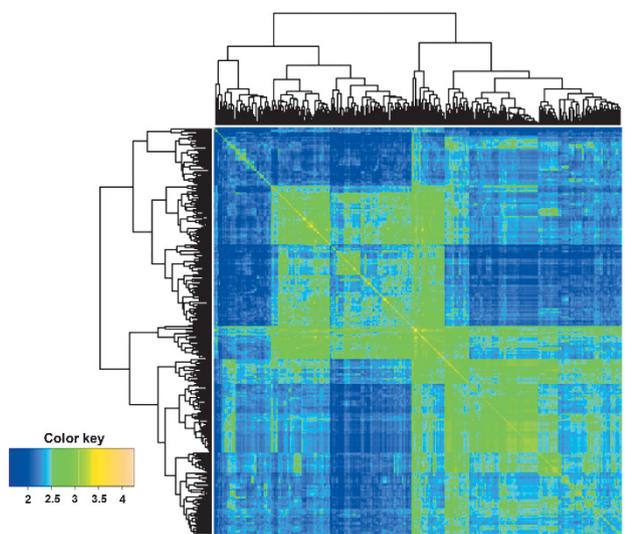
with low-density markers and for simple traits (FFL and MFFL), the predictive ability of the linear model (BL) surpassed those of the nonlinear models (RKHS and RBFNN), whereas with high-density markers, nonlinear models always showed better prediction ability than the linear models. One can speculate that nonlinear models can capture small cryptic epistatic effects better with high marker density, and they cannot be captured with low marker density. Colinearity between high-density markers is another possible reason for the poorer prediction of the linear model compared with the nonlinear model. These results indicate that imputation from low marker density to high marker density, shown by Hickey *et al.* (2012) to be very accurate in maize, is important in the practical implementation of GS, because imputation from low to high marker density will generate economic savings during the breeding cycles while increasing prediction accuracy.

*Wheat.* The summary in Table 2 shows the prediction ability of two wheat data sets for grain yield measured in various environments. Here, the number of genotypes in each data set is 306 and 599, and the number of markers is 1717 and 1279, respectively (Pérez *et al.*, 2012; Burguño *et al.*, 2012). Prediction accuracies were highly consistent in both sets of experiments, with BL with markers and

pedigree (PMBL), and RKHS with pedigree and markers (PMRKHS) giving the best predictions on the data set of 599 wheat lines and with models MRKHS and MRBFNN as the best predictors on the data set



**Figure 1** Heat map of the G matrix for the data set with 306 wheat lines genotyped with 1717 DArTs markers.



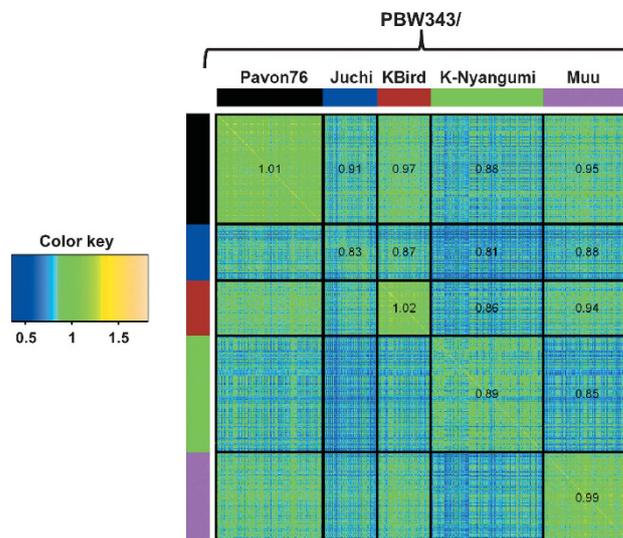
**Figure 2** Heat map of the **G** matrix for the data set with 599 wheat lines genotyped with 1279 DArTs markers.

of 306 wheat lines. Figures 1 and 2 depict the heatmap of the genomic (**G**) matrix for data on 306 and 599 wheat lines, respectively. The 306 lines comprise three groups, one large population (top left), one small subset unrelated to either of the other two and another large population that is closely related to the first large one, apparently with two closely related subgroups (Figure 1). The 599 wheat lines formed two clear large groups, each with several subgroups closely related to each other (Figure 2); the two large subgroups overlap slightly.

The prediction problem addressed in this research concerned the prediction of the entire population. It is interesting to note that markers consistently increased prediction ability over the baseline pedigree-derived model in all environments of both data sets. Another clear result emerging from these two data sets is that considering markers and pedigree together consistently increased the prediction ability of all models, as compared with models with markers only or pedigree only. For other problems (such as prediction between subpopulations), a decrease in prediction ability was expected, especially for subpopulations that are not related, for example, those shown in Figure 1. When one of the large subpopulations in Figure 2 was predicted based on the other large subpopulation, prediction accuracy decreased, as expected (de los Campos, personal communication).

### Predicting genetic values of lines between populations or subpopulations

**Wheat.** In a recent study, we used a stem rust data set from five wheat populations (PBW343/Juchi, PBW343/Pavon76, PBW343/Muu, PBW343/Kingbird and PBW343/K-Nyangumi) for genomic prediction (Ornella *et al.*, 2012). Resistance to stem rust is known to be affected by several major genes and also influenced by slow-rusting genes with small additive effects. All populations were derived from crosses between resistant parents (Juchi, Pavon76, Muu, Kingbird and K-Nyangumi) and PBW343, a moderately susceptible parent. The parents and the  $F_6$  generation recombinant inbred lines were evaluated for reaction to stem rust at different locations. The sample sizes of the five wheat populations were 92 genotypes in PBW343/Juchi, 176 individuals in PBW343/Pavon76, 148 genotypes in



**Figure 3** Heat map of the genomic relationship matrix **G** of five wheat populations: PBW343/Pavon76, PBW343/Juchi, PBW343/Kingbird, PBW343/K-Nyangumi and PBW343/Muu. The numbers indicate the average values of the corresponding elements of **G** within and between populations (from Ornella *et al.*, 2012).

**Table 3** Pair-wise correlations between observed and predicted stem rust values of two models, Bayesian LASSO and the GBLUP, trained in one population and evaluated in the other population for five populations (adapted from Ornella *et al.*, 2012)

	Training <sup>a</sup>					
	PBW343/ Juchi	PBW343/ Kingbird	PBW343/K- Nyangumi	PBW343/ Muu	PBW343/ Pavon76	
Testing						
PBW343/ Juchi	—	0.48	0.14	0.28	0.31	Bayes LASSO
PBW343/ Kingbird	0.53	—	0.29	0.25	0.54	
PBW343/ K-Nyangumi	0.14	0.30	—	0.28	0.28	
PBW343/ Muu	0.18	0.30	0.33	—	0.29	
PBW343/ Pavon76	0.37	0.51	0.22	0.33	—	
GBLUP						

There are five related populations: PBW343/Juchi, PBW343/Kingbird, PBW343/K-Nyangumi, PBW343/Muu and PBW343/Pavon76.

<sup>a</sup>The triangle on the upper-right shows the prediction ability (correlation) of Bayes LASSO, with the rows indicating the training population (that is, PBW343/Juchi) and the columns the testing population (that is, PBW343/Kingbird, 0.48); the triangle on the lower-left gives the prediction ability of GBLUP, with the columns indicating the training population (that is, PBW343/Juchi) and the rows the testing population (that is, PBW343/Kingbird, 0.53).

PBW343/Muu, 90 individuals in PBW343/Kingbird and 176 genotypes in PBW343/K-Nyangumi. Genotypes were molecularly characterized using 1400 Diversity Arrays Technology Pty. Ltd (<http://www.diversityarrays.com/>) (DArT markers).

As depicted in Figure 3, there are five clearly distinct yet related (half-sib) populations; however, lines in the Juchi population do not

**Table 4** Correlations between predicted and observed stem rust values for five wheat populations, when four of them, with different numbers of individuals in the training set, predict each of the others in the testing set using the GBLUP and the Bayesian LASSO (BL) models

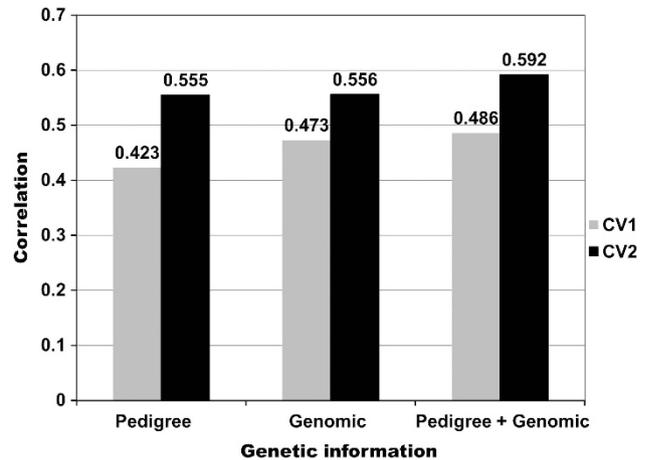
Population	Number of lines in the training set	Number of lines in the testing set	GBLUP	BL
<i>Four populations predicting one population</i>				
PBW343/Muu	534	148	0.59	0.60
PBW343/K-Nyangumi	506	176	0.52	0.53
PBW343/Kingbird	592	90	0.79	0.82
PBW343/Juchi	590	92	0.41	0.39
PBW343/F6Pavon	506	176	0.59	0.62
<i>All populations 50 random partitions with 90:10</i>				
All populations	612	70	0.62	0.64

Mean correlations when considering all five populations combined with a prediction design that considers 50 random partitions with a split of 9:1 for the training set: testing set proportion.

seem to be closely related. Results of the pairwise prediction of one population using the other are shown in Table 3. The accuracy of predictions of one population using another population was relatively high, except for PBW343/K-Nyangumi when predicted by PBW343/Juchi (using BL, 0.14) and PBW343/Juchi when predicted by PBW343/K-Nyangumi (using GBLUP, 0.14). Furthermore, Table 4 indicates that prediction of stem rust data in each individual populations using stem rust data from the other four populations gave relatively high correlations, except for population PBW343/Juchi, which, as shown by Ornella *et al.* (2012), does not have lines that are closely related among themselves or to lines in other populations.

It is interesting to point out that even under the population structure existing in this wheat data set, stem rust prediction between populations, and also within the meta-population comprising all five populations (Table 4), was relatively high. The architecture of the traits has an important role that may help to explain these results. For more complex traits, such as grain yield, this prediction result may not hold; thus population structure could become the main driving force for increasing the prediction accuracy of the meta-population, while prediction between populations significantly decreases.

**Maize.** Recently, Windhausen *et al.* (2012) investigated various aspects of genomic prediction within and between populations, and discussed the potential uses of genomic prediction in maize hybrid development. The authors used 255 diverse maize lines tested in six trials (almost all identical to those used by Crossa *et al.* (2010)), and classified them into eight breeding populations differing in mean performance. The authors examined the problem of separately predicting each breeding population based on marker effects estimated in the other populations; prediction ability was nearly zero. Results suggest that in the current study, prediction comes mainly from pedigree (subpopulation structure within the global set), with negligible contributions from the relationship between the training and validation sets or from LD between markers and causal variants underlying the predicted traits. Potential uses of genomic prediction in maize hybrid development are discussed, emphasizing the need to define the breeding scenario in which genomic prediction should be applied (that is, prediction across or within populations) and the



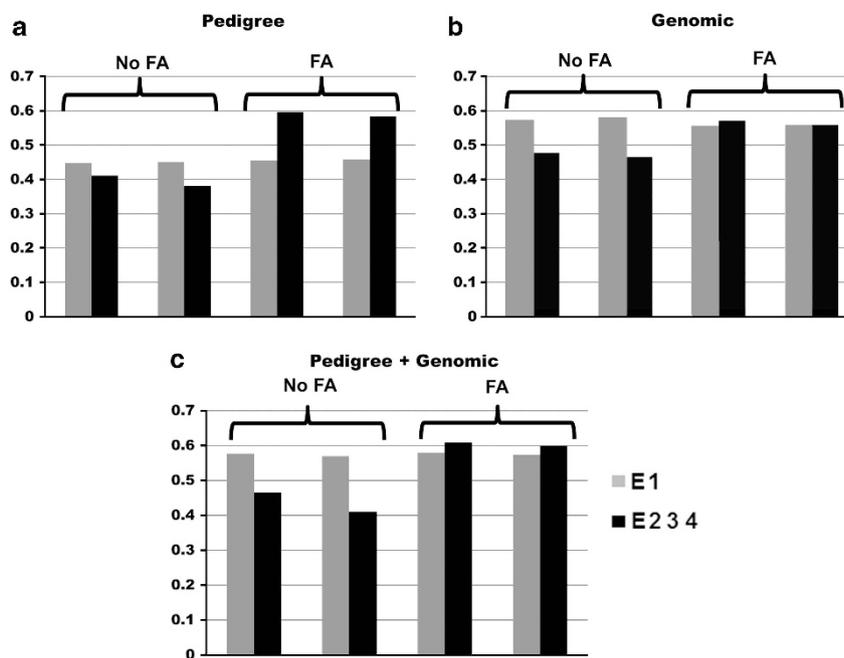
**Figure 4** Mean correlations (across four environments) between predicted and observed grain yield values derived from models using only pedigree, only genomics and pedigree+genomic for two cross-validation schemes (CV1 and CV2) (adapted from Burgueño *et al.*, 2012). Cross-validation CV1 predicts genotypes that have never been evaluated in any environment, and cross-validation CV2 predicts genotypes that were evaluated in some environments but not in other environments.

size of the training sets with a strong genetic relationship to the validation set.

#### Prediction within breeding populations incorporating GE, pedigree and genomic information

**Pedigree- and genomic-derived additive predictions.** Previous results referred to genomic predictions of models for a single environment and did not consider correlated environmental structures due to GE. Crossa *et al.* (2010) evaluated the predictive ability of pedigree, genomics and pedigree + genomic information in single-environment models such as those used for the wheat multi-environment data set. Here, we summarize results from Burgueño *et al.* (2012), who, using the same data set as Crossa *et al.* (2010), were the first to evaluate the impact of modeling GE covariance structures for multi-environment trials. The data comprise 599 wheat lines genotyped with 1279 DArTs and evaluated in four wheat mega environments. Crossa *et al.* (2010) used pedigree, genomic and pedigree + genomic-derived models for a wheat multi-environment trial comprising four environments (E1–E4) with different genetic correlations, E1 not being genetically correlated with the others, and E2 and E4 with intermediate to high genetic correlations. Crossa *et al.* (2010) assessed prediction problems relevant to plant breeders: (1) predicting the performance of genotypes that have not been evaluated in any environment (assessed by cross-validation CV1), and (2) predicting the performance of genotypes that have been evaluated in some environments, but not in others (assessed by cross-validation CV2). Prediction was performed for lines within environments and across environments.

Burgueño *et al.* (2012) found that the predictive ability of genomic-based models was higher than that of pedigree-based models. Their results confirmed the superiority of pedigree + genomic models for GS over pedigree-based predictions or genomic-based predictions alone. Predicting the performance of newly developed lines that have never been evaluated in the field (CV1) is more challenging than predicting the performance of lines that have been evaluated in different but correlated environments (CV2). Figure 4 shows correlations obtained from multi-environment models that model GE using the FA model in two cross-validations schemes (CV1 and CV2).



**Figure 5** Correlations between predicted and observed performance in environment 1 (E1) and average of environments 2, 3 and 4 (E2 3 4) obtained in CV2 using only pedigree (a), only genomics (b) or using pedigree+genomics (c)-based models with different specifications for the residual and genetic covariance matrices (FA = GE modeled using the factor analytic model; no FA = GE not modeled) (adapted from Burgueño *et al.*, 2012).

Correlations for CV1 do not change much and those in CV2 were 31, 17.5 and 21.8% greater than those obtained in CV1, indicating the importance of having information from correlated environments when predicting performance.

Figure 5 shows that the impact of modeling GE in CV2 is marked in environments E2–E4, but not in E1; this is because genetic values in E2–E4 have high genetic correlations, whereas genetic values in E1 exhibit low genetic correlations with those from E2–E4. For correlated environments E2–E4, the benefits in predictive ability come from borrowing information from correlated environments by modeling GE and by using information regarding pedigree and genomic relationships. Interestingly, Burgueño *et al.* (2012) also examined the predictive accuracy of models without using pedigree and genomic relationships, and showed that to predict the environments that cause a great deal of the GE, such as environment E1 (and E4), modeling GE using information on genomic relationships and/or pedigree gives better prediction accuracy than when the GE is not modeled and the information on genomic and pedigree is ignored.

In general, these results also agree with those reported by Burgueño *et al.* (2011), who demonstrated, without any pedigree and genomic information, that modeling GE based solely on phenotypic data is a good thing, as it always gives better predictions than simple linear mixed models without modeling GE.

*Pedigree- and genomic-derived additive and additive × additive predictions.* As proposed by Burgueño *et al.* (2007), pedigree-derived additive and additive × additive relationship information can be combined into a single model that will model GE using the FA model with two random effects: one is a regression on pedigree additive relationships, and the other, a regression on pedigree epistasis additive × additive relationships (see Appendix). Thus, genomic-derived additive and additive × additive relationships can be combined into a single model by extending them into one with two

**Table 5** Mean correlations between the predicted and observed values of genotypes for individual environments (E1, E2, E3 and E4), and for three environments combined (E2, E3 and E4) using five different factor analytic models (FA) for a cross-validation scheme (CV2), each with 10-fold cross-validation

	FA additive pedigree ( $G_{OP}$ )	FA additive genomic ( $G_{OM}$ )	FA pedigree additive ( $G_{OP}$ ), FA genomics additive ( $G_{OM}$ )	FA pedigree additive × additive ( $G_{OPP}$ )	FA genomic additive ( $G_{OM}$ ), FA genomic additive × additive ( $G_{OMM}$ )
E1	0.460	0.552	0.469	0.512	<u>0.554</u>
E2	0.623	0.609	<u>0.638</u>	0.612	0.637
E3	0.633	0.581	0.612	<u>0.634</u>	0.603
E4	0.533	0.513	0.448	<u>0.540</u>	0.465
E2, E3, E4	0.596	0.571	0.566	<u>0.598</u>	0.568

The best predictive model for each environment or environment combination is underlined. Data used were extracted from Crossa *et al.* (2010) and Burgueño *et al.* (2012). Factor analytic-pedigree for  $G_{OP}$ , factor analytic-genomic for  $G_{OM}$ , factor analytic-pedigree ( $G_{OP}$ ) and factor analytic-genomic ( $G_{OM}$ ); factor analytic-pedigree additive ( $G_{OP}$ ) and factor analytic-pedigree additive × additive ( $G_{OPP}$ ); factor analytic-genomic additive ( $G_{OM}$ ) and factor analytic-genomic additive × additive ( $G_{OMM}$ ).

random effects: one with regression on genomic additive relationships, and the other, regression on genomic epistasis additive × additive relationships. We have assessed these predictions to examine whether the inclusion of epistasis will improve prediction ability as compared with using additive effects alone.

Results in Table 5 are for cross-validation CV2 that predicts the genetic performance of 599 lines that were tested in some

environments but not in others. Results show that including pedigree additive relationships and pedigree epistasis additive  $\times$  additive relationships did improve prediction in some environments or combinations of environments. Also, genomic additive relationships and genomic epistasis additive  $\times$  additive relationships did improve predictions in most environments (and their combinations), as compared with models that only include pedigree additive or genomic additive relationships. These results indicate that modeling additive  $\times$  additive epistasis along with GE is important for increasing the accuracy of predictions in wheat breeding populations. These findings agree with those recently reported by Wang *et al.* (2012), who found that modeling epistasis did increase predictive ability.

### Prediction within bi-parental maize populations

One of the key ways in which GS can be used in a maize breeding program is to improve source populations from which candidate inbred lines can be predicted and selected as parents for the next generation; in this manner, genetic gains in a closed multi-parent population derived from three or more elite lines are maximized. It is important to know whether greater gains per year could be achieved at a lower cost and in less time using this approach than using the conventional pedigree breeding approach for the same investment. Prediction within full-sib families is the most favorable situation for GS, with very high LD and no pedigree, family or group structure; therefore, we consider the accuracies estimated for the bi-parental populations as the maximum obtainable in closed rapid-cycle marker-only selection.

In this study, we show results of two bi-parental populations comprising 249 and 250  $F_2$  test-cross individuals, respectively, that were genotyped with 238 and 271 SNPs, respectively, and evaluated under different drought and optimum environmental conditions. Table 6 shows the predictive ability for grain yield across environments measured as a Pearson correlation between the observed value ( $y$ ) and the predicted value ( $\hat{y}$ ) averaged over all 50 validation runs for each of the different combinations of number of individuals and number of markers for two maize bi-parental populations. Broad-sense heritability for grain yield was 0.57 for bi-parental population 1 and 0.28 for bi-parental population 2 when all drought and optimum environments were combined. Prediction accuracy in these bi-parental populations reached around 0.4 at most, when 90 individuals were randomly taken (50 times) as the training population from the entire sample of bi-parental population 2. When 50 individuals were sampled, prediction accuracy dropped to around 0.3. Although research is needed to balance the genotypic and phenotypic costs, it is possible to speculate that these accuracies can still be considered adequate, as they will save costs when large numbers of bi-parental populations are evaluated in the field, and have the additional benefit of potentially increasing selection intensity. Predictions in Table 6 indicate that decreasing the number of markers also has a negative impact on prediction accuracy. However, this effect is much more drastic in bi-parental population 1 (with a heritability for grain yield of 0.27) than in bi-parental population 2 (with a heritability for grain yield of 0.57). Prediction accuracy can be computed by dividing the prediction ability by the square root of the broad-sense heritability of the target trait evaluated in the respective set of environments.

### DISCUSSION

GS has a number of uses in a breeding program; however, its greatest potential use is at points in the breeding program where selection using traditional methods (for example, through the generation of phenotypes via replicated trials) is too expensive, time consuming,

**Table 6** Mean correlations between the predicted grain yield values for three models (BL, RKHS regression and GBLUP) when the numbers of individuals in the training sets randomly taken from the entire population (50 different times) are 30, 40, 50, 70 and 90 for different numbers of SNPs

Model_ number of individuals in the training population	Number of SNPs		
	238 SNPs	100 SNPs	50 SNPs
<i>Bi-parental population 1 (249 <math>F_2</math> individuals; 238 SNPs)</i>			
BL_30	0.1348	0.1009	0.0640
RKHS-KA_30	0.1520	0.1144	0.0703
GBLUP_30	0.1643	0.1173	0.0672
BL_40	0.1539	0.1286	0.0899
RKHS-KA_40	0.1685	0.1380	0.0925
GBLUP_40	0.1789	0.1441	0.0906
BL_50	0.2093	0.1598	0.0918
RKHS-KA_50	0.2004	0.1555	0.0944
GBLUP_50	0.2165	0.1612	—
BL_70	0.2236	0.1814	0.1152
RKHS-KA_70	0.2153	0.1756	0.1147
GBLUP_70	0.2338	0.1839	—
BL_90	0.2484	0.2000	0.1251
RKHS-KA_90	0.2386	0.1952	<u>0.1284</u>
GBLUP_90	<u>0.2574</u>	<u>0.2024</u>	—
<i>Bi-parental population 2 (250 <math>F_2</math> individuals; 271 SNPs)</i>			
	271 SNPs	100 SNPs	50 SNPs
BL_30	0.2994	0.2765	0.2165
RKHS-KA_30	0.2924	0.2725	0.2136
GBLUP_30	0.2982	0.2770	0.2161
BL_40	0.3367	0.3156	0.2447
RKHS-KA_40	0.3371	0.3172	0.2471
GBLUP_40	0.3378	0.3146	—
BL_50	0.3471	0.3264	0.2585
RKHS-KA_50	0.3486	0.3271	0.2621
GBLUP_50	0.3467	—	—
BL_70	0.3717	0.3549	0.2818
RKHS-KA_70	0.3770	0.3605	0.2866
GBLUP_70	0.3714	—	—
BL_90	0.3919	0.3725	0.2998
RKHS-KA_90	<u>0.4066</u>	<u>0.3859</u>	<u>0.3199</u>
GBLUP_90	0.3903	—	—

Abbreviations: BL, Bayesian LASSO; RKHS, reproducing kernel Hilbert space; SNP, single-nucleotide polymorphism. For each bi-parental population and for each number of markers (columns) the best predictive model is underlined.

or not biologically or logistically possible. The most important questions relating to the applicability of GS in the CIMMYT maize and wheat global breeding programs are which prediction problem will better help breeders predict: (1) breeding values of individuals for

rapid selection cycling or (2) genotypic values of advanced lines that are in the last stages of testing. While predicting breeding values require a precise estimation of additive effects, predicting genetic values of advance lines requires using models that account for additive as well as epistatic genetic effects. Results of analyzing vast amounts of data in both CIMMYT breeding programs indicate that pedigree and markers offer opportunities for achieving prediction than can be exploited in the breeding pipeline. However, questions on where and how to use this information remain open. Questions that require further research are, among others, how many individuals and markers are needed per bi-parental population? And can related bi-parental families (half-sibs or quarter-sibs) increase prediction accuracy? The impact of these factors on the prediction accuracy of genomic breeding values on which selection decisions are made can be evaluated by computer simulation.

At CIMMYT, we are placing increasingly greater emphasis on the design of the validation population. In CIMMYT breeding programs, genomic information may be useful for (1) predicting the effect of an unknown population structure, (2) predicting unrecorded pedigree structure, (3) correcting incorrect pedigree and (4) making predictions about the genetic value of Mendelian sampling terms. Each of these four uses of genomics in the breeding programs has different economic value; therefore, design validation schemes should sometimes partition the overall accuracy of a genomic breeding value into the accuracy of these different components.

#### Genetic and environmental structure

Genetic structure (family or breeding populations) brings up questions such as which prediction problem needs to be assessed when applying GS in real plant breeding populations, predicting between families or within family, or both. Likewise, do distantly related (that is, quarter-half-sibs) families contribute (or not) to prediction ability compared with families (that is, half-sibs or full-sibs) that are more closely related and therefore can be better predicted? Experimental maize data indicate that prediction of unrelated lines has low accuracy, and that we can predict between families only if they have some degree of relatedness.

The initial prediction results of CIMMYT maize and wheat breeding trials indicate that promising accuracy can be achieved. In this prediction assessment case, all markers will inherently consider substructures (as the pedigree model does). De los Campos *et al.* (2010) showed that regressing phenotypic values on all available markers is equivalent to regressing phenotypic values on all principal components, thus considering all subpopulations. Furthermore, when combining data from different experiments and populations with different levels of performance, the mixed model will automatically adjust for differences in the mean of these populations and/or experiments. A different prediction problem arises when predicting between subpopulations that are not related. In this case, prediction accuracy dropped significantly and some becomes negligible. However, when all subpopulations are related and form one large global population with a clear structure, the prediction of one population based on the others can be done with relatively high accuracy, as shown by the prediction of stem rust in wheat. However, for a complex trait such as grain yield, these relatively high predictions may not be obtained. A different prediction assessment is concerned with predicting within bi-parental populations; in maize, a prediction ability of around 0.4 with low marker density was achieved. This could save resources during phenotyping.

Although GE has an important role in genomic prediction (Burgu eno *et al.*, 2012), genomic information can only be used if

the **G** matrix from bi-parental populations consisting of an  $F_2$  crossed with a tester (that has no structure) and obtained from around 200 SNP markers contains information that will allow borrowing genotypic information from other individuals in other environments. Increasing the number of markers and/or the sample size might further increase prediction ability. Initial results of bi-parental populations should be assessed in more  $F_2$  populations that have been evaluated using different crosses (related and unrelated) tested in different environmental conditions.

#### Breeding value components in GS

Formally, a breeding value can be partitioned into two components: (1) the parent average (that is, one individual receives 50% of its genome from each of its two parents) and (2) Mendelian sampling, which is the random sampling of the genome of each parent. In efficient breeding programs, genetic progress is primarily driven by the time taken to evaluate the Mendelian sampling term accurately. In the absence of performance on an individual itself or on its descendants, traditional breeding value estimation methods have no ability to evaluate the Mendelian sampling term. Plant breeders have traditionally evaluated the genetic potential of the Mendelian sampling term using replicated field trials of the candidates for selection that are both expensive and time consuming.

Genomic information offers the possibility of accurately estimating the genetic potential of Mendelian sampling terms without having to evaluate individual candidates for selection in the field. This has a great time- and cost-saving potential. While the major benefit of genomic information in an efficient breeding program will likely be its ability to accurately evaluate the Mendelian sampling term, genomic information also has other valuable uses. For example, it can be used to overcome the effect of incorrectly recorded or unrecorded pedigree information, thereby helping to increase the ability to accurately evaluate the parent average component in a breeding value. In some breeding programs, such as those with very low levels of pedigree recording, there may be major population subgroups or structures hidden within the data that could be captured and revealed by the genomic information.

While making selection decisions based on breeding values that contain these pedigree or structure components (as well as some Mendelian sampling components) may be useful, breeders need to be cautious, as the family or population structure with the highest genetic merit could come to dominate the breeding population over time, and useful genetic variation within other families or population subgroups could be lost. These are well-known results using classic BLUP methodology, as BLUP uses both within- and between-family information, and with low heritability traits, between-family selection is dominant. For this reason, methods to control inbreeding in BLUP selection programs are used, so that individuals from more families are chosen.

Therefore, for the reasons mentioned above, when validating the performance of GS within breeding programs, one needs to be aware of the different components possibly contained within a genomic breeding value and evaluate the benefits of genomic information regarding these components. For example, when structure is known, genomic information may not be needed to evaluate its genetic merit; pedigree information is reasonably powerful for evaluating the genetic merit of the parent average component, yet only genomic information or field trials on the selection candidates or their descendants can evaluate the genetic potential of the Mendelian sampling term. Validation of the usefulness of genomic information should place major emphasis on the latter comparison.

## DATA ARCHIVING

There were no data to deposit.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

We would like to thank the numerous cooperators in national agricultural research institutes who carried out the wheat and maize trials in Africa, Asia and Latin America. We also acknowledge financial resources from the Durable Rust Resistant Wheat Project led by Cornell University and the Drought Tolerance Maize for Africa led by CIMMYT; both projects are financed by the Bill and Melinda Gates Foundation.

Bernardo R (2008). Molecular markers and selection of complex traits in plants: learning from the last 20 years. *Crop Sci* **48**: 1649–1664.

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

Burgueño J, Crossa J, Cornelius PL, Trethowan R, McLaren G, Krishnamachari A (2007). Modeling additive  $\times$  environment and additive  $\times$  additive  $\times$  environment using genetic covariances of relatives of wheat genotypes. *Crop Sci* **43**: 311–320.

Burgueño J, Crossa J, Cotes JM, San Vicente F, Das B (2011). Prediction assessment of linear mixed models for multi-environment trials. *Crop Sci* **51**: 944–954.

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

Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL *et al.* (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**: 713–724.

Crossa J, Pérez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokosho C (2011). Genomic selection and prediction in plant breeding. *J Crop Improv* **25**: 239–261.

Daetwyler HD, Villanueva B, Woolliams J (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**: e3395.

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

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

Gianola D, Fernando R, Stella A (2006). Genomic-assisted prediction of genetic values with semiparametric procedures. *Genetics* **173**: 1761–1776.

Gianola D, Okut H, Weigel KA, Rosa GJM (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet* **12**: 87.

Goddard ME, Hayes BJ (2009). Genomic selection. *J Anim Breed Genet* **124**: 323–330.

Gonzalez-Camacho JM, de los Campos G, Pérez P, Gianola D, Cairns J, Mahuku G *et al.* (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor Appl Genet* **125**: 759–771.

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

## APPENDIX

### Pedigree model (P)

In this model, the vector of genetic values  $\mathbf{g} = \{g_i\}$  was assumed to follow a multivariate normal density centered at zero and with a (co)variance matrix proportional to the numerator relationship matrix ( $\mathbf{A}$ ), computed from the pedigree, that is,  $\mathbf{g} \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$ . Here,  $\sigma_a^2$  is the variance of additive effects on the base population and the unknowns are  $\{\mathbf{g}, \sigma_a^2, \sigma_e^2\}$ . Independent scaled-inverse  $\chi^2$  densities,  $\chi^{-2}(\sigma^2 | df, S)$ , were assigned to variance parameters. Therefore, the joint posterior density of model unknowns is

$$p(\mathbf{g}, \sigma_a^2, \sigma_e^2 | \mathbf{y}, H) \propto \prod_{i=1}^n N(y_i | g_i, \sigma_e^2) N(\mathbf{g} | \mathbf{0}, \mathbf{A}\sigma_a^2) \chi^{-2}(\sigma_e^2 | df_e, S_e) \chi^{-2}(\sigma_a^2 | df_a, S_a)$$

where  $H$  denotes the collection of hyperparameters, and  $df$  and  $S$  are prior degree-of-freedom and scale parameters set to one and four,

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* **92**: 433–443.

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning*, 2nd edn. Springer-Verlag: NY, USA.

Heffner EL, Jannink J-L, Sorrells M (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome* **4**: 65–75.

Hickey JM, Crossa J, Babu R, de los Campos G (2012). Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Sci* **52**: 654–663.

Lorenzana RE, Bernardo R (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor Appl Genet* **120**: 151–161.

McKinney BA, Pajewski NM (2012). Six degrees of epistasis: statistical network models for GWAS. *Front Genet* **2**: 1–6.

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

Ornella L, Singh S, Pérez P, Burgueño J, Singh R, Tapia E *et al.* (2012). Genomic prediction of genetic values for resistance to wheat rusts. *The Plant Genome* **5**: 136–148.

Park T, Casella G (2008). The Bayesian LASSO. *J Am Stat Assoc* **103**: 681–686.

Pérez P, de los Campos G, Crossa J, Gianola D (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression Package in R. *The Plant Genome* **3**: 106–116.

Pérez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker S (2012). A comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *Genes Genomes Genet* **2**: 1595–1605.

Riedelshheimer C, Czedik-Eysenberg A, Grieder C, Lisek J, Technow F, Sulpice R *et al.* (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat Genet* **44**: 217–220.

VanRaden PM (2007). Genomic measures of relationship and inbreeding. *Interbull Annu Meeting Proc, Interbull Bulletin* **37**: 33–36.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF *et al.* (2008). Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci* **92**: 16–24.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**: 4414–4423.

Wang D, El-Basyoni IS, Baenziger PS, Crossa J, Eskridge K, Dweikat I (2012). Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity* **109**: 313–319.

Wang D, Eskridge K, Crossa J (2011). Identifying QTLs and epistasis in structured plant breeding populations using adaptive mixed LASSO. *J Agric Biol Environ Stat* **16**: 170–184.

Windhausen VS, Atlin GN, Crossa J, Hickey JM, Grudloyma P, Terekegne A *et al.* (2012). Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *Genes Genomes Genet* **2**: 1427–1436.

Zhang YM, Xu S (2005). A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity* **95**: 96–104.

Zhao Y, Gowda M, Liu W, Würschum T, Maurer HP, Longin FH *et al.* (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theor Appl Genet* **124**: 769–776.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>

respectively. Inferences were based on samples from the above posterior density obtained using a Gibbs sampler.

### Linear regression models

**GBLUP.** One of the first penalized regression methods used in genomic selection was Ridge Regression (RR). This marker-based model (RR-BLUP) is expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1A)$$

where  $\mathbf{X}$  is the genotype matrix for the bi-allelic markers (coded as 0, 1 and 2) and  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ . The marker effects are obtained by solving the optimization problem  $\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \{ \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 + \tilde{\lambda} \sum_{j=1}^p \beta_j^2 \}$ , where  $\tilde{\lambda} \geq 0$  is a regularization parameter. It can be shown that the solution to this problem is given by  $\hat{\boldsymbol{\beta}} = [\lambda \mathbf{I} + \mathbf{X}'\mathbf{X}]^{-1} \mathbf{X}'\mathbf{y}$ . GEBV can be obtained

from model (1A) by setting  $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$  such that the equation 1A can be written as

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon} \quad (2A)$$

Assuming that  $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_{\beta}^2 \mathbf{I})$ , it can be shown that  $\mathbf{g} \sim N(\mathbf{0}, \mathbf{K}\sigma_{\beta}^2)$ , where  $\mathbf{K} = \mathbf{X}\mathbf{X}'$ . When  $\mathbf{G} = \mathbf{K}/c = \mathbf{X}\mathbf{X}'/c$  with  $c = \sum_j 2p_j(1-p_j)$  and  $p_j$  being the minor allele frequency, then  $\sigma_{\beta}^2 = \sigma_{\beta}^2 c$  and is called GBLUP, and the  $\mathbf{G}$  matrix is the genomic-derived relationship matrix, with  $\hat{\mathbf{g}} = \mathbf{X}\mathbf{X}'[\mathbf{X}\mathbf{X}' + \lambda\mathbf{I}]^{-1}\mathbf{y}$ .

*Linear mixed models with genotype  $\times$  environment interaction.* For simplicity, we will use the same notation as in Burgueño *et al.* (2012). The basic linear mixed model for the phenotypic response of  $g$  individuals evaluated in  $J$  environments is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\mathbf{g} + \boldsymbol{\varepsilon} \quad (3A)$$

with  $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_J)'$ ,  $\mathbf{g} = (\mathbf{g}'_1, \dots, \mathbf{g}'_J)'$ ,  $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \dots, \boldsymbol{\varepsilon}'_J)'$ , where  $\mathbf{y}_j = \{y_{ij}\}$  ( $i = 1, 2, \dots, g$ ) is a column vector of  $g$  phenotypic records (or means) collected in each of the  $j$ th environments ( $j = 1, \dots, J$ ),  $\mathbf{X}_j$  and  $\mathbf{Z}_j$  are incidence matrices vectors of systematic effects,  $\boldsymbol{\theta}_j$  and random genetic effects  $\mathbf{g}_j$ , for each environment.

*Pedigree and genomic linear mixed models.* The vector of random effects had a covariance structure  $\text{Cov}(\mathbf{g}, \mathbf{g}') = \mathbf{G}_0 \otimes \mathbf{A}$ ,  $\text{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}') = \mathbf{R}_0 \otimes \mathbf{I}$  and  $\text{Cov}(\mathbf{g}, \boldsymbol{\varepsilon}') = \mathbf{0}$ , where  $\mathbf{G}_0$  is a  $J \times J$  variance-covariance of the genetic effect on environments,  $\otimes$  is the Kronecker product,  $\mathbf{A} = \{a(i, j')\}$  is a numerator relationship matrix,  $\mathbf{R}_0 = \{\text{Cov}(\varepsilon_{ij}, \varepsilon_{j'})\}$  is a  $g \times g$  covariance matrix (within-environments, across-genotypes) of model residuals and  $\mathbf{I}$  is an identity matrix of size  $J$ . This model can easily be extended to accommodate heterogeneous residual variances.

The marginal density of the data is multivariate normal with mean and variance-covariances as

$$[\mathbf{y} | \boldsymbol{\theta}, \mathbf{R}_0, \mathbf{G}_0] \sim N[\mathbf{X}\boldsymbol{\theta}, \mathbf{Z}(\mathbf{G}_0 \otimes \mathbf{A})\mathbf{Z}' + \mathbf{R}_0 \otimes \mathbf{I}] \quad (4A)$$

In the previous models,  $\mathbf{A} = \{a(i, j')\}$  represents a matrix of additive relationships derived from a pedigree ( $\mathbf{A}_P$ ). Alternatively, this matrix can be derived from molecular marker information, that is,  $\mathbf{A}_M = \mathbf{G}$  (the genomic relationship matrix in GBLUP; see VanRaden (2008) for more details).

Burgueño *et al.* (2012) outlined how pedigree and marker information can be combined in a single model by extending (3A) to a model with two random effects, one of which,  $\mathbf{g}_P \sim N(\mathbf{0}, \mathbf{G}_{0P} \otimes \mathbf{A}_P)$ , represents a regression on pedigree additive relationships, and the other represents a regression on marker additive relationships,  $\mathbf{g}_M \sim N(\mathbf{0}, \mathbf{G}_{0M} \otimes \mathbf{A}_M)$ . Thus, (3A) and (4A) become

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}(\mathbf{g}_P + \mathbf{g}_M) + \boldsymbol{\varepsilon} \quad (5A)$$

and

$$[\mathbf{y} | \boldsymbol{\theta}, \mathbf{G}_{0P}, \mathbf{G}_{0M}, \mathbf{R}_0] \sim N[\mathbf{X}\boldsymbol{\theta}, \mathbf{Z}(\mathbf{G}_{0P} \otimes \mathbf{A}_P + \mathbf{G}_{0M} \otimes \mathbf{A}_M)\mathbf{Z}' + \mathbf{R}_0 \otimes \mathbf{I}] \quad (6A)$$

Similar to Burgueño *et al.* (2007), pedigree-derived additive and additive  $\times$  additive relationship information can be combined into a single model by extending (5A) to a model with two random effects, one of which,  $\mathbf{g}_P \sim N(\mathbf{0}, \mathbf{G}_{0P} \otimes \mathbf{A}_P)$ , is a regression on pedigree additive relationships. The other represents a regression on pedigree epistasis additive  $\times$  additive relationships,  $\mathbf{g}_{PP} \sim N(\mathbf{0}, \mathbf{G}_{0PP} \otimes \tilde{\mathbf{A}}_P)$ , where  $\tilde{\mathbf{A}}_P = \mathbf{A}_P \# \mathbf{A}_P$  (where  $\#$  is the element-wise multiplication operator, that is, Hadamard product). Therefore (5A) and (6A) become

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}(\mathbf{g}_P + \mathbf{g}_{PP}) + \boldsymbol{\varepsilon} \quad (7A)$$

and

$$[\mathbf{y} | \boldsymbol{\theta}, \mathbf{G}_{0P}, \mathbf{G}_{0PP}, \mathbf{R}_0] \sim N[\mathbf{X}\boldsymbol{\theta}, \mathbf{Z}(\mathbf{G}_{0P} \otimes \mathbf{A}_P + \mathbf{G}_{0PP} \otimes \tilde{\mathbf{A}}_P)\mathbf{Z}' + \mathbf{R}_0 \otimes \mathbf{I}] \quad (8A)$$

Habier *et al.* (2007) have shown that  $\mathbf{A}_M$  approaches  $\mathbf{A}_P$ ; therefore it is possible to assume that  $\tilde{\mathbf{A}}_M$  tends to  $\tilde{\mathbf{A}}_P$ . Thus, the genomic-derived additive and additive  $\times$  additive relationships can be combined into a single model by extending (5A) to a model with two random effects, one with  $\mathbf{g}_M \sim N(\mathbf{0}, \mathbf{G}_{0M} \otimes \mathbf{A}_M)$ , representing a regression on genomic additive relationships, and the other representing a regression on genomic epistasis additive  $\times$  additive relationships,  $\mathbf{g}_{MM} \sim N(\mathbf{0}, \mathbf{G}_{0MM} \otimes \tilde{\mathbf{A}}_M)$ . Thus (7A) and (8A) become

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}(\mathbf{g}_M + \mathbf{g}_{MM}) + \boldsymbol{\varepsilon} \quad (9A)$$

and

$$[\mathbf{y} | \boldsymbol{\theta}, \mathbf{G}_{0M}, \mathbf{G}_{0MM}, \mathbf{R}_0] \sim N[\mathbf{X}\boldsymbol{\theta}, \mathbf{Z}(\mathbf{G}_{0M} \otimes \mathbf{A}_M + \mathbf{G}_{0MM} \otimes \tilde{\mathbf{A}}_M)\mathbf{Z}' + \mathbf{R}_0 \otimes \mathbf{I}] \quad (10A)$$

*Modeling genotype  $\times$  environment covariance structures using the factor analytic model.* The factor analytic (FA) model that defines matrices for  $\mathbf{g}_P$  and  $\mathbf{g}_M$  in (5A) and (6A) is  $(\mathbf{A}_P \mathbf{A}'_P + \boldsymbol{\Psi}_P) \otimes \mathbf{A}_P = [\text{FA}(k)]_P \otimes \mathbf{A}_P$  for pedigree and  $(\mathbf{A}_M \mathbf{A}'_M + \boldsymbol{\Psi}_M) \otimes \mathbf{A}_M = [\text{FA}(k)]_M \otimes \mathbf{A}_M$  for markers where  $\mathbf{A}_P$  (or  $\mathbf{A}_M$ ) is a matrix of order  $J \times k$ , with the  $k$ th column containing the environment loadings for the  $k$ th latent factor, and  $\boldsymbol{\Psi}$  being a diagonal matrix  $(\sigma_{\delta_1}^2, \sigma_{\delta_2}^2, \dots, \sigma_{\delta_j}^2)$  of the order  $J \times J$ . When only one factor is considered,  $k = 1$ , the model has one multiplicative term and is denoted as FA(1); for  $k = 2$ , FA(2) has two multiplicative components, and so on.

The factor analytic model that defines  $(\mathbf{g}_P + \mathbf{g}_{PP})$  in (7A) and (8A) is  $(\mathbf{A}_P \mathbf{A}'_P + \boldsymbol{\Psi}_P) \otimes \mathbf{A}_P + (\mathbf{A}_{PP} \mathbf{A}'_{PP} + \boldsymbol{\Psi}_{PP}) \otimes \tilde{\mathbf{A}}_P = [\text{FA}(k)]_P \otimes \mathbf{A}_P + [\text{FA}(k)]_{PP} \otimes \tilde{\mathbf{A}}_P$ . Similarly, the factor analytic model that defines  $(\mathbf{g}_M + \mathbf{g}_{MM})$  in (9A) and (10A) is  $(\mathbf{A}_M \mathbf{A}'_M + \boldsymbol{\Psi}_M) \otimes \mathbf{A}_M + (\mathbf{A}_{MM} \mathbf{A}'_{MM} + \boldsymbol{\Psi}_{MM}) \otimes \tilde{\mathbf{A}}_M = [\text{FA}(k)]_M \otimes \mathbf{A}_M + [\text{FA}(k)]_{MM} \otimes \tilde{\mathbf{A}}_M$ . In this model, only one of  $\boldsymbol{\Psi}_M$  or  $\boldsymbol{\Psi}_{MM}$  ( $\boldsymbol{\Psi}_P$  or  $\boldsymbol{\Psi}_{PP}$ ) can be estimated in order to make the parameter identifiable.

*Bayesian LASSO (BL).* Marker effects in (1A) are assumed independent and identically distributed *a priori*. The BL assigns the same double exponential distribution to all marker effects (conditionally on a regularization parameter), that is  $p(\beta_j | \lambda, \sigma_e) = DE(\beta_j | 0, \frac{\lambda}{\sigma_e^2})$ , where  $\lambda \geq 0$  is a regularization parameter. Compared with the Normal distribution, the DE assigns a bigger mass near 0, shrinking small marker effects. The joint posterior density of models unknowns is given by:

$$p(\boldsymbol{\beta}, \sigma_e^2, \lambda | \mathbf{y}, \mathbf{X}) \propto \prod_{i=1}^n N\left(y_i | \sum_{j=1}^p x_{ij} \beta_j, \sigma_e^2\right) \times \prod_{j=1}^p DE\left(\beta_j | 0, \frac{\lambda}{\sigma_e^2}\right) p(\sigma_e^2) p(\lambda^2).$$

It is usual to assign a prior distribution to the square of the regularization parameter ( $\lambda^2$ ); for example, it can be a beta or a gamma distribution (de los Campos *et al.*, 2009) but it also can be estimated by using cross-validation (Park and Casella, 2008). It is also usual to assign a scaled-inverse  $\chi^2$  distribution,  $\chi^{-2}(df_e, S_e)$  to  $\sigma_e^2$ . The DE distribution does not conjugate with the normal distribution, but it can be expressed as a mixture of scaled normal densities, which allows us to represent the original problem as a hierarchical model (Park and Casella, 2008) and then the posterior distributions can be obtained using a Gibbs sampler.

The BL is explained in detail in several articles, such as Park and Casella (2008), de los Campos *et al.* (2009), Crossa *et al.* (2010; 2011), Pérez *et al.* (2010) and Gonzalez-Camacho *et al.* (2012). Guidelines for setting the hyperparameters for the prior distributions for the regularization parameters and  $\sigma_e^2$  are given in Pérez *et al.* (2010) and Gonzalez-Camacho *et al.* (2012).

**Principal component regression—lower-dimension representation of matrices K and G**

De los Campos *et al.* (2010) showed that the symmetric matrix **K** and the rectangular marker matrix **X** can be represented using spectral decomposition and singular value decomposition, respectively. Based on these eigen analyses, the authors show that the regression of the phenotypes on the markers is equivalent to the regression of the phenotypes on the principal components. Thus, a model that includes all the markers implicitly considers there may be potential population structures in the data.

A general prediction model can be represented as

$$\mathbf{y} = \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{11A}$$

where **y** is the phenotypic response (centered), **K** is an incidence matrix that could be the same as the realized relationship matrix (**G**), or the same as **XX'** or the same as matrix **X**;  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ , and  $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{K}^{-1}\sigma_g^2)$ . This last assumption helps obtain results related to the spectral decomposition and singular value decomposition of **K** and **X** (de los Campos *et al.*, 2010). However,  $\boldsymbol{\alpha}$ , as previously defined, does not strictly denote the effects of the markers.

Defining  $\mathbf{g} = \mathbf{K}\boldsymbol{\alpha}$ , de los Campos *et al.* (2010) showed that

$$\hat{\mathbf{g}} = \mathbf{K}\hat{\boldsymbol{\alpha}} = \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$$

Note that this is similar to equation. 2 of vanRaden (2008), that is,  $\hat{\mathbf{g}} = \hat{\mathbf{a}}$  ( $\hat{\mathbf{a}}$  denotes the GEBV according to vanRaden (2008) and, in this case,  $\mathbf{K} = \mathbf{G}$ ). According to de los Campos *et al.* (2010),  $\hat{\mathbf{a}} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$ .

*Spectral decomposition.* Following the previous line of reasoning, and as  $\mathbf{g} = \mathbf{K}\boldsymbol{\alpha}$ , then equation. 11A can be written as  $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$ . Owing to the fact that  $\mathbf{K} = \mathbf{G}$  is a symmetric matrix, it can be expressed as

$\mathbf{K} = \boldsymbol{\Psi}\boldsymbol{\Delta}\boldsymbol{\Psi}'$ , where  $\boldsymbol{\Delta}$  contains the eigenvectors of  $\mathbf{K}(\boldsymbol{\Delta}\boldsymbol{\Delta}' = \mathbf{I})$ , and  $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}_j\}$  is a diagonal matrix with the eigenvalues of **K** (with  $j = 1, 2, \dots, g$ ). Then model  $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$  can be reformulated as

$$\mathbf{y} = \boldsymbol{\Delta}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

where  $\boldsymbol{\delta} = \boldsymbol{\Psi}\boldsymbol{\Delta}'\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{\Psi}\boldsymbol{\sigma}_g^2)$  and  $Var(\boldsymbol{\Delta}\boldsymbol{\delta}) = \mathbf{K}\boldsymbol{\sigma}_g^2$ . Evidently, as  $\boldsymbol{\delta} \sim N(\mathbf{0}, \boldsymbol{\Psi}\boldsymbol{\sigma}_g^2)$ , then  $\boldsymbol{\Delta}\boldsymbol{\delta} \sim N(\mathbf{0}, \mathbf{K}\boldsymbol{\sigma}_g^2)$ .

*Singular value decomposition.* Note that  $\mathbf{y} = \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$ ,  $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$  and  $\mathbf{y} = \boldsymbol{\Delta}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$  are not generally used in GS. Usually, in GS the model expressing the regression of phenotype on markers is defined in equation 1A, where **X** was already defined and  $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{I}\sigma_\beta^2)$  is the vector with the marker effects. Then, the singular value decomposition of **X** is  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$ , where **U** and **V** denote the left and right eigenvectors of matrix **X**, and  $\mathbf{D} = \{\zeta_j^2\}$  is the diagonal matrix with the singular values of matrix **X**. Defining  $\boldsymbol{\delta} = \mathbf{D}\mathbf{V}'\boldsymbol{\beta}$ , it can be shown that (1A) can be written as in de los Campos *et al.* (2010):

$$\mathbf{y} = \mathbf{U}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

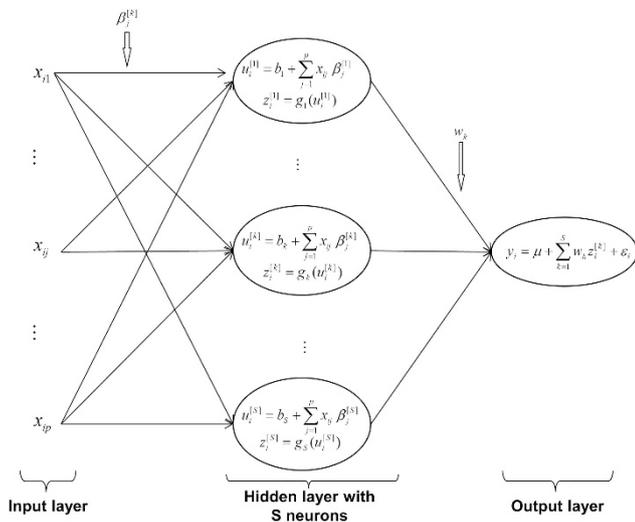
where  $\boldsymbol{\delta} = \mathbf{D}\mathbf{V}'\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{D}\boldsymbol{\sigma}_g^2)$ . This indicates that regressing phenotypes on all markers is equivalent to regressing phenotypes on all principal components. This reparameterization of the original model based on eigenvalue decomposition accounts for possible unrevealed population structures existing in the training set.

**Nonlinear models**

*Single hidden layer feed-forward neural network (NN).* We follow the symbols and sequence outlined by Pérez *et al.* (2012) with the structure of the NN depicted in Figure A1 (originally presented by Gonzalez-Camacho *et al.* (2012), and later adapted by Perez *et al.* (2012)) This NN can be thought of as a two-step regression (for example, Hastie *et al.*, 2009). In the first step, in the hidden layer, input variables  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$  ( $j = 1, \dots, p$  markers) are combined for each neuron ( $k = 1, \dots, S$  neurons), using a linear function,  $u_i^{[k]} = b_k + \sum_{j=1}^p x_{ij} \beta_j^{[k]}$ , and subsequently transformed using a nonlinear activation function, yielding a set of inferred scores,  $z_i^{[k]} = g_k(u_i^{[k]})$ , where  $g_k(\cdot)$  is the activation function that maps the inputs into the real line in the closed interval  $[-1, 1]$ . One computes  $z_i^{[k]} = g_k(b_k + \sum_{j=1}^p x_{ij} \beta_j^{[k]})$ , where  $b_k$  is an intercept, and  $(\beta_1^{[1]}, \dots, \beta_p^{[1]}, \dots, \beta_1^{[S]}, \dots, \beta_p^{[S]})'$  is a vector of regression coefficients or 'weights' of each neuron  $k$  in the hidden layer. In the second step, these scores are used in the output layer as basis functions to regress the response using the linear activation function on the data-derived predictors, such that  $y_i = \sum_{k=1}^S w_k z_i^{[k]} + \varepsilon_i = \sum_{k=1}^S w_k g_k(b_k + \sum_{j=1}^p x_{ij} \beta_j^{[k]}) + \varepsilon_i$ .

*Radial Basis Function Neural Network (RBFNN).* The architecture of a single hidden layer RBFNN with  $S$  nonlinear neurons is similar to that of the single hidden layer feed-forward neural network. The only difference is that each nonlinear neuron in the hidden layer has a Gaussian RBF defined as  $z_i^{[k]} = \exp[-h_k \|\mathbf{x}_i - \mathbf{c}_k\|^2]$ , where  $\|\mathbf{x}_i - \mathbf{c}_k\|$  is the Euclidean norm between the input vector,  $\mathbf{x}_i$ , and the center vector,  $\mathbf{c}_k$ , and  $h_k$  is the bandwidth of the Gaussian RBF. Then, in the linear output layer, phenotypes are regressed on the data-derived features,  $\{z_i^{[k]}\}$ ; thus the response phenotype is  $y_i = \sum_{k=1}^S w_k z_i^{[k]} + \varepsilon_i$ , where  $\varepsilon_i$  is a model residual.

*Reproducing Kernel Hilbert Spaces (RKHS) Regression.* RKHS models have been suggested as an alternative to multiple linear regressions for capturing complex interaction patterns that may be



**Figure A1** Structure of a single-layer feed-forward neural network (from Pérez *et al.*, 2012; originally presented by Gonzalez-Camacho *et al.*, 2012).

difficult to account for in a linear model framework (Gianola *et al.*, 2006). In a RKHS model, the regression function takes the following form:

$$g(\mathbf{x}_i) = \sum_{i'=1}^n \alpha_{i'} K(\mathbf{x}_i, \mathbf{x}_{i'})$$

where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  and  $\mathbf{x}_{i'} = (x_{i'1}, \dots, x_{i'p})'$  are input vectors of marker genotypes in individuals  $i$  and  $i'$ ,  $\alpha_i$  are regression coefficients,

and  $K(\mathbf{x}_i, \mathbf{x}_{i'}) = \exp(-h\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2)$  is the reproducing kernel defined (here) with a Gaussian basis function, where  $h$  is a bandwidth parameter and  $\|\mathbf{x}_i - \mathbf{x}_{i'}\|$  is the Euclidean norm between pairs of input vectors. The strategy termed 'kernel averaging' for selecting optimal values of  $h$  within a set of candidate values was implemented using the Bayesian approach described in de los Campos *et al.* (2010). Similarities and connections between the RKHS and the RBFNN are given in Gonzalez-Camacho *et al.* (2012).