Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers

José Crossa[*, 1, 2], Gustavo de los Campos*[, †, 2], Paulino Pérez[§,*, 2]; Daniel Gianola[†, ‡], Gary Atlin*, Juan Burgueño[§,]*, José Luis Araus*, Dan Makumbi*, Jianbing Yan*, Vivi Arief ¶, Marianne Banziger*, and Hans-Joachim Braun*

* International Maize and Wheat Improvement Center (CIMMYT), México;
† Department of Biostatistics, University of Alabama-Birmingham, USA;
‡ Departments of Dairy Science, Animal Sciences, and Biostatistics and Medical Informatics, University of Wisconsin-Madison, USA;
§ Colegio de Postgraduados, Montecillos, México;
¶ School of Land Crop and Food Sciences of the University of Queensland, Australia.

ABSTRACT

The use of molecular marker data has become an important aid in plant breeding, and the availability of dense markers has made possible the use of genomic selection for enhancing the prediction of genetic values. However, the evaluation of models for genomic selection (GS) in real populations is very limited. This article evaluates two parametric models and one semi-parametric model for GS in two extensive datasets. The first dataset contains historical phenotypic records of a series of wheat (*Triticum aestivum* L.) trials and recently generated genomic data. The other dataset pertains to international maize (*Zea mays*) trials in which different traits were measured in maize lines evaluated under severe drought and well-watered conditions. The findings of this study, which used extensive cross-validations, showed that models including marker information yield high correlations between predicted and observed phenotypic outcomes and produce important gains in predictive ability relative to pedigree-based models; these gains, in the wheat dataset, ranged from 7.7% to 35.7%. Estimates of marker effects were different across environmental conditions indicating that genotype × environment interaction is an important component of genetic variability. These results indicate that GS in plant breeding can be an effective strategy for selecting among lines whose phenotypes have yet to be observed. Denser markers will become available soon, and this may further improve the ability to predict genetic values for complex traits in plant breeding.

INTRODUCTION

Genetic improvement of complex traits in plants and animals has been based mainly on the standard additive infinitesimal model of quantitative genetics (FISHER, 1918). Animal breeders have used this model for predicting breeding values either in a mixed model (BLUP; HENDERSON, 1984) or a Bayesian framework (GIANOLA AND FERNANDO, 1986). More recently, plant breeders have incorporated pedigree data into linear mixed models for predicting breeding values (e.g., PIEPHO *et al.*, 2007; BURGUEÑO *et al*., 2007; CROSSA *et al*., 2006, 2007).

The use of marker data in animal and plant breeding has become an important aid for mapping quantitative trait loci (QTL) as well as for marker-assisted selection (MAS) of major genes. Linkage disequilibrium between molecular markers and a QTL can be used for identifying genomic regions influencing the trait of interest, as well as for improving the

prediction of genetic values. In standard models for genomic selection (GS; e.g., MEUWISSEN *et al.*, 2001), phenotypic outcomes are regressed on marker genotypes, and knowledge of the extent of linkage disequilibrium is not necessary (e.g., LANGE and WHITTAKER, 2001; GIANOLA *et al*., 2003).

The availability of thousands of genome wide molecular markers has made it possible to use GS for enhancing the prediction of genetic values (MEUWISSEN *et al.*, 2001) in both plant (e.g., BERNARDO and YU, 2007; PIEPHO, 2009) and animal breeding (GONZALEZ-RECIO *et al*., 2008; HAYES *et al.*, 2009; VANRADEN *et al.*, 2008; DE LOS CAMPOS *et al*., 2009a).

Coping with the curse of dimensionality and with co-linearity, two issues that arise when the number of molecular markers (p) is large relative to the number of observations (n), are two important challenges for models for GS. Another challenge is how models can accommodate the complexity of quantitative traits (e.g., diverse forms and degrees of interaction between genes) as well as the peculiarities of breeding populations in which the standard assumptions of an infinitesimal model (such as linkage equilibrium, no natural or artificial selection, and assortative mating) do not hold. Parametric (e.g., MEUWISSEN *et al.*, 2001) and semi-parametric (e.g., GIANOLA *et al*., 2006; GIANOLA and VAN KAMM, 2008) procedures address these problems differently.

In standard genetic models, phenotypic outcomes, $y_i \ \left(i=1,...,n\right)$, are viewed as the sum of a genetic value, g_i, and a model residual, ε_i; that is, $y_i = g_i + \varepsilon_i$. One method for incorporating markers in models for GS is to define g_i as a parametric regression on marker covariates x_{ij} (that can take on values of 1, 0, or -1 for a bi-allelic marker of a segregating population or values of 1 and -1 for inbred lines) of the form $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$, such that $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$ (j=1,…,p) (or $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ in matrix notation), where β_j is the regression of y_i on the j^{th} marker covariates x_{ij}.

Estimation of $\boldsymbol{\beta}$ via multiple regression by ordinary least squares (OLS) is not feasible when $p>n$. Subset regression can be implemented by selecting a small set of markers and estimating the effect of the selected markers either by single-marker regression or multiple regression via OLS. However, when marker effects are estimated one at a time, there is a strong tendency to overestimate the absolute value of the marker effect. This produces over-fitting and estimates of breeding values that have a low correlation with the true breeding value (GODDARD and HAYES, 2007).

A commonly used alternative is to estimate marker effects jointly using penalized methods such as ridge regression (HOERL and KENNARD, 1970) or the least absolute value selection and shrinkage operator (LASSO; TIBSHIRANI, 1996), or their Bayesian counterparts. This approach yields greater accuracy of estimated genetic values, and can be coupled with geostatistical models commonly used in plant breeding field trials where the covariance structure depends on the spatial proximity of field plots (PIEPHO, 2009).

In the Bayesian view of ridge regression, marker effects are random variables, all drawn from the same normal distribution and, when variance components are known, estimates of marker effects are best linear unbiased predictors or BLUP (e.g., SCHAEFFER, 2006). However, in ridge regression or its Bayesian counterpart, the extent of shrinkage is homogeneous across markers, which may not be appropriate if some markers are located in regions that are not associated with genetic variance, while markers in other regions may be

linked to QTLs (GODDARD and HAYES, 2007). To overcome this limitation, many authors have proposed methods that use marker-specific shrinkage. In a Bayesian setting, this can be implemented by using priors of marker effects that are mixtures of scaled-normal densities. Examples of this are the Bayes A of MEUWISSEN *et al.* (2001) and the Bayesian LASSO of PARK and CASELLA (2008).

An alternative to parametric regressions is to use semi-parametric methods such as reproducing kernel Hilbert spaces (RKHS) regression (GIANOLA and VAN KAMM, 2008). The Bayesian RKHS regression regards genetic values as random variables coming from a Gaussian process with a (co)variance structure that is proportional to a kernel matrix **K** (DE LOS CAMPOS *et al.*, 2009b), that is, $\mathrm{Cov}(g_i,g_j)\propto \mathrm{K}(\mathbf{x}_i,\mathbf{x}_j)$, where $\mathbf{x}_i$, $\mathbf{x}_j$ are vectors of marker genotypes for the i^{th} and j^{th} individuals, respectively, and $K(.,.)$ is a positive definite function evaluated in marker genotypes. One of the most attractive features of RKHS regression is that the methodology can be used with almost any information set (e.g., covariates, strings, images, graphs). This is particularly important at a time when techniques for characterizing genomes are changing rapidly. A second advantage is that with RKHS the model is represented in terms of n unknowns, which gives RKHS a great computational advantage relative to parametric methods, when $p>>n$.

Usually, in addition to phenotypic records and marker genotypes, a pedigree is also available. The parametric and semi-parametric models described above can be extended to include a regression on a pedigree under the standard assumptions of the infinitesimal additive model (e.g., DE LOS CAMPOS *et al.,* 2009a). In this approach, genetic values can be represented as the sum of two components, $g_i = f_i + u_i$, where f_i specifies some form of parametric or semi-parametric regression on marker genotypes, and u_i is a standard infinitesimal effect with a (co)variance structure modeling the expected resemblance between relatives under an infinitesimal model.

Successful application of GS in plant breeding programs requires comprehensive phenotypic information. The most important quantitative trait in plant breeding for any species is grain yield; other traits, such flowering time in maize, are also very important because they reflect the plants' adaptation to different environmental conditions (BUCKLER *et al*., 2009) (such as drought, altitude, and rainfall) and affect grain yield. In plant breeding, predicting breeding values of tested lines can be based on historical series of variety trials, which, in the case of the Global Maize and Wheat Breeding Programs of the International Maize and Wheat Improvement Center (CIMMYT), are in plentiful supply. These two global breeding programs focus on producing stable, high yielding, and widely adapted advanced breeding lines or varieties. In maize and wheat, grain yield under drought conditions, flowering time, and disease resistance are the primary selection criteria, and the discovery of molecular markers that are closely associated to those traits may speed up breeding progress.

Recently, many lines evaluated by CIMMYT's Global Maize and Wheat Breeding Programs have been genotyped. This study presents parameter estimates and an evaluation of predictive ability of several GS models using two extensive datasets. One contains phenotypic records of a series of wheat trials and recently generated genomic data. The other dataset pertains to international maize trials in which different traits were measured in maize lines evaluated under severe drought and well-watered conditions.

MATERIALS AND METHODS

Experimental data

Two distinct datasets were used: the first one comprises information from a collection of 599 historical CIMMYT wheat lines, and the second one includes information on 300 maize CIMMYT lines. The specifics of each dataset are described below.

Wheat dataset. The wheat dataset is from CIMMYT's Global Wheat Program. Historically, this program has conducted numerous international trials across a wide variety of wheat-producing environments. For this study, we used a subset of 599 wheat lines derived from 25 years of Elite Spring Wheat Yield Trials (ESWYT) conducted from 1979 through 2005. The environments represented in these trials were grouped into four basic target sets of environments (mega-environments) (E1-E4). The phenotypic trait considered here was grain yield (GY) evaluated in each of the four mega-environments. Hereinafter we will refer to this dataset as Wheat-Grain Yield (W-GY).

A pedigree tracing back many generations was available, and the Browse application of the International Crop Information System (ICIS), as described in http://cropwiki.irri.org/icis/index.php/TDM_GMS_Browse (MCLAREN *et al*., 2005), was used for deriving the relationship matrix **A** among the 599 lines.

Wheat lines were genotyped using 1447 Diversity Array Technology (DArT) markers generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au). The DArT markers may take on two values, denoted by their presence or absence. In this dataset, the overall mean frequency of the minor allele was 0.5607, with a minimum of 0.0083 and a maximum of 0.9866. Markers with a minor allele (coded as 1) frequency lower than 0.05 were removed. Missing genotypes were imputed using samples from the marginal distribution of marker genotypes, that is, $x_{ij} \sim Bernoulli(\hat{p}_j)$, where $\hat{p}_j$ is the estimated allele frequency computed from the non-missing genotypes. The number of DArT markers after edition was 1,279.

Maize dataset. The maize dataset is from the Drought Tolerance Maize for Africa project of CIMMYT's Global Maize Program. This project focuses on developing drought tolerant maize for Africa and comprises several maize breeding programs operating in different West African countries in coordination with the tropical maize breeding program established in Mexico. The data used here come from a large study aimed at detecting chromosomal regions affecting drought tolerance and adaptive traits identified in global maize germplasm based on analyses of available marker data from genotyping a total of 300 tropical inbred lines genotyped with 1,148 SNPs.

No pedigree was available for this data. Traits analyzed for this study were grain yield (GY), female flowering (FFL) (or days to silking), and male flowering time (MFL) (or days to anthesis), as well as the anthesis-silking interval (ASI) evaluated on lines under severe drought stress (SS) and well-watered (WW) environments. Hereinafter we will refer to these datasets as Maize-Flowering (M-F) and Maize-Grain Yield (M-GY). The number of lines in the M-F dataset was 284, whereas 264 lines were available in M-GY. The average minor allele frequency in these datasets was 0.20. Markers in each of these datasets were subjected to the edition and imputation procedures described above; after editing, the numbers of SNPs available for analysis were 1,148 and 1,135 in M-F and M-GY, respectively.

Statistical models

This study evaluated several models for GS that differ depending on the type of information used for constructing predictions (pedigree, markers, or both) and on how molecular markers were incorporated into the model (parametric vs. semi-parametric). All the unknowns in the model were trait-environment specific, and, consequently, separate models

were fitted to each trait-environment combination. For ease of presentation, models are described for a generic trait-environment.

Likelihood function. In all models, phenotypic records were described as $y_i = \mu + g_i + \varepsilon_i$, where μ is an intercept, g_i is the genetic value of the i^{th} line, and ε_i is a model residual. In all environments, the response variable was standardized to a sample variance equal to one. The joint distribution of model residuals was $p(\boldsymbol{\varepsilon}) = \prod_{i=1}^{n} N\left(\varepsilon_i | 0, \frac{\sigma^2_\varepsilon}{n_i}\right)$, where n_i is the number of replicates used for computing the mean value of the i^{th} genotype in the corresponding environments. Under this assumption, the likelihood function becomes

$$p(\mathbf{y}|\mu, \mathbf{g}, \sigma^2_\varepsilon) = \prod_{i=1}^{n} N\left(y_i | \mu + g_i, \frac{\sigma^2_\varepsilon}{n_i}\right), \qquad [1]$$

where $\mathbf{g} = \{g_i\}$ is a vector of genetic values. Models differed on how pedigree and molecular marker information was used to describe g_i. In the following sections, we present the different classes of models used to incorporate pedigree and marker data and either parametric or semi-parametric methods for describing g_i.

Standard infinitesimal models. In this model, denoted as P (standing for pedigree), $g_i = u_i$ and $p(\mathbf{u}|\sigma^2_u) = N(\mathbf{u}|\mathbf{0}, \mathbf{A}\sigma^2_u)$, where $\mathbf{A}$ is the additive relationship matrix computed from the pedigree and σ^2_u is an additive genetic variance. Following standard assumptions, the joint prior of model unknowns in P was

$$p(\mu, \mathbf{u}, \sigma^2_\varepsilon, \sigma^2_u | df_\varepsilon, S_\varepsilon, df_u, S_u) \propto N(\mathbf{u}|\mathbf{0}, \mathbf{A}\sigma^2_u)\chi^{-2}(\sigma^2_\varepsilon | df_\varepsilon, S_\varepsilon)\chi^{-2}(\sigma^2_u | df_u, S_u) \qquad [2a]$$

where $\chi^{-2}(\sigma^2_\cdot | df_\cdot, S_\cdot)$ are Scaled Inverse Chi-squared priors assigned to the variance parameters. The prior scale and degree of freedom parameters were set to $S_\cdot = 1$ and , $df_\cdot = 4$, respectively. This prior has finite variance and an expectation of 0.5 (the density plot for this prior is given in Figure 1A of Appendix A). Combining [1] and [2a], the joint posterior distribution of P is

$$p(\mu, \mathbf{u}, \sigma^2_\varepsilon, \sigma^2_u | \mathbf{y}, H) \propto \prod_{i=1}^{n} N\left(y_i | \mu + g_i, \frac{\sigma^2_\varepsilon}{n_i}\right) \times N(\mathbf{u}|\mathbf{0}, \mathbf{A}\sigma^2_u)\chi^{-2}(\sigma^2_\varepsilon | df_\varepsilon, S_\varepsilon)\chi^{-2}(\sigma^2_u | df_u, S_u) \qquad [2b]$$

Above, H denotes all hyper-parameters indexing the prior distribution. This posterior distribution has no closed form; however, samples from the above model can be obtained from a Gibbs sampler, as described, for example, in SORENSEN and GIANOLA (2002). An R-program (R DEVELOPMENT CORE TEAM, 2009) that implements a Gibbs sampler for this model is provided in supplementary materials. No pedigree data were available for the maize dataset; therefore, this model was only evaluated in the wheat dataset.

Parametric genomic models. Two parametric GS models were used, the standard Best Linear Unbiased Prediction (BLUP) (MEUWISSEN *et al.*, 2001; BERNARDO and YU, 2007) and the Bayesian LASSO (BL) (PARK and CASELLA, 2008). These two models regress phenotypic

outcomes on marker covariates; they differ in the way marker effects are estimated, as discussed below.

Consider a linear regression of the form $\mathbf{y}=\mu+\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ and assume $p\left(\boldsymbol{\varepsilon},\boldsymbol{\beta}\middle|\sigma_\varepsilon^2,\sigma_\beta^2\right)=N\left(\boldsymbol{\varepsilon}\middle|\mathbf{0},Diag\left\{n_i^{-1}\right\}\sigma_\varepsilon^2\right)N\left(\boldsymbol{\beta}\middle|\mathbf{0},\mathbf{I}\sigma_\beta^2\right)$. From this model, the BLUP estimates of marker effects are (e.g., ROBINSON, 1991)

$$\begin{aligned}E\left(\boldsymbol{\beta}\middle|\mathbf{y},\mu,\sigma_\varepsilon^2,\sigma_\beta^2\right)&=Cov\left(\boldsymbol{\beta},\mathbf{y}'\right)Var\left(\mathbf{y}\right)^{-1}\left(\mathbf{y}-\mathbf{1}\mu\right)\\&=Cov\left(\boldsymbol{\beta},\mathbf{1}'\mu+\boldsymbol{\beta}'\mathbf{X}'+\boldsymbol{\varepsilon}'\right)Var\left(\mathbf{1}\mu+\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right)^{-1}\left(\mathbf{y}-\mathbf{1}\mu\right)\\&=\sigma_\beta^2\mathbf{X}'\left[\sigma_\beta^2\mathbf{X}\mathbf{X}'+Diag\left\{\frac{\sigma_\varepsilon^2}{n_i}\right\}\right]^{-1}\left(\mathbf{y}-\mathbf{1}\mu\right).\end{aligned}$$

Computation of BLUPs requires knowledge of $\left\{\mu,\sigma_\varepsilon^2,\sigma_\beta^2\right\}$. To this end we fitted a random effects model, $y_{ik}=\mu+g_i+\varepsilon_{ik}$, where y_{ik} is the observed phenotype of the k^{th} replicate of the i^{th} genotype ($i=1,...,g\,;k=1,...,n_i$). This model yields estimates of $\left\{\mu,\sigma_\varepsilon^2,\sigma_g^2\right\}$, where $Var\left(g_i\right)=\sigma_g^2$. An estimate of σ_β^2 was obtained by plugging the estimate of σ_g^2 in $\sigma_\beta^2=\dfrac{\sigma_g^2}{\sum_j 2\theta_j\left(1-\theta_j\right)}\approx\dfrac{\sigma_g^2}{2p\bar{\theta}\left(1-\bar{\theta}\right)}$ (e.g., MEUWISSEN *et al.*, 2001; VANRADEN, 2007) where θ_j is the allelic frequency of the j^{th} marker, and $\bar{\theta}$ is an average (across markers) allele frequency, which in our case was estimated from the marker data.

The Bayesian LASSO method

The Bayesian LASSO (BL) of PARK and CASELLA (2008) provides an alternative way of estimating marker effects. Moreover, this model can be extended to accommodate an infinitesimal effect as well, as described in DE LOS CAMPOS *et al*. (2009a). When an infinitesimal effect is included (u_i) together with a regression on marker genotypes, $g_i=\sum_{j=1}^{p}x_{ij}\beta_j+u_i$, the data-equation is $y_i=\mu+\sum_{j=1}^{p}x_{ij}\beta_j+u_i+\varepsilon_i$, and the joint prior density of the model unknowns (upon assigning a flat prior to μ) is

$$\begin{aligned}p\left(\mu,\mathbf{u},\boldsymbol{\beta},\lambda,\sigma_\varepsilon^2,\sigma_u^2\middle|\,r,\delta,df_\varepsilon,S_\varepsilon,df_u,S_u\right)\propto{}&N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{A}\sigma_u^2\right)\left\{\prod_{j=1}^{p}N\left(\beta_j\middle|0,\sigma_\varepsilon^2\tau_j^2\right)\right\}\\&\times\left\{\prod_{j=1}^{p}Exp\left(\tau_j^2\middle|\lambda^2\right)\right\}G\left(\lambda^2\middle|r,\delta\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_u^2\middle|df_u,S_u\right)\end{aligned}\qquad[3a]$$

Above, marker effects are assigned independent Gaussian priors with marker-specific variances, $\beta_j\sim N\left(\beta_j\middle|0,\sigma_\varepsilon^2\tau_j^2\right)$. At the next level of the hierarchical model, the τ_j^2's are assigned IID exponential priors, $\tau_j^2\overset{IID}{\sim}Exp\left(\tau_j^2\middle|\lambda^2\right)$. At a deeper level of the hierarchy, the regularization parameter, λ^2, is assigned a Gamma prior with rate (δ) and shape (r), which in this study were set to $\delta=1\times10^{-4}$ and $r=0.6$. Finally, independent Scaled Inverse Chi-squared priors were assigned to the variance parameters, and the scale and degree of freedom

parameters were set to $S_u = S_\varepsilon = 1$ and , $df_\varepsilon = df_u = 4$, respectively. The above model is referred as to PM-BL.

The effect of the prior choice for λ in the BL has been addressed in DE LOS CAMPOS *et al*. (2009a); these authors studied the influence of the choice of hyper-parameters for λ on inferences of several items and concluded that, even when the prior for λ had influence on inferences about this unknown, model goodness of fit and estimates of genetic values were robust with respect to the choice of $p(\lambda)$. Figure 2A (Appendix A) depicts the prior density of λ corresponding to the values of the hyper-parameters used in this study; this prior gave high density over a wide range of values of λ. Also, as shown later, the posterior mean of λ did change between traits and datasets, indicating that the posterior distribution moved away from the prior.

Combining the assumptions of the likelihood [1] and the prior described in [3a], the joint posterior distribution is

$$\begin{aligned} p(\mu,\mathbf{u},\boldsymbol{\beta},\lambda,\sigma^2_\varepsilon,\sigma^2_u|\mathbf{y},H) \propto & \left\{\prod_{i=1}^{n} N\left(y_i|\mu+g_i,\frac{\sigma^2_\varepsilon}{n_i}\right)\right\} N\left(\mathbf{u}|\mathbf{0},\mathbf{A}\sigma^2_u\right) \\ & \times\left\{\prod_{j=1}^{p} N\left(\beta_j|0,\sigma^2_\varepsilon\tau^2_j\right)\right\}\left\{\prod_{j=1}^{p} Exp\left(\tau^2_j|\lambda^2\right)\right\} \\ & \times G\left(\lambda^2|r,\delta\right)\chi^{-2}\left(\sigma^2_\varepsilon|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma^2_u|df_u,S_u\right) \end{aligned} \qquad [3b]$$

This posterior distribution has no closed form; however, samples from the above model can be obtained from a Gibbs sampler, as described in DE LOS CAMPOS *et al.* (2009a). An R-program (R DEVELOPMENT CORE TEAM, 2009) that implements a Gibbs sampler for this model is provided as supplementary material.

The marker-based model, M-BL, is a special case of the above model [3b] with $\mathbf{u} = \mathbf{0}$, which implies that $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$.

Semi-parametric models (RKHS). As previously stated, in RKHS, the genetic values are viewed as a Gaussian process. When markers and a pedigree are available, genetic values can be modeled as the sum of two components, $g_i = u_i + f_i$, where u_i is as before and f_i is a Gaussian process with a (co)variance function proportional to the evaluations of a reproducing kernel, $\mathrm{K}(\mathbf{x}_i,\mathbf{x}_j)$, evaluated in marker genotypes; here $\mathbf{x}_i$ and $\mathbf{x}_j$ are vectors of marker genotype codes for the i^{th} and j^{th} individuals, respectively. The joint prior distribution of $\mathbf{u} = \{u_i\}$, $\mathbf{f} = \{f_i\}$ and the associated variance parameters σ^2_ε, σ^2_u, and σ^2_f, is as follows:

$$\begin{aligned} p(\mu,\mathbf{u},\mathbf{f},\sigma^2_\varepsilon,\sigma^2_u,\sigma^2_f|df_\varepsilon,S_\varepsilon,df_u,S_u,df_f,S_f) \propto & N\left(\mathbf{u}|\mathbf{0},\mathbf{A}\sigma^2_u\right)N\left(\mathbf{f}|\mathbf{0},\mathbf{K}\sigma^2_f\right) \\ & \times\chi^{-2}\left(\sigma^2_\varepsilon|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma^2_u|df_u,S_u\right)\chi^{-2}\left(\sigma^2_f|df_f,S_f\right) \end{aligned} \qquad [4a]$$

Above, **K** is a kernel-matrix, which is symmetric and positive. In this study, the entries of these matrices were the evaluations of a Gaussian kernel, $\mathrm{K}(\mathbf{x}_i,\mathbf{x}_j) = \exp\left\{-2\frac{d_{ij}}{q_{.5}}\right\}$, where $d_{ij} = \sum_{k=1}^{p}(x_{ik}-x_{jk})^2$ is a squared-Euclidean distance, and $q_{.5}$ is the sample median of the

matrix of sampled squared-Euclidean distances $\{d_{ij}\}$. Note that if $d_{ij}=q_{.5}$, this choice yields $\mathrm{K}(\mathbf{x}_i,\mathbf{x}_j)=\exp(-2)\approx 0.13$, which implies that a prior correlation of 0.13 is assigned to pairs of lines whose squared-Euclidean distances are equal to the median squared-Euclidean distance, and higher (lower) prior correlation is assigned to pairs of lines that are closer (farther apart) than $q_{.5}$, in the sense of the squared-Euclidean distance. This choice of kernel may not yield the highest predictive ability; however, addressing the problem of kernel choice is beyond the scope of the current study. The scale and degree of freedom parameters of the prior described in [4a] were $S_\varepsilon = S_u = S_f = 1$ and, $df_\varepsilon = df_u = df_f = 4$.

Combining the assumptions in [1] and [4a], the joint posterior distribution of this marker and pedigree model (PM-RKHS) is

$$\begin{aligned} p\left(\mu,\mathbf{u},\mathbf{f},\sigma^2_\varepsilon,\sigma^2_u,\sigma^2_f\middle|\mathbf{y},H\right) \propto & \left\{\prod_{i=1}^{n} N\left(y_i\middle|\mu+f_i+u_i,\frac{\sigma^2_\varepsilon}{n_i}\right)\right\} \\ & \times N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{A}\sigma^2_u\right)N\left(\mathbf{f}\middle|\mathbf{0},\mathbf{K}\sigma^2_f\right) \\ & \times \chi^{-2}\left(\sigma^2_\varepsilon\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma^2_u\middle|df_u,S_u\right)\chi^{-2}\left(\sigma^2_f\middle|df_f,S_f\right) \end{aligned} \qquad [4b]$$

The above joint posterior distribution has no closed form; however, draws samples from this posterior distribution can be obtained using a slightly modified version of the Gibbs sampler that implements the pedigree model in [2a]. An R-program (R DEVELOPMENT CORE TEAM, 2009) that implements a Gibbs sampler for this model is provided in supplementary materials.

As with parametric methods, a marker-based model, M-RKHS, can be obtained as a particular case of PM-RKHS, described in [4b], with $\mathbf{u}=\mathbf{0}$, which implies $g_i = f_i$.

Data Analysis

Full-data analysis. Models were first fitted using all lines in the training set, and inferences for each fit were based on 30,000 samples (obtained after discarding 5,000 samples as burn-in; the thinning interval between consecutive observations used in all simulations was 10). Convergence was checked by inspecting trace plots of variance parameters.

Using estimates of marker effects of M-BL from the full-data analysis, we performed principal component analysis of estimated marker effects. Results are displayed in a biplot of the first two principal component axes for estimates of molecular markers and the trait-environment combination (see Appendix B for further discussion of this).

Cross-validation. Cross-validation (CV) methods can be used to evaluate the ability of a model to predict future outcomes. Here, we designed the CV scheme so as to address the following question: What is the expected performance of a genotype with yet-to-be observed phenotypes (e.g., newly developed lines)? This is one of the most important prediction problems plant breeders face. Predictions of performance of lines whose phenotypes are yet to be observed can be used, for example, to decide which of the newly generated lines will be evaluated in field trials. To this end, we divided the data into ten folds, by using an index variable, $I_i \in \{1,...,10\}$, *i=1,..,n*, that randomly assigns observations to ten disjoint folds, $F_j = \{i : I_i = j\}$, *j=1,..,10*. CV predictions of the observations in the first fold, $F_1 = \{i : I_i = 1\}$, can be obtained by fitting models with all lines in fold 1 regarded as missing data. This yields CV predictions of lines in the first fold, that is, $\{\hat{y}_i : I_i = 1\}$. Repeating this exercise for the 2nd, 3rd,.., 10*th* folds 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 our dataset, the response is the average

performance of each line; thus each line appears only once in the data. As a result, CV predictions obtained as described above are based completely on the information on other genotypes' performance.

The wheat and maize experimental data, and all the required computer programs written in R (R DEVELOPMENT CORE TEAM, 2009) that fit the BL and RKHS models described above are made available in supplementary materials.

Multivariate analysis of estimated marker effects

Parametric models as those described in the previous section (BLUP, M-BL, and PM-BL) yield estimates of marker effects which, in our case, are environment-specific. An important volume of literature describes how biplots from singular-value decomposition can be used to assess G×E (e.g., CORNELIUS *et al.,* 2001). We used these techniques (see Appendices B and C) to study G×E at the level of estimated marker effects from PM-BL or M-BL models.

RESULTS

We begin the section by presenting parameter estimates obtained when models BL and RKHS were fitted using all available records (i.e., no line was regarded as a missing value; in other words, a full data model) in the training set. The evaluation of predictive ability is introduced later on.

Variance components

Tables 1a and 1b give the estimates of posterior means of variance parameters and of the BL parameter λ, by trait-environment combination and models. The posterior mean of the residual variance (σ^2_ε) can be used to assess model goodness of fit. Since the response variable was standardized within trait-environment combinations, the estimate of the residual variance gives an indication of the percentage of the phenotypic variance that is attributable to model residuals. In the GY-W dataset (Table 1a), RKHS models fitted data markedly better (smaller σ^2_ε) than P, M-BL, or PM-BL; M-BL had a posterior mean of residual variance that was either close to or higher than P, while PM-BL fitted the data better than P.

Results from the maize datasets were mixed: M-BL fitted the data much better than M-RKHS for FFL and FLM, regardless of moisture conditions, but the opposite was true (i.e., M-RKHS fitted data better than M-BL) for ASI and GY (Table 1b).

The variances of u_i and f_i can be used as a measure of the relative contribution of each of these components to the conditional expectation function in models where these components are present (i.e., P, M-RKHS, and PM-RKHS). From [4a], $Var(u_i) = a(i,i)\sigma^2_u$, where $a(i,i)$ is the ith diagonal element of matrix **A**, and $Var(f_i) = \mathrm{K}(\mathbf{x}_i, \mathbf{x}_i)\sigma^2_f$; $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_i)$ is a standardized kernel, with $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_i) = 1$. This does not occur in $a(i,i)$; here $a(i,i) = 1 + F_i$, where F_i is the coefficient of inbreeding for the i^{th} individual. In the wheat population, the average value of $a(i,i)$ was 1.98. For the W-GY dataset, the posterior mean of σ^2_u was smaller in PM-BL and PM-RKHS relative to P models for all four environments (Table 1a), and the ratio $\frac{\sigma^2_f}{a(i,i)\sigma^2_u}$ evaluated at $a(i,i) = 1.98$ and at the posterior mean of σ^2_f and of σ^2_u was 3.49 (GY-E1), 2.74 (GY-E2), 2.13 (GY-E3), and 2.54 (GY-E4). These results indicate that in PM

models the regression on the markers (f_i) makes a much stronger contribution to estimates of genetic values than the regression on the pedigree (u_i).

TABLE 1a

Estimates of posterior means of parameters σ_ε^2 , σ_u^2 , σ_f^2 , and λ from the full-data analysis of grain yield (GY) of 599 historical wheat lines genotyped with 1,279 DArTs molecular markers. Five models* were fitted to each trait (GY) and environment (E1, E2, E3, E4) combination.

Trait-environment	Model	Parameter			
		σ_ε^2	σ_u^2	σ_f^2	λ
GY-E1	P	0.562	0.286	---	---
	M-RKHS	0.272	---	0.825	---
	PM-RKHS	0.197	0.108	0.746	---
	M-BL	0.554	---	---	20.389
	PM-BL	0.434	0.141	---	20.747
GY-E2	P	0.581	0.248	---	---
	M-RKHS	0.394	---	0.720	---
	PM-RKHS	0.364	0.115	0.531	---
	M-BL	0.574	---	---	21.994
	PM-BL	0.501	0.117	---	24.927
GY-E3	P	0.492	0.342	---	---
	M-RKHS	0.317	---	0.888	---
	PM-RKHS	0.283	0.148	0.625	---
	M-BL	0.667	---	---	26.924
	PM-BL	0.479	0.237	---	37.423
GY-E4	P	0.517	0.300	---	---
	M-RKHS	0.330	---	0.771	---
	PM-RKHS	0.298	0.118	0.594	---
	M-BL	0.612	---	---	24.725
	PM-BL	0.471	0.169	---	27.503

***** The five models are: pedigree model (P), molecular marker regression model using Bayesian LASSO (M-BL), pedigree (P) plus the molecular marker model regression using Bayesian LASSO (PM-BL), molecular marker model using reproducing kernel Hilbert space (M-RKHS) regression, and pedigree (P) plus molecular marker model using reproducing kernel Hilbert space (PM-RKHS) regression. Estimates of posterior standard deviations (across traits and models) ranged from 0.041, 0.028, 0.093 and 2.73, to 0.057, 0.060, 0.132 and 11.73 for σ_ε^2 , σ_u^2 , σ_f^2 , and λ , respectively.

TABLE 1b

Estimates of posterior means of parameters σ^2_ε, σ^2_f, and λ from the full-data analysis of female flowering time (FFL), male flowering time (MFL), the MFL to FFL interval (ASI) of 284 maize genotypes and 1,148 SNPs, and grain yield (GY) of 264 genotypes and 1,135 SNPs. Two models* were fitted to each of the trait (FFL, MFL, ASI, and GY) and environment (SS=severe stress; WW=well-watered) combinations.

		Parameter		
Trait-environment	Model	σ^2_ε	σ^2_f	λ
MFL-WW	M-RKHS	0.761	0.262	---
	M-BL	0.315	---	28.2
MFL-SS	M-RKHS	0.402	0.645	---
	M-BL	0.169	---	18.6
FFL-WW	M-RKHS	0.793	0.241	---
	M-BL	0.323	---	28.4
FFL-SS	M-RKHS	0.489	0.566	---
	M-BL	0.179	---	18.9
ASI-WW	M-RKHS	0.231	0.700	---
	M-BL	0.467	---	41.8
ASI-SS	M-RKHS	0.183	0.747	---
	M-BL	0.370	---	32.9
GY-WW	M-RKHS	0.252	0.725	---
GY-WW	M-BL	0.369	---	31.069
GY-SS	M-RKHS	0.212	0.836	---
GY-SS	M-BL	0.431	---	33.365

* The two models are: molecular marker (M) regression model using Bayesian LASSO (M-BL) and molecular marker (M) using reproducing kernel Hilbert space (M-RKHS) regression. Estimates of posterior standard deviations (across traits and models) ranged from 0.049, 0.096 and 4.014, to 0.124, 0.168 and 8.619 for σ^2_ε, σ^2_f, and λ, respectively.

Marker effects

A multivariate analysis of estimated marker effects can be performed using the singular value decomposition of a matrix whose rows pertain to markers and columns give the estimated effect of the markers for different traits or environments. Appendix B describes the methodology in detail. We illustrate its use below with a multivariate analysis of estimates of marker effects from the GY wheat dataset.

The first and second components from the singular value decomposition on the matrix of estimated effects, $\hat{\mathbf{B}} = \left[\hat{\boldsymbol{\beta}}_1,..,\hat{\boldsymbol{\beta}}_4\right] = \left\{\hat{\beta}_{jk}\right\}$, computed from the PM-BL model in each

environment, $\hat{\boldsymbol{\beta}}_k$ ($k=1,..,4$), of W-GY data are depicted in the biplot of Fig. 1. The first two component axes explained 74.44% of the total variability in estimated DArT effects; the phenotypic correlations between estimated effects in the four environments showed that E2 and E3 had a correlation of 0.661, whereas E2 and E4, and E3 and E4 had correlations of 0.411 and 0.388, respectively (Table S1 in supplementary materials). On the other hand, estimates of marker effects for E1 had lower correlations with those for E2, E3, and E4 (-0.020, -0.193, and -0.123, respectively).

The pattern of correlations between estimated DArT effects reflects the patterns of phenotypic correlations observed for W-GY. Environment E1 causes a great deal of the interaction between molecular marker effects and the other environments due to its low and even negative correlations with E2, E3, and E4. It should be pointed out that the correlation between the four mega-environments in this study may not reflect their associations in later years very well due to the dynamics of climate change prevailing in many regions of the world.

The variance of marker effects was slightly smaller in E4, as can be inferred by the length of the corresponding vector in Fig. 1. The vast majority of the estimated effects are located around the center of the figure (i.e., estimated effects were small, in absolute value), which reflects shrinkage of the BL model. However, some DArTs had estimated effects that were large in absolute value; those DArTs are identified by name in Fig. 1, and their effects are shown in Table S2 (supplementary materials).

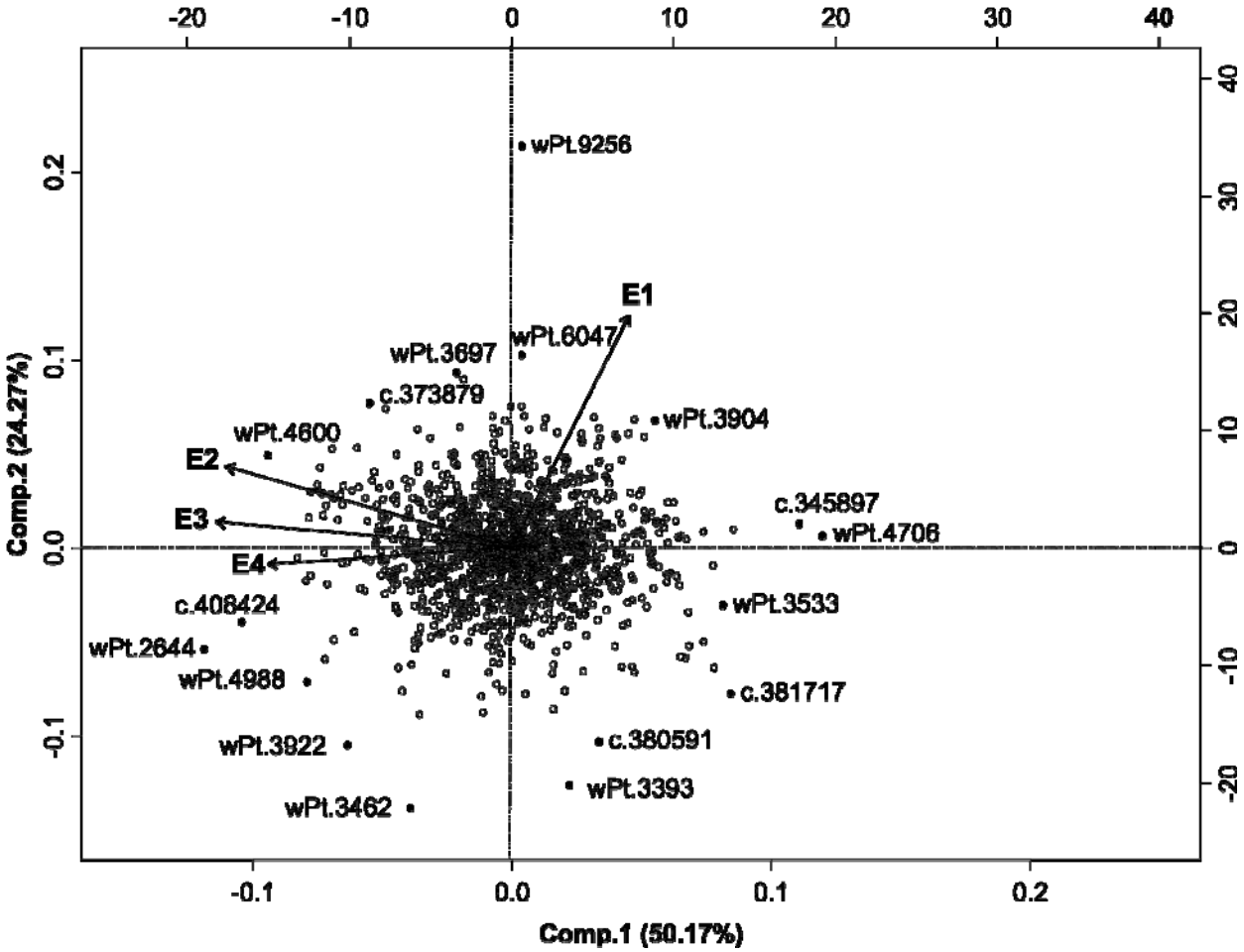
The estimated effect of the presence of a DArT in GY for a given environment can be obtained by orthogonal projection of the marker effect displayed in Fig. 1 on the vector of the corresponding environment. To illustrate this, consider E1, where the presence of DArTs wPt.9256, wPt.6047, and wPt.3904 is expected to increase GY (Fig. 1 and Table S2, supplementary materials); in contrast, the presence of DArTs wPt.3462, wPt.3922, and wPt.4988 (located in the opposite direction of E1) is expected to reduce GY.

Those DArTs whose presence is expected to increase or decrease GY across environments can be viewed as contributing to positive genetic correlations in GY between environments. Examples of this group are DArTs wPt.9256, wPt.6047, and c.373879, whose presence increased GY in the four environments; and wPt.3393, c.380591, and c.381717, whose presence decreased GY in all environments. However, some DArTs act in an 'antagonistic' fashion, that is, the presence of a DArT increases (decreases) GY in some environments and decreases (increases) GY in others. Examples of this group are c.408424, wPt.2644, wPt.4988, and wPt.3462, whose presence decreased GY in E1 and is predicted to increase GY in all other environments; and wPt.3904, c.345897, and wPt.4706, whose presence is expected to increased GY in E1 and to decrease GY in all other environments.

The effects of all 1,279 DArTs (with the corresponding chromosome number) in each of the four environments (E1-E4) can be found in Table S2a (supplementary materials). The scores of the first two components axes of the 1,279 DArTs are shown in Table S6 (supplementary materials).

Results from multivariate analysis of estimated effects on the maize flowering and grain yield datasets are given in Appendix C.

Figure 1. Biplot of the first and second principal component axes (Comp. 1 and Comp. 2) of the grain yield (GY) effect of the 1,279 DArTs estimated from the full data model PM-BL of the wheat dataset in each of four mega-environments (E1, E2, E3, E4). Only the effects of 17 DArTs that are located far from the center of the biplot were identified with their corresponding DArT's name (filled-in circles).



Predictive ability

Tables 2a and 2b show the estimated correlations between phenotypic outcomes and cross-validation (CV) predictions for W-GY, M-F, and M-GY datasets, respectively. Overall, the values of these correlations, especially those obtained with BL or RKHS methods, were large across models, datasets, and traits, indicating that genomic selection can be effective for predicting the genetic value of lines. Nevertheless, predictive ability was different between models, as discussed next.

In the W-GY, correlations ranged from 0.355 (BLUP in E3) to 0.608 (PM-RKHS in E1) (Table 2a), and relative to the P model, the PM-RKHS model showed the highest percent gain in CV-correlation in three out of four environments. BLUP was outperformed by BL and RKHS methods across environments. For these data, PM models had better predictive ability than P models, and the magnitude of the gain in predictive ability attained by including markers in the model varied, from a modest 7.7% (PM-BL in GY-E3) to a very important 35.7% (PM-RKHS in GY-E1) (Table 2a). In general, RKHS outperformed BL in both M and PM, and BLUP outperformed P models in three out of four environments (all but E3); however, as stated, BLUP was outperformed by BL and RKHS.

In the M-F, correlations ranged from 0.464 (BLUP for MFL-SS) to 0.790 (M-BL for MFL-WW) (Table 2b). BLUP was systematically outperformed by BL and RKHS for these traits, while M-BL yielded better predictions than M-RKHS, with relatively high correlation values, ranging from 0.774 to 0.790. However, for ASI under severe drought stress and well-

watered conditions, correlations were not as strong as those found for the other flowering time traits, and M-RKHS outperformed M-BL, with correlation values of 0.547 and 0.572, respectively (Table 2b).

Predictive correlations in M-GY (Table 2b) were somehow smaller than those obtained in flowering traits, and the differences between methods were not as marked. Here, CV correlations ranged from 0.415 (M-BL GY under drought stress) to 0.525 (M-BL GY well-watered). These traits did not yield a clear ranking of models: BL was best for GY under well-watered conditions, and RKHS was best for GY under drought stress; however, as stated, in M-GY, the differences in predictive ability between methods were not marked.

TABLE 2a

Predictive ability measured as the correlation between predicted and actual phenotypes, obtained in a 10-fold cross-validation, from the analysis of grain yield (GY) of 599 historical ESWYT wheat lines genotyped with 1,279 DArT molecular markers. Six models* were fitted to GY measured in four environments (E1, E2, E3, E4). Changes relative to the pedigree model (P) are presented as percentages.

	Model					
Trait-environment	P	M-RKHS	PM-RKHS	M-BL	PM-BL	BLUP[+]
			Correlation			
GY-E1	0.448	0.601	0.608	0.518	0.542	0.480
GY-E2	0.417	0.494	0.497	0.493	0.501	0.488
GY-E3	0.417	0.445	0.478	0.403	0.449	0.355
GY-E4	0.449	0.524	0.524	0.457	0.495	0.464
			% change (relative to P)			
GY-E1	---	34.2	35.7	15.6	21.0	7.1
GY-E2	---	18.5	19.2	18.2	20.1	17.0
GY-E3	---	6.7	14.6	-3.4	7.7	-14.9
GY-E4	---	16.7	16.7	1.8	10.2	3.3

* The six fitted models are: pedigree model (P), molecular marker regression model using Bayesian LASSO (M-BL), pedigree (P) plus molecular marker model regression using Bayesian LASSO (PM-BL), molecular marker model using reproducing kernel Hilbert space (M-RKHS) regression, pedigree (P) plus molecular marker model using reproducing kernel Hilbert space (PM-RKHS) regression, and the BLUP method.

[+] The range of genetic variance components used for BLUP estimation was 0.8065-0.9141.

TABLE 2b

Predictive ability measured as the correlation between predicted and actual phenotypes, obtained in a 10-fold cross-validation from evaluating female flowering (FFL), male flowering (MFL), the MFL to FFL interval (ASI) of 284 maize genotypes and 1,148 SNPs, and grain yield (GY) of 264 maize genotypes and 1,135 SNPs. Each of the three models* was fitted to each combination of four traits (FFL, MFL, ASI, and GY) and two environments (SS=severe drought stress, WW=well-watered).

	Model		
Trait-environment	M-RKHS	M-BL	BLUP[#]
MFL-WW	0.607	0.790	---[+]
MFL-SS	0.674	0.778	0.464
FFL-WW	0.588	0.781	---
FFL-SS	0.648	0.774	0.521
ASI-WW	0.547	0.513	0.469
ASI-SS	0.572	0.517	0.481
GY-WW	0.514	0.525	0.515
GY-SS	0.453	0.415	0.442

* The two fitted models are: molecular marker (M) regression model using Bayesian LASSO (M-BL), molecular marker (M) using reproducing kernel Hilbert space (M-RKHS) regression, and the BLUP method.

[+] BLUPs were not computed since the estimated genetic variances were negligible.

[#] Ranges of genetic variance components used for BLUP estimation were 0.000-0.319 for flowering, and 0.017-0.206 for grain yield.

DISCUSSION

Results found in this study are encouraging and indicate that, even with a modest number of molecular markers, models for GS can attain relatively high predictive ability for genetic values of traits of economic interest in contrasting environmental conditions. This indicates that GS using BL and RKHS models with pedigree and molecular marker information can be used effectively for selecting individuals whose phenotypes for various traits have yet to be observed. These predictions can be used for developing several rounds of selection without phenotyping and for pre-selecting lines that will be evaluated in field trials. Overall, such practices can contribute to an important reduction in the generation interval and a simultaneous substantial reduction of phenotyping costs.

Our results show important gains in predictive ability relative to pedigree-based models. However, how good are these CV correlations relative to a maximum attainable predictive ability? Answering this question requires knowledge of the ('true') underlying model and of parameter values. As an exercise, let us assume that the model $y_i = g_i + \varepsilon_i$ holds, and also, as the best (unlikely) scenario, that CV predictions, $\hat{g}_{i,CV}$, are such that $\hat{g}_{i,CV} = g_i$. If so, the maximum attainable correlation is $Cor(g_i, y_i) = \left(\sigma_g^2 + \sigma_\varepsilon^2\right)^{-\frac{1}{2}} \sigma_g = h$, where h is the square root of the heritability of the trait. Thus, if heritability is 0.5, then the maximum correlation is 0.707. Above we have assumed that only one replicate is available. For repeated measures the maximum correlation is: $Cor(g_i, y_i) = \left(\sigma_g^2 + n_i^{-1}\sigma_\varepsilon^2\right)^{-\frac{1}{2}} \sigma_g > h$. CV correlations in this study ranged from 0.40 to 0.79; these values are below the theoretical maxima, and larger gains in predictive ability may be expected when more markers are available. However, even

with complete sequencing, the maximum correlation may not be attained due to, for example, inability of the model/experimental design to completely uncover genetic signals.

Predictive ability of models. In general, M and PM in wheat had similar predictive abilities; this is in agreement with previous findings (e.g., DE LOS CAMPOS *et al*., 2009a) and occurs because there is some redundancy between the regression on the pedigree and the regression on markers (e.g., HABIER *et al*., 2009). As the number of molecular markers increases, it is reasonable to expect that the relative contribution of pedigree information will decrease. Marker-based models in this study included BL, RKHS, and BLUP. Overall, BL and RKHS outperformed BLUP in most instances, except in M-GY, where differences in predictive ability of each of these methods were small.

The high predictive ability found for maize flowering and grain yield under drought stress conditions is encouraging for the usefulness of GS selection under these conditions, which are becoming the rule rather than the exception all over the world due to the dynamics of climate change.

As previously stated, the comparison between BL and RKHS is not strictly fair, because results from RKHS could be improved if the kernel is selected based on predictive ability or some other criterion, an issue not addressed in this study. However, it should be noted that while in some trait-environment combinations RKHS outperformed BL (in the wheat dataset, and for GY under severe water stress and ASI in maize), in others (e.g., FFL, MFL, and GY under well-watered conditions), the opposite was true. This indicates that the problem of model choice is population-trait-environment specific and that a 'one-size-fits-all' approach to model choice in GS is not appropriate.

An advantage of parametric methods such as M-BL and PM-BL is that, in addition to estimating genetic values, these models also provide information on 'marker effects' that can be used to gain a better understanding of the underlying architecture of the traits and of genotype × environment interaction. The multivariate study of estimates of marker effects obtained from single-trait models presented in this study provides a way of identifying which markers contribute to positive genetic correlations, and which act in an 'antagonistic' fashion. This should be taken into account when constructing selection indices that consider multiple breeding goals (e.g., yield and yield stability across environments).

CONCLUDING REMARKS

Genomic selection appears to be a very promising tool for plant breeding. However, most studies so far come from simulations, and only a few studies have quantified the predictive ability of models for GS in real plant populations evaluated under different environmental conditions. We studied the performance of GS in two extensive international wheat and maize trials that include different traits and environments. The extensive cross-validation scheme used in this study showed that models including markers or markers and pedigrees yield relatively high correlations between predicted and observed phenotypic outcomes, and that the inclusion of molecular markers in pedigree-based models yielded an important increase in predictive ability, relative to pedigree-based models. This occurred even when a relatively modest number of markers was available.

Evidence from this study indicates that GS can be an effective strategy for selecting among lines whose phenotypes have yet to be observed. Denser markers will become available soon, and this may further improve the ability of GS to predict genetic values.

Results of this study show that while RKHS outperformed BL in some traits/environments, the opposite was true in other traits/environments. On the other hand, the standard BLUP was outperformed by BL and RKHS in almost all traits, except in M-GY,

where differences between models were only modest. Although the predictive ability of RKHS models may be improved by suitable choice of the kernel, results seem to indicate that a 'one-size-fits-all' approach to the problem of model choice is not appropriate. The relatively promising results from RKHS indicate that designing methods to address the problem of kernel choice is a relevant area of research in the context of semi-parametric models for GS.

In this study, separate models were fitted to each trait-environment combination. Multiple-trait models are ubiquitous in plant and animal breeding, and the development and evaluation of multiple-trait and multiple-environment models for GS where marker effects and genomic values for several traits are estimated jointly appears as a relevant area of research.

ACKNOWLEDGMENTS

The present article benefited from the valuable comments of the associate editor, Dr. Matias Krist, and two anonymous reviewers. The maize dataset used in this study comes from the Drought Tolerance Maize for Africa project financed by the Melinda and Bill Gates Foundation. The authors would like to thank the numerous cooperators in national agricultural research institutes who carried out the maize trials in Africa and the Elite Spring Wheat Yield Trials (ESWYT), and provided the phenotypic data analyzed in this article. We also thank the International Nursery and Seed Distribution Units in CIMMYT-Mexico for preparing and distributing the seed and computerizing the data. Financial support by the Wisconsin Agriculture Experiment Station and grant DMS-NSF DMS-044371 to Gustavo de los Campos and Daniel Gianola is acknowledged.

REFERENCES

BERNARDO, R., and J. YU. 2007 Prospects for genome-wide selection for quantitative traits in maize. Crop Science **47**: 1082-1090.

BUCKLER, E.S., J.B. HOLLAND, P.J. BRADBURY, C.B. ACHARYA, P.J. BROWN *et al.* 2009 The genetic architecture of maize flowering time. Science **325**:714-718.

BURGUEÑO, J., J. CROSSA, P.L. CORNELIUS, R. TRETHOWAN, G. MCLAREN *et al*. 2007 Modeling additive × environment and additive × additive × environment using genetic covariances of relatives of wheat genotypes. Crop Science **43**, 311-320.

CORNELIUS, P.L., J. CROSSA, M.S. SEYEDSADR, G. LIU and K. VIELE. 2001 Contributions to multiplicative model analysis of genotype-environment data. In: Statistical Consulting Section, American Statistical Association, Joint Statistical Meetings. Atlanta, GA, Aug. 7, 2001.

CROSSA, J., J. BURGUEÑO, P.L. CORNELIUS, G. MCLAREN, R. TRETHOWAN *et al*. 2006 Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Science **46**,1722-1733.

CROSSA, J., J. BURGUEÑO, S. DREISIGACKER, M. VARGAS, S.A. HERRERA-FOESSEL *et al*. 2007 Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics **177**, 1889-1913.

DE LOS CAMPOS, G., H. NAYA, D. GIANOLA, J. CROSSA, A. LEGARRA *et al*. 2009a Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics **182**, 375-385.

DE LOS CAMPOS, G., D. GIANOLA, and G.J.M. ROSA. 2009b Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J. Anim. Sci. 87: 1883-1887.

FISHER, R.A. 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. **52**, 399-433.

GIANOLA, D., and R.L. FERNANDO. 1986 Bayesian methods in animal breeding theory. J. Anim. Sci. **63**, 217-244.

GIANOLA, D., M. PEREZ-ENCISO, and M.A. TORO. 2003 On marker-assisted prediction of genetic value: beyond the ridge. Genetics **163**: 347-365.

GIANOLA, D., R.L. FERNANDO, and A. STELLA. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics **173**, 1761-1776.

GIANOLA, D., and J.B.C.H.M VAN KAMM. 2008 Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics **178**, 2289-2303.

GODDARD, M.E., and B.J. HAYES. 2007 Genomic selection. J. Anim. Breed. Genet. **124**, 323-330.

GONZALEZ-RECIO, O., D. GIANOLA, N. LONG, K. WIEGEL, G.J.M. ROSA *et al*. 2008 Non parametric methods for incorporating genomic information into genetic evaluation: An application to mortality in broilers. Genetics **178**:2305-2313.

HABIER, D., R.L. FERNANDO, and J.C.M. DECKKERS. 2009 Genomic selection using low-density marker panels. Genetics **182**:343-353.

HAYES, B.J., P.J. BOWMAN, A.J. CHAMBERLAIN, and M.E. GODDARD. 2009 Invited review: Genomic selection in dairy cattle: Progress and challenges. J. of Dairy Science **92**, 433-443.

HENDERSON, C.R. 1984 Application of Linear Models in Animal Breeding. University of Guelph, Guelph, Ontario, Canada.

HOERL, A.E., and R.W. KENNARD. 1970 Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**:55-67.

LANGE, C., and J.C. WHITTAKER. 2001. On prediction of genetic values in marker-assisted selection. Genetics **159**:1375-1381.

MEUWISSEN, T.H.E., B.J. HAYES, and M.E. GODDARD. 2001 Prediction of total genetic values using genome-wide dense marker maps. Genetics **157**, 1819-1829.

MCLAREN, C.G., R. BRUSKIEWICH, A.M. PORTUGAL, and A.B. COSICO. 2005 The International Rice Information System. A platform for meta-analysis of rice crop data. Plant Physiology **139**: 637-642.

PARK, T., and G. CASELLA. 2008 The Bayesian LASSO. J. Am. Stat. Assoc. **103**, 681-686.

PIEPHO, H.P., J. MÖHRING, A.E. MELCHINGER, and A. BÜCHSE. 2007 BLUP for phenotypic selection in plant breeding and variety testing. Euphytica **161**:209-228.

PIEPHO, H.P. 2009 Ridge regression and extensions for genome-wide selection in maize. Crop Sci. **49**:1–12.

R DEVELOPMENT CORE TEAM. 2009 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

ROBINSON, G.K. 1991 That BLUP is a good thing: The estimation of random effects. Statistical Science 6(1): 15-51.

SORENSEN, D. and D. GIANOLA. 2002 Likelihood, Bayesian, and MCMC methods in quantiative genetics. Springer-Verlag, New York.

SCHAEFFER, L.R. 2006 Strategy for applying genome-wide selection in dairy cattle. J. Anim Breed. Genet. **123**, 218-223.

TIBSHIRANI, R. 1996 Regression shrinkage and selection via the LASSO. J. Royal. Statist. Soc. B. **58**: 267-288.

VANRADEN, P.M. 2007 Genomic measures of relationship and inbreeding. In: Interbull Annual Meeting Proceedings, Interbull Bulletin **37**:33-36.

VANRADEN, P.M., C.P. VAN TASSELL, G.R. WIGGANS, T.S. SONSTEGARD, R.D. SCHNABEL *et al*. 2008 Invited review: Reliability of genomic predictions for North American Holstein bulls. J. of Dairy Science **92**, 16-24.

Appendix A

Figure 1A. Plot of the prior density for variance components corresponding to the values of σ_ε^2 used in this study.

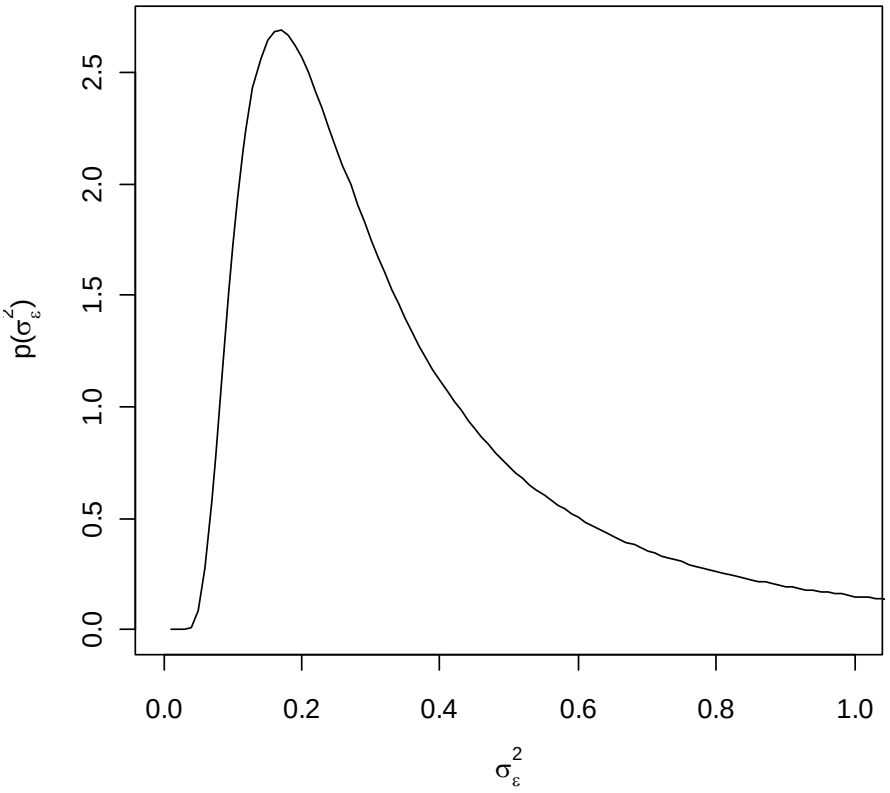
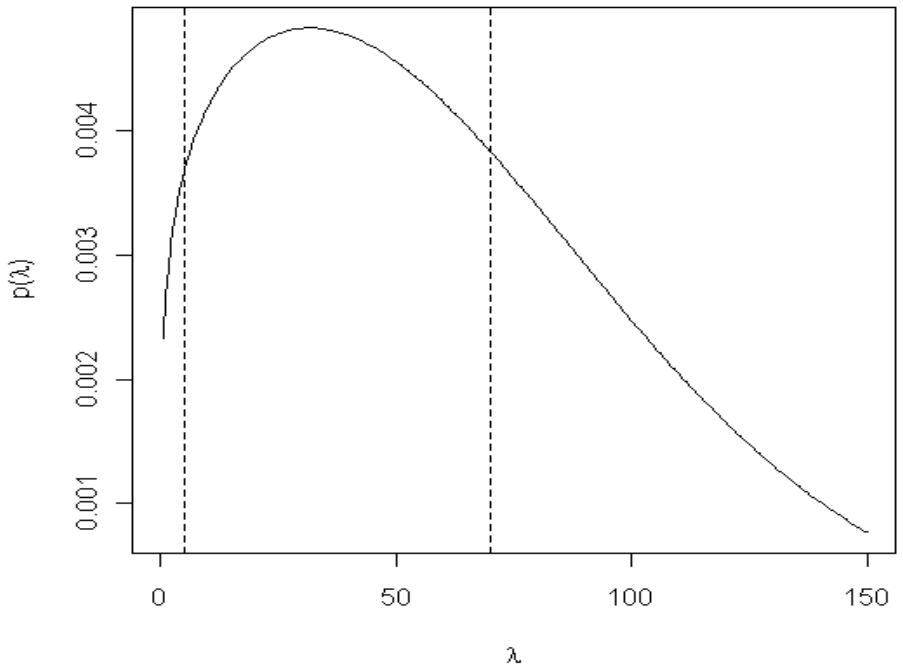


Figure 2A. Plot of the prior density of lambda, $p(\lambda)$, corresponding to the values of lambda (λ) used in this study.



Appendix B

Multivariate analysis of estimated marker effects

Consider a matrix of estimated molecular marker effects, $\hat{\mathbf{B}}_{p\times q} = \left[\hat{\boldsymbol{\beta}}_1,..,\hat{\boldsymbol{\beta}}_q\right] = \left\{\hat{\beta}_{jk}\right\}$, whose columns, $\hat{\boldsymbol{\beta}}_k$, $k = 1,...,q$, are estimates of the effects of p markers in q different environments. The singular value decomposition of this matrix is $\hat{\mathbf{B}} = \mathbf{UDV}'$, where $\mathbf{U}_{p\times q} = \left[\boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_q\right] = \left\{\alpha_{jk}\right\}$ and $\mathbf{V}_{q\times q} = \left[\boldsymbol{\gamma}_1,...,\boldsymbol{\gamma}_q\right] = \left\{\gamma_{kl}\right\}$ are ortho-normal matrices that span the row (marker) and column (environment) spaces of $\hat{\mathbf{B}}$, respectively, and $\mathbf{D}_{q\times q}$ is a diagonal matrix whose non-null entries are the singular values of $\hat{\mathbf{B}}$, that is, $\mathbf{D} = Diag\left\{\lambda_k\right\}$.

The biplot is constructed using the first and second components, that is, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. Points in the biplot are the marker effects projected in the first two components, and are displayed using the coordinates provided by $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. The "environmental effects" are displayed as vectors whose coordinates are given by $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. The length of the vectors approximates the variance accounted for by the specific molecular marker and "environmental effect." Molecular markers represented in the same direction as the environments had positive effects on those environments, whereas molecular markers located in the opposite direction of the environmental vectors had negative effects on those environments. The cosine of the angle between two environments (or molecular marker effect) approximates the correlation of the two environments (or molecular marker), with an angle of zero indicating a correlation of +1, an angle of 90° (or -90°) a correlation of 0, and an angle of 180° a correlation of –1.

Appendix C

Marker effects for maize flowering data

The display of the first two component axes (accounting for 85.79% of the total variability in estimated SNPs effects) on estimated effects of the SNP markers in the six trait-environment combinations (MFL-SS, MFL-WW, FFL-SS, FFL-WW, ASI-SS, and ASI-WW) of the M-F dataset obtained from the M-BL model is depicted in Fig. 1C. The correlation between trait-environment combinations using marker effects and phenotypic data, and the effects of the SNP markers most distant from the center of the biplot are in Tables S3 and S4 (supplementary materials), respectively. Clearly the two groups of trait-environment combinations are dominated more by the trait (ASI vs FFL and MFL) and less by the environment (SS and WW). Phenotypic outcomes and estimates of marker effects for ASI showed relatively small correlations with those of FFL and MFL; this occurs because ASI is defined as the difference between FFL and MFL, and these two traits are positively correlated. The pattern of correlations between estimated SNP effects reflects the patterns of observed phenotypic correlations (Table S3 in supplementary materials).

Markers with relatively large (in absolute value) estimated effects are identified by name in Fig. 1C, and their effects are shown in Table S4 (supplementary materials). Interpretation of the marker effect on these traits should be different than their effect on grain yield, since the favorable marker allele decreases both flowering times, whereas for ASI, the optimal marker should give an ASI of 0. The alleles coded as 1 of SNPs whose estimated effects are located in the left and upper left corner of the biplot (i.e., PZA03551.1, PZA03578.1, PZA03222.1, PZA03385.1, PZB01201.1, and PZB00118.2) increase FFL, MFL, and ASI (they all have positive effects in all trait-environments combinations), whereas those SNPs located on the opposite side of the biplot (lower right corner) (i.e., PZA02587.16, PZA00236.7, PZB0255.1, and PZA00676.2) decrease the value of FFL, MFL, and ASI. Those SNPs whose presence is expected to increase or decrease traits across environments can be viewed as contributing to positive genetic correlations in FFL, MFL, and ASI between environments.

Despite the high heritability (between 0.74 and 0.87) found for flowering time and ASI in this maize trial, results show substantial interaction between molecular marker effects and environment. The biplot in Fig. 1C shows SNPs that had very contrasting effects across environments. For example, the minor alleles of SNPs whose estimated effects are located in the upper right corner of the biplot (PZA03592.3, PZB01077.3, and PZB02076.1) increase the anthesis-silking interval under drought and well-watered conditions (Table S4 in supplementary materials), but decrease days to male and female flowering. In contrast, the minor alleles of SNPs whose estimated effects are located in the opposite quadrant of the biplot (lower left corner) (PZB00592.1, PHM13183.12, and PZB01964.5) showed a complete rank reversal with respect to the effects of SNPs PZA03592.3, PZB01077.3, and PZB01077.3 on those trait-environment combinations, i.e., a decrease in ASI under SS and WW, and an increase in male and female flowering times. These results are suggestive of important molecular marker effect × environment interaction, which in turn causes genotype × environment interaction. On the other hand, BUCKLER *et al*. (2009) reported low levels of genotype × environment interaction for the same traits; however, our study covers a more diverse genetic background, and the selection history of the population considered here is different than the one used in BUCKLER *et al*. (2009).

The effects of all 1,148 SNPs (with their corresponding chromosome numbers) in each of the six trait-environment combinations can be found in Table S4a (supplementary

materials). The scores of the first two component axes of the 1,148 SNPs are in Table S7 (supplementary materials).

Marker effects for maize grain yield under stress and well-watered environments

Since only two trait-environment combinations (GY-WW and GY-SS) are available for the M-GY dataset, no principal component analysis was performed, and only the effect of the 10 SNPs with the largest positive effects and the 10 SNPs with the largest negative effects in SS and WW environments are presented in Table S5 (supplementary materials). The phenotypic correlations between GW-WW and GY-SS, as well as the correlations between the estimated marker effects for grain yield, were low (0.250). This indicates important context-dependent effects due to genotype × environment interaction. This was confirmed by the fact that none of the 10 SNPs with the largest/smallest effects in the GY-WW environment was among those with the largest/smallest effects under GY-SS conditions, and by the relatively low broad-base heritability of 0.510 and 0.381 under SS and WW environmental conditions, respectively.

The effects of 1,135 SNPs (with their corresponding chromosome numbers) in each of the two environments (SS and WW) can be found in Table S5a (supplementary materials).

Figure 1C. Biplot of the first and second principal component axes (Comp. 1 and Comp. 2) of maize female flowering (FFL) and male flowering (MFL) effects of the 1,148 SNPs estimated from the full data model M-BL of the maize dataset in each of two environments, severe water stress (SS) and well-watered (WW). A total of six trait-environment combinations (FFL-SS, FFL-WW, MFL-SS, MFL-WW, SS-ASI, and WW-ASI) were formed. Only the effects of the 19 SNPs that are located far from the center of the biplot were identified with their corresponding SNP's name (filled-in circles).