GENOMIC PREDICTION OF QUANTITATIVE TRAITS IN PLANT BREEDING USING MOLECULAR MARKERS AND PEDIGREE

Jose Crossa[1], Paulino Perez[1,2], Gustavo de los Campos[1,3], George Mahuku[1], Susanne Dreisigacker[1], and Cosmos Magorokosho[1]

[1] International Maize and Wheat Improvement Center, Apdo. Postal 6-641, 06600, México DF, México.
[2] Colegio de Postgraduados, Montecillos, México.
[3] Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, USA.

ABSTRACT

The availability of thousands of genome wide molecular markers has made possible the use of genomic selection in plants and animals. However, the evaluation of models for genomic selection in plant breeding populations is very limited. In this study, we provide an overview of several models for genomic selection, whose predictive ability we investigated using two plant data sets. One data set contains the historical phenotypic records of a series of wheat (*Triticum aestivum* L.) trials and recently generated genomic data. The other data set pertains to international maize (*Zea mays* L.) trials in which two disease traits (*Exserohilum turcicum* and *Cercospora zeae-maydis*) were measured in maize lines evaluated in five international environments. Results showed that models including marker information yield important gains in predictive ability, relative to that of a pedigree-based model—this, with a modest number of markers. Estimates of marker effects were different across environmental conditions, indicating that genotype × environment interaction is an important component of genetic variability. Overall, the study provides evidence from real populations indicating that genomic selection can be an effective tool for improving traits of economic importance in commercial crops.

INTRODUCTION

Selection in plant breeding is usually based on estimates of breeding values obtained with pedigree-based mixed models (Piepho et al., 2007; Piepho, 2009). In their multivariate formulation, these models can also accommodate genotype × environment (GE) interaction (Crossa et al., 2006; Burgueño et al., 2007). These models have been used successfully for predicting breeding values in plants and animals. However, pedigree-based models cannot account for mendelian segregation, a term that under an infinitesimal additive model (e.g. Fisher, 1918) and in absence of inbreeding, explains one half of the genetic variability. Molecular markers (MM) allow tracing mendelian segregation at several positions of the genome, which gives them enormous potential in terms of increasing the accuracy of estimates of genetic values and the genetic progress attainable when these predictions are used for selection purposes.

Marker-assisted selection (MAS) has been widely used in plant breeding to improve a few traits controlled by major genes. However, adoption of the technology has been limited because the bi-parental populations used for mapping quantitative trait loci (QTL) are not easily used in breeding applications. Also, MAS presents limitations for improving traits controlled by many loci with small effects. On the other hand, genomic selection (GS) (or genome-wide selection) is an approach for improving quantitative traits (Meuwissen et al., 2001) that uses all available MM across the genome to estimate genetic values.

Reports on the use of GS in plants are few and refer mainly to computer simulation studies such as the research of Bernardo and Yu (2007), who concluded that GS is superior to marker-assisted selection in maize. In a recent article, de los Campos et al. (2009a) used Bayesian estimates from genomic regression and showed that models using MM had better prediction accuracy of grain yield in wheat than those based on pedigree. Genomic selection has been validated in animal breeding for predicting breeding values (Gonzalez-Recio et al., 2008; Hayes et al., 2009; VanRaden et al., 2009; de los Campos et al., 2009a).

In a usual genetic model, the phenotypic response of the *i^th^* individual (y_i) is described as the sum of a genetic value, g_i , and a model residual, ε_i , such that the linear model for *n* genotypes $(i = 1,...,n)$ is represented as $y_i = g_i + \varepsilon_i$. One method for incorporating markers in models for GS is to define g_i as a parametric regression on marker covariates x_{ij} (which can take values of 1, 0 or -1 for a biallelic marker of a segregating population or values of 1 and 0 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*=1,2,…,*p*), where β_j is the regression of y_i on the j^{th} marker covariate. In matrix notation, the model is expressed as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Usually, the number of markers exceeds the number of individuals, and estimation of marker effects via ordinary least squares (OLS) is not feasible. In OLS, estimates are obtained to maximize model goodness of fit to the training set, and model complexity is not considered. When the number of MM is large, this typically yields high mean-squared error of estimates of marker effects and poor predictive ability.

Penalized estimation methods such as ridge regression (Hoerl and Kennard, 1970) or the absolute shrinkage and selection operator LASSO (Tibshirani, 1996) usually yield higher predictive ability. In ridge regression, estimates are obtained by minimizing the residual sum of squares, $\min_{\boldsymbol{\beta}}\left\{\sum (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right\}$, subject to the following constraint: $\sum_j \beta_j^2 \leq t$, or equivalently, $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left\{\sum (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda(t)\sum_j \beta_j^2\right\}$. The solution to this optimization problem can be shown to be $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda(t)\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$, where $\lambda = \lambda(t) > 0$ is a regularization parameter that induces shrinkage of estimates of effects towards zero. Even though these estimates are biased, the sampling variance is reduced yielding smaller mean-squared error and better predictive ability (e.g., Hastie et al., 2009).

From the Bayesian perspective, estimates of marker effects from ridge regression can be viewed as the posterior mean of a model with a Gaussian likelihood and a prior distribution of marker effects that is the product of *p* normal densities, that is, $p\left(\mathbf{y},\boldsymbol{\beta}|\sigma_\varepsilon^2,\sigma_\beta^2\right) = \prod_i N\left(y_i | \sum_j x_{ij}\beta_j, \sigma_\varepsilon^2\right)\prod_j N\left(\beta_j|0,\sigma_\beta^2\right)$, where σ_ε^2 and σ_β^2 are the residual variance and the prior precision variance of marker effects, respectively. When variance components are known, estimated marker effects are best linear unbiased predictors or BLUP (e.g., Schaeffer, 2006). A disadvantage of ridge regression, or its Bayesian counterpart, is that the extent of shrinkage is homogeneous across markers, which may not be appropriate in GS if some markers are located in regions that are not associated with genetic variance while others may be linked to QTLs (Goddard and Hayes, 2007).

An alternative to ridge regression is to use LASSO (Tibshirani, 1996). Estimates in LASSO are obtained by minimizing the residual sum of squares, $\min_{\boldsymbol{\beta}}\left\{\sum\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\}$, subject to the following constraint: $\sum_j\left|\beta_j\right|\leq t$, or equivalently, $\tilde{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}\left\{\sum(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2+\lambda(t)\sum_j\left|\beta_j\right|\right\}$. Unlike the quadratic penalty of ridge regression, $\sum_j\beta_j^2$, the absolute-value penalty of LASSO, $\sum_j\left|\beta_j\right|$, induces selection and shrinkage simultaneously. From a Bayesian perspective, LASSO estimates can be viewed as the posterior mode of a Bayesian model with a Gaussian likelihood and with a prior on marker effects that is the product of p IID double-exponential (DE) distributions, that is $p\left(\mathbf{y},\boldsymbol{\beta}\middle|\sigma_\varepsilon^2,\sigma_\beta^2\right)=\prod_i N\left(y_i\middle|\sum_j x_{ij}\beta_j,\sigma_\varepsilon^2\right)\prod_j DE\left(\beta_j\middle|0,\lambda\right)$. Relative to the Gaussian prior of the Bayesian ridge regression, the DE has higher peaks of mass at zero and thicker tails. Here, $\lambda(t)$ controls the shape of DE density, with higher values increasing the peak of mass at zero. A computationally convenient representation of this model (hereinafter called BL, for Bayesian LASSO) was presented by Park and Casella (2008) and used in GS by de los Campos et al. (2009a). The latter authors discussed the connections between the BL and other models for GS such as those proposed by Meuwissen et al. (2001). Contrary to ridge regression, the BL and the models proposed by Meuwissen et al. (2001) induce marker-specific shrinkage.

One variant of the traditional LASSO is the elastic net LASSO (Zou and Hastie, 2005), which differs from LASSO in that it uses two penalties. The advantage of the elastic net method is that, by adding another penalty, it stabilizes the LASSO solution when some predictors are highly correlated; such is the case of MM used in GS. Two other variants of LASSO are the group LASSO and the fused LASSO. Group LASSO (Yuan and Lin, 2006) selects variables at a group level such that some groups of predictors are selected together; this may be useful when instead of examining the effect of individual MM, the researcher wishes to examine the effects of haplotypes (genes) comprising several MM in high linkage disequilibrium. Group LASSO could also be useful when there are more than two alleles for each MM, and the breeder wishes to keep all the alleles of the same MM active in the model. The fused LASSO focuses on adjacent predictors such that the value effects tend to be the same for adjacent predictors. This can be useful when there is a natural ordering of predictors, for example, when markers have been ordered on a common map.

An alternative to parametric regressions is to use semi-parametric methods such as reproducing kernel Hilbert spaces (RKHS) regression (Gianola et al., 2006; Gianola and van Kamm, 2008). A Bayesian RKHS regression for molecular markers 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, $\text{Cov}(g_i, g_j) \propto \text{K}(\mathbf{x}_i, \mathbf{x}_j)$, 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 advantages of RKHS regression is that it 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 in RKHS, the model is represented in terms of *n* unknowns, which gives RKHS a great computational advantage relative to parametric methods, when $p>>n$.

In this chapter, we evaluate several models for predicting genetic values that differ in the type of genetic information used (MM, pedigree, or both) and in the way that MM data are incorporated into the model (either parametric regression via Bayesian LASSO or semi-parametric regression via RKHS). Models are compared based on predictive ability estimated using cross-validation methods and evaluated in two data sets form CIMMYT's Global Wheat and Maize Programs. In the remaining sections, the data and models are described, the results presented, and a discussion of the use of MM in GS provided.

EXPERIMENTAL DATA

Wheat experimental data set

Two CIMMYT wheat international multi-environment trials were studied: Elite Spring Wheat Yield Trials (ESWYT) 20 and 24. ESWYT20 had 47 lines, some of them form 7 sets of sister lines, each with 2-4 sisters, whereas ESWYT24 comprised 46 lines with some of them forming 8 sets of sister lines, each with 2-4 sisters. A total of 93 lines were analyzed. In general, the data for each ESWYT are balanced, although sometimes only one replicate could be recorded. There are no lines in common among the historical sequence of ESWYTs. A total of ten international sites (E1-E10) common to both ESWYT20 and ESWYT24 were included in this study; the trait analyzed was average (across replicates) grain yield, standardized to have a sample variance equal to one in each environment.

Diversity Array Technology (DArT) markers were generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au), which is a whole-genome profiling service laboratory. In total, 234 loci were scored as present (1) or absent (0) across the 93 wheat lines. The coefficient of parentage (COP) between individuals, i.e., the probability of being identical by descent, was derived from the pedigree using the Browse application (http://cropwiki.irri.org/icis/index.php/TDM_GMS_Browse) of the International Crop Information System (ICIS) described in McLaren et al. (2005) and accounts for selection as well as inbreeding.

Maize experimental data set

The maize data set is from the Drought Tolerance Maize (DTMA) 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 southern, eastern and western African countries in coordination with the tropical maize breeding program in Mexico. The data used in this study come from a large study aimed at detecting chromosomal regions affecting drought tolerance and adaptive traits identified in global maize germplasm based on analysis of available marker data from genotyping 300 tropical inbred lines with 1152 single nucleotide polymorphisms (SNPs). No pedigree was available for this data set.

Two traits were analyzed in this study: (1) northern corn leaf blight (NCLB) disease caused by the fungus *Exserohilum turcicum* and evaluated in three international environments: El Batán (Mexico), Harare (Zimbabwe), and Mpongue (Zambia); and (2) gray leaf spot (GLS) caused by the fungus *Cercospora zeae-maydis* and evaluated in San Pedro Lagunillas (Mexico) and Pereira (Colombia). NCLB is a major foliar disease of maize that occurs worldwide, virtually everywhere maize is grown. The disease is polycyclic in nature and can cause extensive defoliation during the grain-filling period, resulting in grain yield losses of 50% or more (Welz and Geiger, 2000). Pandemic in Africa, GLS is now recognized as one of the most significant yield-limiting diseases of maize worldwide and is associated with the rapid adoption of conservation agriculture techniques (Ward et al., 1999). For the El Batán site, the NCLB trait analyzed was the area under the disease progress curve, which is a measure of quantitative disease resistance that integrates all aspects of disease progress in relation to host development and growth. It therefore characterizes the overall patterns of disease increase over time or over time and space. For the Harare, Mpongue, Pereira, and San Pedro Lagunillas sites, the NCLB and GLS traits were analyzed using an ordinal scale from 1 (no disease) to 5 (complete infection). The Box-Cox (Box and Cox, 1964) transformation was applied to the original data to make their distribution more symmetric; after transformation, data were standardized to get a unit sample variance within each environment.

STATISTICAL MODELS

In this study, the response consisted of the average performance of each line within each environment, that is, $\bar{y}_i = \frac{1}{SD \times n_i}\sum_{k=1}^{n_i} y_{ik}$, where n_i is the number of replicates available for the i^{th} line and SD is the (sample) standard deviation of $\{\bar{y}_i\}$. With this, and extending the model to allow for an intercept, the data-equation becomes:

$$\bar{y}_i = \mu + g_i + \varepsilon_i . \qquad [1]$$

We adopted Gaussian assumptions for model residuals, and allowed for heterogeneous residual variances to accommodate unbalanced data. Specifically, the joint distribution of model residuals in [1] is $p(\boldsymbol{\varepsilon}) = \prod_{i=1}^{n} N\left(\varepsilon_i \middle| 0, \frac{\sigma_e^2}{n_i}\right)$, where n_i is the number of replicates used for computing the mean value of the i^{th} genotype in the corresponding environment. Together with [1], this gives the following likelihood function: $p(\mathbf{y}|\mu, \mathbf{g}, \sigma_\varepsilon^2) = \prod_{i=1}^{n} N\left(y_i \middle| \mu + g_i, \frac{\sigma_\varepsilon^2}{n_i}\right)$.

Models will differ in the type of information and the methods used to describe g_i. In the following sections, we present the different classes of models used to incorporate pedigree and marker data using either parametric or semi-parametric methods.

Standard infinitesimal model

A standard additive infinitesimal model (e.g., Fisher, 1918, Henderson, 1975) postulates that genetic values are multivariate normal, centered at zero, and with a co-variance matrix proportional to the numerator relationship matrix computed from the pedigree, that is, $g_i = a_i$, where $\mathbf{a} = (a_1, ..., a_n) \sim N(\mathbf{a}|\mathbf{0}, \mathbf{A}\sigma_a^2)$, where σ_a^2 is the additive variance. In a Bayesian setting, a prior is assigned to variance parameters as well. Following standard assumptions, we chose independent scaled inverse Chi-square distributions for the residual and the additive variance. Collecting assumptions, the joint prior of the pedigree-based model (P) becomes:

$$p\left(\mu,\mathbf{a},\sigma_\varepsilon^2,\sigma_a^2\middle|df_\varepsilon,S_\varepsilon,df_a,S_a\right)\propto N\left(\mathbf{a}\middle|\mathbf{0},\mathbf{A}\sigma_a^2\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_a^2\middle|df_a,S_a\right), \qquad [2a]$$

where $df_.$ and $S_.$ are prior degree-of-freedom and scale parameters, respectively. The joint posterior distribution for this model is obtained combining the likelihood and prior such that:

$$\begin{aligned}p\left(\mu,\mathbf{a},\sigma_\varepsilon^2,\sigma_a^2\middle|\bar{\mathbf{y}},H\right)\propto&\prod_{i=1}^{n}N\left(\bar{y}_i\middle|\mu+a_i,\frac{\sigma_\varepsilon^2}{n_i}\right)\\&\times N\left(\mathbf{a}\middle|\mathbf{0},\mathbf{A}\sigma_a^2\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_a^2\middle|df_a,S_a\right),\end{aligned} \qquad [2b]$$

where H denotes all hyper parameters. This joint posterior distribution does not have closed form, but a Gibbs sampler (e.g., Sorensen and Gianola, 2002) can be used to draw samples from it. This model was fitted using $S_a=S_\varepsilon=1$ and $df_\varepsilon=df_a=4$ for the scale and degree-of-freedom parameters, respectively.

Pedigree was not available for the maize data set. An alternative is to replace **A** in [2a] and [2b] with a kinship matrix (**U**) estimated using marker genotypes. A kinship-based infinitesimal model (K) is obtained using $g_i=u_i$, where $\mathbf{u}=(u_1,...,u_n)'\sim N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{U}\sigma_u^2\right)$ and σ_u^2 is the associated variance parameter. The joint prior and posterior distributions of this model are:

$$p\left(\mu,\mathbf{u},\sigma_\varepsilon^2,\sigma_u^2\middle|df_\varepsilon,S_\varepsilon,df_u,S_u\right)\propto N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{U}\sigma_u^2\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_u^2\middle|df_u,S_u\right) \qquad [3a]$$

and

$$\begin{aligned}p\left(\mu,\mathbf{u},\sigma_\varepsilon^2,\sigma_u^2\middle|\bar{\mathbf{y}},H\right)\propto&\prod_{i=1}^{n}N\left(\bar{y}_i\middle|\mu+u_i,\frac{\sigma_\varepsilon^2}{n_i}\right)\\&\times N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{U}\sigma_u^2\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_u^2\middle|df_u,S_u\right),\end{aligned} \qquad [3b]$$

respectively. This model was fitted using $S_u=S_\varepsilon=1$ and $df_\varepsilon=df_u=4$ as values for the scale and degree-of-freedom parameters, respectively. A main difference between the pedigree-based model and the model defined by [3a] is that in the latter, the matrix of relationships, **U**, is computed using marker genotypes; it therefore accounts for mendelian segregation.

Parametric GS model using the Bayesian LASSO (BL)

As stated in this model, marker data are introduced using a parametric regression on marker genotypes. The standard Bayesian LASSO (BL) of Park and Casella (2008) can be extended to accommodate pedigree data as well, as described in de los Campos et al. (2009a). For this model, we can have two cases depending whether the matrix **A** or the Kinship matrix (**U**) is available, as described below.

PM–BL model

When a pedigree is available, genetic values can be described as the sum of a regression on marker covariates and a regression on pedigree (e.g., de los Campos et al., 2009a). The data equation becomes $\bar{y}_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + a_i + \varepsilon_i$. The joint prior density of the unknowns of the linear model (upon assigning a flat prior to μ) is:

$$p\left(\mu, \mathbf{a}, \boldsymbol{\beta}, \lambda, \sigma_\varepsilon^2, \sigma_a^2 \middle| r, \delta, df_\varepsilon, S_\varepsilon, df_a, S_a\right) \propto N\left(\mathbf{a}\middle|\mathbf{0}, \mathbf{A}\sigma_a^2\right)\left\{\prod_{j=1}^{p} N\left(\beta_j \middle| 0, \sigma_\varepsilon^2, \tau_j^2\right)\right\}$$
$$\times\left\{\prod_{j=1}^{p} Exp\left(\tau_j^2\middle|\lambda^2\right)\right\} G\left(\lambda^2\middle|r, \delta\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon, S_\varepsilon\right) \qquad [4a]$$
$$\times\chi^{-2}\left(\sigma_a^2\middle|df_a, S_a\right)$$

In [4a], 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 $r = 0.6$. The prior for the vector **a** is as in the standard infinitesimal model (see [3a]). Finally, independent scaled-inverse chi-squared priors are assigned to the variance parameters, and the scale and degree-of-freedom parameters were set to $S_a = S_\varepsilon = 1$ and $df_\varepsilon = df_a = 4$, respectively. The above model is referred to as PM-BL (for pedigree and marker models using the BL). Combining the Gaussian likelihood with the prior assumptions described in [4a], the posterior distribution of the PM-BL model is:

$$p\left(\mu, \mathbf{a}, \boldsymbol{\beta}, \lambda, \sigma_\varepsilon^2, \sigma_a^2 \middle| \bar{\mathbf{y}}, H\right) \propto \left\{\prod_{i=1}^{n} N\left(\bar{y}_i\middle|\mu + \sum_{j=1}^{p} x_{ij}\beta_j + a_i, \frac{\sigma_\varepsilon^2}{n_i}\right)\right\} N\left(\mathbf{a}\middle|\mathbf{0}, \mathbf{A}\sigma_a^2\right)\left\{\prod_{j=1}^{p} N\left(\beta_j\middle|0, \sigma_\varepsilon^2, \tau_j^2\right)\right\}$$
$$\times\left\{\prod_{j=1}^{p} Exp\left(\tau_j^2\middle|\lambda^2\right)\right\} G\left(\lambda^2\middle|r, \delta\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon, S_\varepsilon\right)$$
$$\times\chi^{-2}\left(\sigma_a^2\middle|df_a, S_a\right).$$

[4b]

A set of fully conditional distributions that can be used to implement a Gibbs sampler for the above model is given in de los Campos et al. (2009a).

KM–BL model

When a pedigree is not available, the relationship matrix **A** can be replaced by the kinship matrix (**U**) computed from marker genotypes. The data equation of this model (KM-BL) is $\bar{y}_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + u_i + \varepsilon_i$, where $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_p\right)'$ and $\mathbf{u} = \left(u_1, ..., u_n\right)'$ are as described above. The joint prior and posterior distributions of KM-BL are:

$$p\left(\mu, \mathbf{u}, \boldsymbol{\beta}, \sigma_\varepsilon^2, \lambda, \sigma_u^2 \middle| r, \delta, df_\varepsilon, S_\varepsilon, df_u, S_u\right) \propto N\left(\mathbf{u}\middle|\mathbf{0}, \mathbf{U}\sigma_u^2\right)\left\{\prod_{j=1}^{p} N\left(\beta_j\middle|0, \sigma_\varepsilon^2, \tau_j^2\right)\right\}\left\{\prod_{j=1}^{p} Exp\left(\tau_j^2\middle|\lambda^2\right)\right\} \qquad [5a]$$
$$\times G\left(\lambda^2\middle|r, \delta\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon, S_\varepsilon\right)\chi^{-2}\left(\sigma_u^2\middle|df_u, S_u\right),$$

and

$$p\left(\mu, \mathbf{u}, \boldsymbol{\beta}, \sigma_\varepsilon^2, \lambda, \sigma_u^2 \middle| \bar{\mathbf{y}}, H\right) \propto \left\{ \prod_{i=1}^{n} N\left(\bar{y}_i \middle| \mu + \sum_{j=1}^{p} x_{ij}\beta_j + u_i, \frac{\sigma_\varepsilon^2}{n_i} \right) \right\} N\left(\mathbf{u}\middle|\mathbf{0}, \mathbf{U}\sigma_u^2\right)$$
$$\times \left\{ \prod_{j=1}^{p} N\left(\beta_j \middle| 0, \sigma_\varepsilon^2, \tau_j^2\right) \right\} \left\{ \prod_{j=1}^{p} Exp\left(\tau_j^2 \middle| \lambda^2\right) \right\} \qquad [5b]$$
$$\times G\left(\lambda^2 \middle| r, \delta\right) \chi^{-2}\left(\sigma_\varepsilon^2 \middle| df_\varepsilon, S_\varepsilon\right) \chi^{-2}\left(\sigma_u^2 \middle| df_u, S_u\right),$$

The above model is equivalent to that described by [4b] with **U** replacing **A**; therefore, the algorithm described by de los Campos et al. (2009a) can also be used to draw samples from the posterior distribution defined by [5b]. This model was fitted using the following hyper-parameters values: $\delta = 1\times10^{-4}$, $r = 0.6$. $S_u = S_\varepsilon = 1$, and $df_\varepsilon = df_u = 4$.

M–BL model

A special case of PM-BL (see [4b]) is obtained by removing a_i from the data equation, such that $\bar{y}_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$; this model is denoted as model M-BL, which stands for marker-based model using the Bayesian LASSO. The prior and posterior distributions of this model are:

$$p\left(\mu, \boldsymbol{\beta}, \sigma_\varepsilon^2, \lambda \middle| r, \delta, df_\varepsilon, S_\varepsilon\right) \propto \left\{ \prod_{j=1}^{p} N\left(\beta_j \middle| 0, \sigma_\varepsilon^2, \tau_j^2\right) \right\} \left\{ \prod_{j=1}^{p} Exp\left(\tau_j^2 \middle| \lambda^2\right) \right\} \chi^{-2}\left(\sigma_\varepsilon^2 \middle| df_\varepsilon, S_\varepsilon\right) G\left(\lambda^2 \middle| r, \delta\right), \qquad [6a]$$

and

$$p\left(\mu, \boldsymbol{\beta}, \sigma_\varepsilon^2, \lambda \middle| \bar{\mathbf{y}}, H\right) \propto \left\{ \prod_{i=1}^{n} N\left(\bar{y}_i \middle| \mu + \sum_{j=1}^{p} x_{ij}\beta_j + u_i, \frac{\sigma_\varepsilon^2}{n_i} \right) \right\} \left\{ \prod_{j=1}^{p} N\left(\beta_j \middle| 0, \sigma_\varepsilon^2, \tau_j^2\right) \right\} \left\{ \prod_{j=1}^{p} Exp\left(\tau_j^2 \middle| \lambda^2\right) \right\} \qquad [6b]$$
$$\times G\left(\lambda^2 \middle| r, \delta\right) \chi^{-2}\left(\sigma_\varepsilon^2 \middle| df_\varepsilon, S_\varepsilon\right),$$

Semi-parametric GS model using Bayesian reproducing kernel Hilbert spaces (RKHS) regression

As stated earlier, an alternative to the parametric regression is to introduce marker information using RKHS, an alternative that is discussed next.

PM–RKHS model

When markers and pedigree are available, genetic values can be modeled as the sum of two components, $g_i = a_i + f_i$, where a_i is a regression on the pedigree, as described earlier, and f_i is a semi-parametric regression on marker genotypes. In RKHS the assumption is that $\mathbf{f} = \left(f_1, ..., f_n\right)'$, a Gaussian process with null-mean and a (co)variance function proportional to a reproducing kernel, $\mathbf{K}_{\text{RKHS}}\left(\mathbf{x}_i, \mathbf{x}_j\right)$, 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 $\mathbf{a}$, **f**, and the associated variance parameters σ_ε^2, σ_a^2, and σ_f^2 is as follows:

$$\begin{aligned} p\left(\mu,\mathbf{a},\mathbf{f},\sigma_\varepsilon^2,\sigma_a^2,\sigma_f^2 \middle| df_\varepsilon,S_\varepsilon,df_a,S_a,df_f,S_f\right) &\propto N\left(\mathbf{a}\middle|\mathbf{0},\mathbf{A}\sigma_a^2\right)N\left(\mathbf{f}\middle|\mathbf{0},\mathbf{K}_{\mathrm{RKHS}}\sigma_f^2\right) \\ &\times\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_a^2\middle|df_a,S_a\right)\chi^{-2}\left(\sigma_f^2\middle|df_f,S_f\right) \end{aligned} \qquad [7a]$$

In general, any positive definite function, i.e., any function satisfying $\sum_i\sum_j \alpha_i\alpha_j\mathbf{K}_{\mathrm{RHKS}}\left(\mathbf{x}_i,\mathbf{x}_j\right)$ for the all non-null sequence $\{\alpha_i\}$, is a valid choice of kernel. In this study, we chose $\mathbf{K}_{\mathrm{RHKS}}\left(\mathbf{x}_i,\mathbf{x}_j\right)$ to be the Gaussian kernel $\mathbf{K}_{\mathrm{RHKS}}\left(\mathbf{x}_i,\mathbf{x}_j\right)=\exp\left\{-2\frac{d_{ij}}{q_{.5}}\right\}$, where $d_{ij}=\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2$ is a squared-Euclidean distance, and $q_{0.5}$ is the sample median of the matrix of sampled squared-Euclidean distances $\{d_{ij}\}$. Note that, with our choice of kernel, if $d_{ij}=q_{.5}$ $\mathrm{K}_{\mathrm{RHKS}}\left(\mathbf{x}_i,\mathbf{x}_j\right)=\exp(-2)\approx 0.13$, 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 (or farther apart) than $q_{.5}$. The scale and degree-of-freedom parameters of the prior described in [7a] were set to $S_\varepsilon=S_a=S_f=1$ and $df_\varepsilon=df_a=df_f=4$, respectively

Combining the prior assumptions in [7a] with those of the likelihood, the fully conditional distribution of this model becomes:

$$\begin{aligned} p\left(\mu,\mathbf{a},\mathbf{f},\sigma_\varepsilon^2,\sigma_a^2,\sigma_f^2 \middle| \bar{\mathbf{y}},H\right) &\propto \left\{\prod_{i=1}^{n} N\left(\bar{y}_i\middle|\mu+a_i+f_i,\frac{\sigma_\varepsilon^2}{n_i}\right)\right\}N\left(\mathbf{a}\middle|\mathbf{0},\mathbf{A}\sigma_a^2\right)N\left(\mathbf{f}\middle|\mathbf{0},\mathbf{K}_{\mathrm{RKHS}}\sigma_f^2\right) \\ &\times\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_a^2\middle|df_a,S_a\right)\chi^{-2}\left(\sigma_f^2\middle|df_f,S_f\right) \end{aligned} \qquad [7b]$$

A slight extension of the algorithm used to draw samples from the pedigree model (e.g., Sorensen and Gianola, 2002) can be used to draw samples from the above distribution (see Appendix).

KM–RKHS model

As before, when a pedigree is not available, one can replace **a** and **A** in [7a] with **u** and **U**, respectively, and the prior and posterior distribution become:

$$\begin{aligned} p\left(\mu,\mathbf{u},\mathbf{f},\sigma_\varepsilon^2,\sigma_u^2,\sigma_f^2 \middle| df_\varepsilon,S_\varepsilon,df_u,S_u,df_f,S_f\right) &\propto N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{U}\sigma_u^2\right)N\left(\mathbf{f}\middle|\mathbf{0},\mathbf{K}_{\mathrm{RKHS}}\sigma_f^2\right) \\ &\times\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_u^2\middle|df_u,S_u\right)\chi^{-2}\left(\sigma_f^2\middle|df_f,S_f\right) \end{aligned} \qquad [8a]$$

and

$$\begin{aligned} p\left(\mu,\mathbf{u},\mathbf{f},\sigma_\varepsilon^2,\sigma_u^2,\sigma_f^2 \middle| \bar{\mathbf{y}},H\right) &\propto \left\{\prod_{i=1}^{n} N\left(\bar{y}_i\middle|\mu+u_i+f_i,\frac{\sigma_\varepsilon^2}{n_i}\right)\right\}N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{U}\sigma_u^2\right)N\left(\mathbf{f}\middle|\mathbf{0},\mathbf{K}_{\mathrm{RKHS}}\sigma_f^2\right) \\ &\times\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_u^2\middle|df_u,S_u\right)\chi^{-2}\left(\sigma_f^2\middle|df_f,S_f\right) \end{aligned} \qquad [8b]$$

The same algorithm that is used to draw samples from [7b] can be used to obtain samples from the posterior distribution defined in [8b] (see Appendix).

M–RKHS model

A marker-based RKHS (M-RKHS) model can be obtained by removing a_i from the data equation in [7a] and [7b]. The prior and posterior distribution of this model are given by

$$p\left(\mu, \mathbf{f}, \sigma_\varepsilon^2, \sigma_a^2, \sigma_f^2 \middle| df_\varepsilon, S_\varepsilon, df_a, S_a, df_f, S_f\right) \propto N\left(\mathbf{f} \middle| \mathbf{0}, \mathbf{K}_{\text{RKHS}} \sigma_f^2\right) \chi^{-2}\left(\sigma_\varepsilon^2 \middle| df_\varepsilon, S_\varepsilon\right) \times \chi^{-2}\left(\sigma_a^2 \middle| df_a, S_a\right) \chi^{-2}\left(\sigma_f^2 \middle| df_f, S_f\right) \qquad [9a]$$

and

$$p\left(\mu, \mathbf{f}, \sigma_\varepsilon^2, \sigma_a^2, \sigma_f^2 \middle| \bar{\mathbf{y}}, H\right) \propto \left\{ \prod_{i=1}^{n} N\left(\bar{y}_i \middle| \mu + f_i, \frac{\sigma_\varepsilon^2}{n_i} \right) \right\} N\left(\mathbf{f} \middle| \mathbf{0}, \mathbf{K}_{\text{RKHS}} \sigma_f^2\right) \times \chi^{-2}\left(\sigma_\varepsilon^2 \middle| df_\varepsilon, S_\varepsilon\right) \chi^{-2}\left(\sigma_a^2 \middle| df_a, S_a\right) \chi^{-2}\left(\sigma_f^2 \middle| df_f, S_f\right), \qquad [9b]$$

respectively.

Note that the kinship-based model, whose posterior distribution is given by [3b], has exactly the same structure as that of the above model but with $\mathbf{K}_{\text{RKHS}}$ replaced with $\mathbf{U}$, the Gaussian kernel and the kinship matrix are simply two different (co)variance functions.

Inferences for all the above models were based on 30,000 samples (obtained after discarding 5,000 samples as burn-in). Convergence was checked by inspecting trace plots of variance parameters.

CROSS-VALIDATION

Cross-validation (CV) is used to assess how the results of a statistical model will generalize to another data set—for example, how a fitted model will predict data that were not used to fit the model. Predicting the performance of genotypes with phenotypes yet to be observed (e.g., newly developed lines or lines that have been evaluated in few environments) is central in plant breeding; thus, cross-validation appears as a natural way of assessing model performance from the breeder's perspective. A simple approach for evaluating predictive ability consists of dividing the data into a *training* and a *validation sample*, sometimes also referred to as a *testing set*. Models are fitted using the *training sample*, and the fitted models are used to predict outcomes in the *validation sample*. This approach is appropriate for large data sets but is not recommended for small data sets, because the size of the training and validation samples becomes too small (e.g., Hastie et al., 2009).

A *k*-fold cross-validation is a generalization of the training/testing evaluation described above. Here, the data set is divided into *k* groups; this is done by assigning observations $\{i=1,..,n\}$ into *k* disjoint sets $\{S_1,...,S_k\}$. Each of these sets can then be used to measure predictive ability. For example, using the *1st* set, the data can be divided so that the *training set* contains all the observations in $\{S_2,...,S_k\}$and the *testing set* those in S_1. Subsequently, models are fitted using data into$\{S_2,...,S_k\}$ and the fitted model is used to obtain predictions for observations in $\{S_1\}$, that is, $\{\hat{y}_i : i \in S_1\}$. Repeating this exercise for the 2nd, 3rd,.., *kth* 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.

RESPONSE PATTERNS OF MARKER EFFECTS ON ENVIRONMENTS FOR THE WHEAT AND MAIZE DATA

Principal component analysis was performed using estimates of marker effects obtained from the PM-BL fitted to the entire wheat and maize data sets. Results are displayed in a biplot of the first two principal components axes. A large number of articles on the use of biplot analysis for studying genotype × environment interaction has been published (Cornelius et al., 2001; Crossa et al., 2001, 2004). For this study, consider a set of estimated molecular marker effects measured in various environments arranged in a two-way table $\mathbf{B} = \{\beta_{jh}\}$. Such a two-way table may be analyzed through singular value decomposition (or principal component) analysis as

$$\beta_{jh} = \sum_{m=1}^{t} \xi_m \alpha_{jm} \gamma_{hm} \qquad [10]$$

where β_{jh} is the estimated effect of the j^{th} molecular marker (j=1,2,…,p) in the h^{th} environment (h=1,2,…,s); ξ_m's ($\xi_1 \geq \xi_2 \geq ... \geq \xi_t$) are scaling constants (singular values) that allow the imposition of orthonormality constraints on the singular vectors for molecular markers, $\boldsymbol{\alpha}_m = (\alpha_{1m},...,\alpha_{pm})'$, and for environments, $\boldsymbol{\gamma}_m = (\gamma_{1m},...,\gamma_{sm})'$, such that $\sum_{j=1}^{p} \alpha_{jm}^2 = \sum_{h=1}^{s} \gamma_{hm}^2 = 1$ and $\sum_{j=1}^{p} \alpha_{jm}\alpha_{jm'} = \sum_{h=1}^{s} \gamma_{hm}\gamma_{hm'} = 0$ for $m \neq m'$.

The biplot graphs vectors α_{j1} and γ_{h1} versus vectors α_{j2} and γ_{h2}, where molecular markers and environments are represented as vectors in a two-dimensional space (Gabriel, 1971, 1978). The length of the vectors approximates the variance accounted for by the specific molecular marker and environment. Molecular markers represented in the same direction as the environments had positive effects on those environments, whereas molecular markers located in the opposite direction to the environments have negative effects on those environments (note that the relevance of a marker is given by the absolute value of marker effect; the sign of the estimated effect only indicates which allele should be favored in selection. For example, a positive effect means that substituting the allele coded with 0 by that coded with 1 is expected to increase the trait of interest, something desirable for grain yield and undesirable for disease traits). 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.

RESULTS AND DISCUSSION

Estimates of Variance Components

The estimated posterior means and the posterior standard deviation of the variance parameters ($\sigma_\varepsilon^2, \sigma_a^2, \sigma_u^2, \sigma_f^2$), and of λ obtained when models were fitted using all available records (full-data model) are presented in Tables 1, 2, and 3 for the wheat data set and the two traits of the maize data set, respectively. The posterior mean of the residual variance can be used to assess model goodness of fit. Here, since data were standardized to have a sample variance equal to one, the estimate of the residual variance gives an indication of the proportion of phenotypic variance that is attributable to model residuals, and 1-σ_ε^2 gives the proportion of phenotypic variance attributable to differences between genotypes. In the wheat data set (Table

1), the pedigree model always gave the worst fit (i.e., larger posterior mean of residual variance), and the RKHS models fitted the data better than molecular-based models using the BL. Also, in this data set, models including markers and pedigree (PM) almost always had a better fit than those based on molecular markers or pedigree only.

The differences in goodness of fit between models fitted to the maize data sets were not as marked as those in the wheat data set. This occurs because, unlike in the case of the wheat data set, in maize all models are marker-based models; the comparison involves different ways of incorporating markers, rather than different information sets such as pedigree vs markers. In the analysis of NCLB disease, the kinship model (K) fitted the data slightly better than the M-BL and KM-BL models in Harare and Mpongwe sites, whereas for GLS disease, models M-RHKS and KM-RHKS gave the best fit to the data in both sites, Pereira and San Pedro Lagunillas.

For the wheat data set, estimates of variance components, σ_a^2 and σ_f^2, obtained from PM-RHKS can be used to assess the relative contribution to genetic values of the regression on markers (f_i) and that of the regression on the pedigree (a_i). From the model described by [7a], $Var(a_i) = a(i,i)\sigma_a^2$, where $a(i,i) = 1 + F_i$ is the i^{th} diagonal element of matrix $\mathbf{A}$ and F_i is the coefficient of inbreeding for the i^{th} line, and $Var(f_i) = \mathrm{K}_{\mathrm{RKHS}}(\mathbf{x}_i, \mathbf{x}_i)\sigma_f^2 = \sigma_f^2$ because $\mathrm{K}_{\mathrm{RKHS}}(\mathbf{x}_i, \mathbf{x}_i)$ is a standardized kernel, with $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_i) = 1$. In the wheat data set, the average value of $1 + F_i$ was close to 2 (1.98); thus, comparison of the relative contribution of each component to the regression can be based on the ratio $\frac{\sigma_f^2}{\bar{a}(i,i)\sigma_a^2} \approx \frac{\sigma_f^2}{2\sigma_a^2}$. In the wheat data this ratio was larger than 1.6 in all environments except in E6, where it was 0.6753. In E1 and E8, this ratio was 3.18 and 2.44, respectively (Table 1). These results indicated that in the PM-RHKS model for the wheat data set, markers contributed more to estimates of genetic values than did the pedigree. In agreement with this, we also observed that the posterior mean of σ_a^2 for the P model is always larger than the posterior mean of σ_a^2 for models PM-BL and PM-RHKS, indicating that inclusion of markers, using either BL or RKHS, reduced the contribution of a_i to the conditional expectation.

Table 1. Posterior estimates of the mean and standard deviation (SD) of parameters $\sigma_\varepsilon^2, \sigma_a^2, \sigma_f^2$, and λ from full-data analysis of grain yield of 93 wheat lines genotyped with 234 DArT molecular markers. Five models* were fitted to each of 10 international environments (E1-E10).

Environment	Model*	σ_ε^2		σ_a^2		σ_f^2		λ	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
E1	P	0.814	0.156	0.198	0.094	---	---	---	---
	M-BL	0.419	0.108	---	---	---	---	8.008	2.557
	PM-BL	0.406	0.104	0.018	0.005	---	---	7.749	2.439
	M-RKHS	0.226	0.085	---	---	0.853	0.212	---	---
	PM-RKHS	0.193	0.083	0.126	0.052	0.802	0.208	---	---
	P	0.528	0.124	0.330	0.120	---	---	---	---

	M-BL	0.407	0.097	---	---	---	---	8.170	2.386
E2	PM-BL	0.377	0.095	0.017	0.005	---	---	7.560	2.541
	M-RKHS	0.265	0.075	---	---	0.680	0.158	---	---
	PM-RKHS	0.183	0.068	0.156	0.062	0.571	0.159	---	---
	P	0.483	0.120	0.366	0.129	---	---	---	---
	M-BL	0.428	0.104	---	---	---	---	8.417	2.610
E3	PM-BL	0.405	0.099	0.018	0.005	---	---	8.056	2.551
	M-RKHS	0.248	0.072	---	---	0.704	0.165	---	---
	PM-RKHS	0.159	0.062	0.176	0.068	0.565	0.159	---	---
	P	0.913	0.148	0.144	0.060	---	---	---	---
	M-BL	0.506	0.113	---	---	---	---	10.418	3.589
E4	PM-BL	0.499	0.109	0.021	---	---	0.165	10.349	---
	M-RKHS	0.459	0.109	---	0.007	0.508	---	---	3.303
	PM-RKHS	0.376	0.116	0.142	0.059	0.515	0.173	---	---
	P	0.521	0.124	0.332	0.120	---	---	---	---
	M-BL	0.454	0.110	---	---	---	---	8.725	2.857
E5	PM-BL	0.432	0.105	0.019	0.006	---	---	8.658	3.086
	M-RKHS	0.298	0.090	---	---	0.699	0.180	---	---
	PM-RKHS	0.176	0.071	0.180	0.069	0.583	0.165	---	---
	P	0.311	0.099	0.525	0.147	---	---	---	---
	M-BL	0.516	0.112	---	---	---	---	10.912	3.479
E6	PM-BL	0.489	0.100	0.021	---	---	0.184	10.474	---
	M-RKHS	0.347	0.104	---	0.007	0.619	---	---	3.447
	PM-RKHS	0.213	0.075	0.273	0.112	0.368	0.149	---	---
	P	0.656	0.138	0.264	0.107	---	---	---	---
	M-BL	0.446	0.114	---	---	---	---	8.743	3.173
E7	PM-BL	0.425	0.102	0.019	0.005	---	---	8.696	2.847
	M-RKHS	0.321	0.091	---	---	0.631	0.172	---	---
	PM-RKHS	0.216	0.081	0.166	0.068	0.567	0.168	---	---
	P	0.671	0.134	0.253	0.102	---	---	---	---
	M-BL	0.462	0.132	---	---	---	---	8.673	4.330
E8	PM-BL	0.448	0.121	0.020	0.008	---	---	8.326	3.815
	M-RKHS	0.250	0.085	---	---	0.797	0.200	---	---
	PM-RKHS	0.214	0.088	0.142	0.057	0.695	0.204	---	---
	P	0.779	0.141	0.198	0.085	---	---	---	---
	M-BL	0.592	0.125	---	---	---	---	12.409	4.163
E9	PM-BL	0.596	0.133	0.027	0.013	---	---	13.990	6.661
	M-RKHS	0.379	0.108	---	---	0.610	0.189	---	---
	PM-RKHS	0.278	0.104	0.151	0.064	0.595	0.188	---	---
	P	0.577	0.128	0.254	0.106	---	---	---	---
	M-BL	0.508	0.127	---	---	---	---	11.741	5.258
E10	PM-BL	0.472	0.084	0.021	---	---	0.192	10.681	---
	M-RKHS	0.249	0.115	---	0.007	0.730	---	---	4.166

PM-RKHS	0.172	0.073	0.164	0.064	0.616	0.176	---	---

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

Table 2. Posterior estimates of the mean and standard deviation (SD) of parameters $\sigma_{\varepsilon}^{2}, \sigma_{u}^{2}, \sigma_{f}^{2}$, and λ from the full-data analysis of *Exserohilum turcicum* on maize lines genotyped with 1152 SNP molecular markers. Five models* were fitted to each of three international environments: El Batán (Mexico), Harare (Zimbabwe), and Mpongwe (Zambia).

Environment	Model*	σ_{ε}^{2}		σ_{u}^{2}		σ_{f}^{2}		λ	
		Mean	SD	Mean	SD	mean	SD	mean	SD
El Batán (México)	K	0.190	0.086	0.395	0.081	---	---	---	---
	M-BL	0.186	0.070	---	---	---	---	18.018	4.719
	KM-BL	0.201	0.076	0.012	0.002	---	---	19.014	5.125
	M-RKHS	0.170	0.079	---	---	0.862	0.168	---	---
	KM-RKHS	0.165	0.072	0.257	0.093	0.353	0.184	---	---
Harare (Zimbabwe)	K	0.371	0.098	0.304	0.071	---	---	---	---
	M-BL	0.431	0.091	---	---	---	---	33.078	7.591
	KM-BL	0.452	0.090	0.008	0.002	---	---	35.319	7.835
	M-RKHS	0.253	0.094	---	---	0.786	0.155	---	---
	KM-RKHS	0.269	0.092	0.190	0.069	0.390	0.172	---	---
Mpongwe	K	0.283	0.072	0.265	0.054	---	---	---	---
	M-BL	0.343	0.067	---	---	---	---	32.04	6.315
	KM-BL	0.331	0.062	0.006	0.001	---	---	30.88	5.966
	M-RKHS	0.170	0.059	---	---	0.684	0.108	---	---
	KM-RKHS	0.184	0.063	0.142	0.051	0.396	0.142	---	---

*The five models are: kinship model (K); molecular marker regression model using Bayesian LASSO (M-BL); kinship model (K) plus the molecular marker model regression using Bayesian LASSO (KM-BL); molecular marker model using reproducing kernel Hilbert space (M-RKHS) regression; and kinship model (K) plus molecular marker model using reproducing kernel Hilbert space regression (KM-RKHS).

Table 3. Posterior estimates of the mean and standard deviation (SD) of parameters $\sigma_{\varepsilon}^{2},\sigma_{u}^{2},\sigma_{f}^{2}$, and λ from the full-data analysis of *Cercospora zeae-maydis* on maize lines genotyped with 1152 SNP molecular markers. Five models* were fitted to each of two international environments: Pereira (Colombia) and San Pedro Lagunillas, (Mexico).

Environment	Model*	σ_{ε}^{2}		σ_{u}^{2}		σ_{f}^{2}		λ	
		Mean	SD	mean	SD	mean	SD	mean	SD
Pereira (Colombia)	K	0.119	0.039	0.351	0.044	---	---	---	---
	M-BL	0.205	0.062	---	---	---	---	20.314	5.104
	KM-BL	0.208	0.052	0.005	0.001	---	---	20.618	4.149
	M-RKHS	0.100	0.034	---	---	0.780	0.087	---	---
	KM-RKHS	0.106	0.034	0.252	0.060	0.229	0.112	---	---
San Pedro Lagunillas (Mexico)	K	0.245	0.067	0.265	0.052	---	---	---	---
	M-BL	0.377	0.064	---	---	---	---	37.011	7.160
	KM-BL	0.377	0.069	0.006	0.001	---	---	37.474	8.360
	M-RKHS	0.211	0.064	---	---	0.612	0.109	---	---
	KM-RKHS	0.204	0.060	0.191	0.052	0.230	0.101	---	---

*The five models are: kinship model (K); molecular marker regression model using Bayesian LASSO (M-BL); kinship model (K) plus the molecular marker model regression using Bayesian LASSO (KM-BL); molecular marker model using reproducing kernel Hilbert space (M-RKHS) regression; and kinship model (K) plus molecular marker model using reproducing kernel Hilbert space regression (KM-RKHS).

Predictive ability

Tables 4 and 5 show correlations between the phenotypic outcomes and the predicted values for each model in each environment and each trait. For the wheat data, PM models gave sizable gains in predictive ability relative to P; the only exception was E6, where PM-BL performed similarly to P. This provides evidence that indicates genomic selection can be effective for predicting genetic values of wheat lines. In this data set, in most environments (E2, E3, and E5-E10) RKHS outperformed BL; however, in E4 the opposite was true, indicating that model choice needs to be assessed for each trait/environment (Table 4). In general, correlations between predictive and observed values were higher for the PM-BL model than for the M-BL model, but this was not the case for PM-RKHS versus M-RKHS. Previous studies (e.g., de los Campos et al., 2009a) have also shown only slight differences between the predictive ability of the M and PM models, indicating that once markers are included, using pedigree information may not affect, or may increase only slightly, the ability of the model to predict future data.

For the NCLB maize data (Table 5), the CV correlations ranged from 0.44 to 0.52. These values are smaller than those obtained for the wheat data. More importantly, the differences in CV correlations between models were not as marked as in the wheat data set. This was expected because, as stated, in this data set all models (including K) are marker based. For this trait, as it was for the wheat data set, in some environments (Harare and Mpongue) RKHS outperformed BL, but in El Batán the opposite was true.

The analysis of predictive ability of models for GLS in maize (Table 5) gave CV correlations ranging from 0.425 to 0.600. Also for this trait, predictive ability seemed to be higher in Pereira than in San Pedro Lagunillas. And using either the kinship matrix or the Gaussian kernel described above, RKHS slightly outperformed the Bayesian LASSO.

Table 4. Predictive ability measured as the correlation between predicted and actual phenotypes, obtained in a 10-fold cross-validation, from data analysis of grain yield of 93 wheat lines genotyped with 234 DArT molecular markers. Five models* were fitted to each of 10 international environments (E1-E10).

	P	M-BL	PM-BL	M-RKHS	PM-RKHS
E1	-0.085	0.497	0.499	0.543	0.471
E2	0.494	0.590	0.596	0.686	0.673
E3	0.515	0.567	0.568	0.691	0.685
E4	-0.067	0.514	0.509	0.482	0.446
E5	0.519	0.527	0.539	0.636	0.657
E6	0.559	0.531	0.554	0.607	0.666
E7	0.341	0.582	0.591	0.650	0.624
E8	0.255	0.466	0.469	0.617	0.540
E9	0.204	0.449	0.450	0.596	0.561
E10	0.445	0.506	0.530	0.659	0.665

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

Table 5. Predictive ability measured as the correlation between predicted and actual phenotypes, obtained in a 10-fold cross-validation, from the analysis of *Exserohilum turcicum* (first three rows) and *Cercospora zeae-maydis* (last two rows) in maize. Lines were genotyped with 1152 SNP molecular markers. Five models* were fitted to each trait/environment combination.

Environment	K	M-BL	KM-BL	M-RKHS	KM-RKHS
----------------------------*Exserohilum turcicum*-------------------------------					
El Batán (Mexico)	0.472	0.509	0.512	0.472	0.484
Harare (Zimbabwe)	0.447	0.465	0.462	0.500	0.475
Mpongue (Zambia)	0.464	0.452	0.450	0.521	0.499
------------------------*Cercospora zeae-maydis*------------------------------					
Pereira (Colombia)	0.602	0.549	0.550	0.585	0.600
San Pedro Lagunillas (Mexico)	0.451	0.425	0.425	0.465	0.460

*The five models are: kinship model (K), molecular marker regression model using Bayesian LASSO (M-BL); kinship model (K) plus the molecular marker model regression using Bayesian LASSO (KM-BL); molecular marker model using reproducing kernel Hilbert pace (M-RKHS) regression; and kinship model (K) plus molecular marker model using reproducing kernel Hilbert space regression (KM-RKHS).

Patterns of marker effects in ten environments for the wheat data

Principal component analysis on the estimated effects of the presence of each DArT computed from the PM-BL model for the wheat data is depicted in the biplot in Fig. 1. The first two principal components explained 82.30% of the total variability in estimated DArT effects. The pattern of correlations between estimated DArT effects reflects the patterns of phenotypic correlations observed for the wheat data (see Fig. 1 and Table 6). These patterns allow classifying environments into three groups. Group 1 includes environments E2, E3, E6, E7, E8, E9, and E10, which form a compact cluster located on the left-hand side of the biplot; these environments have strong correlations with each other based on phenotypic as well as estimated marker effect data (Table 6). Group 2 has E5 located on the other side of the biplot; it has strong negative correlations (phenotypic and based on estimates of marker effects) with the environments in Group 1

(E2, E3, E6, E7, E8, E9, and E10). Finally, E4 and E1 of Group 3 showed a moderate to low correlation with each other and with the environments in Group 1 and Group 2.

Evidently, E5 causes a great deal of the interaction between molecular marker effects and environments. This is evidenced in its correlation with E1 and E4, which is low in absolute value, and in its highly negative correlations with the other environments located on the opposite side of the biplot (e.g., E2 and E3). E1 and E4, which had very low phenotypic correlations with E2, E3, E6, E7, E8, E9, and E10, also showed that genotype × environment is an important component of genetic variability. It should be pointed out that the correlations among 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 majority of the estimated effects of the 234 DArT markers are located around the center of the biplot (i.e., estimated effects were small, in absolute value), which reflects shrinkage of the BL model. However, some DArTs (identified by name in Fig. 1) had estimated effects that were large in absolute value (shown in Table 7). The estimated effect of the presence of a DArT in 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, the presence of DArT markers wPt.4223 and wPt.1859, among others, is expected to increase grain yield performance in E1, E4, and E5 and strongly decrease grain yield in all environments in Group 1. Those DArTs whose presence is expected to increase or decrease grain yield across environments can be viewed as contributing to positive genetic correlations in grain yield among environments. On the other hand, DArTs whose presence is expected to increase (or decrease) grain yield in one environment, and decrease (or increase) in others can be viewed as causes of genotype × environment interaction.

Figure 1. Biplot of the first and second principal component axes (Comp. 1 and Comp. 2) of the grain yield effect of 234 DArTs estimated from the full-data PM-BL model for the wheat data set in each of 10 environments (E1-E10). Only the effects of 21 DArTs located far from the center of the biplot were identified with their corresponding DArT name (filled-in circles). Three groups of environments and molecular markers are delineated as Groups 1, 2, and 3.

Fig. 1

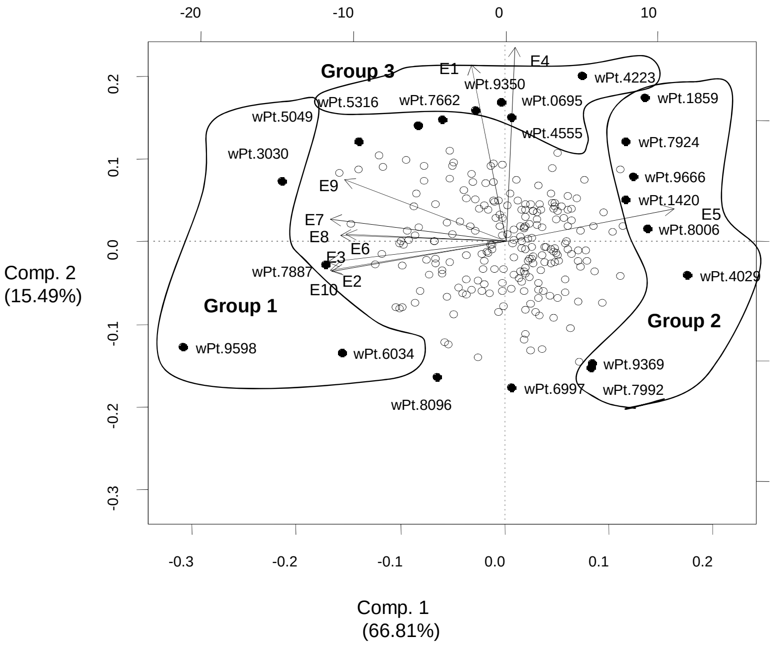


Table 6. Correlation between phenotypic data (upper triangular) and between estimates of marker effects (lower triangular) from data analysis of wheat grain yield from 10 international environments (E1-E10).

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
E1	---	0.195	0.219	0.441	-0.221	0.199	0.295	0.240	0.305	0.176
E2	0.073	---	0.927	-0.001	-0.901	0.802	0.845	0.826	0.784	0.897
E3	0.100	0.918	---	0.015	-0.895	0.811	0.869	0.860	0.804	0.918
E4	0.432	-0.146	-0.117	---	0.046	0.073	0.172	0.060	0.248	-0.044
E5	-0.109	-0.876	-0.869	0.169	---	-0.778	-0.830	-0.812	-0.707	-0.880

E6	0.135	0.746	0.760	-0.007	-0.732	----	0.827	0.768	0.735	0.774
E7	0.219	0.833	0.865	-0.037	-0.816	0.832	----	0.805	0.842	0.815
E8	0.147	0.771	0.830	-0.016	-0.764	0.731	0.812	----	0.783	0.821
E9	0.287	0.751	0.790	-0.205	-0.660	0.720	0.849	0.784	---	0.765
E10	0.094	0.908	0.919	-0.158	-0.872	0.747	0.844	0.798	0.768	----

Table 7. Estimated effects of 21 DArT molecular markers located farthest from the center of the biplot (Fig. 1) of principal component analysis of the marker effects in each of 10 international environments (E1-E10) for wheat grain yield data.

DArT	Group*	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
wPt.3030	1	0.059	0.146	0.116	0.038	-0.113	0.090	0.126	0.146	0.102	0.126
wPt.5049	1	0.086	0.065	0.085	0.036	-0.024	0.097	0.128	0.077	0.088	0.056
wPt.6034	1	-0.041	0.104	0.124	-0.067	-0.119	0.044	0.107	0.168	0.025	0.076
wPt.7887	1	-0.039	0.082	0.104	0.007	-0.122	0.056	0.105	0.201	0.074	0.076
wPt.9598	1	-0.013	0.242	0.284	-0.048	-0.229	0.069	0.205	0.109	0.083	0.196
wPt.1420	2	0.040	-0.067	-0.099	0.019	0.040	-0.027	-0.067	-0.098	-0.032	-0.061
wPt.1859	2	0.027	-0.110	-0.099	0.108	0.146	-0.036	-0.049	-0.053	-0.001	-0.084
wPt.4029	2	-0.053	-0.075	-0.107	0.003	0.082	-0.063	-0.083	-0.176	-0.063	-0.083
wPt.7924	2	0.028	-0.109	-0.114	0.070	0.068	-0.045	-0.039	-0.008	-0.014	-0.081
wPt.7992	2	-0.065	-0.056	-0.040	-0.088	0.033	-0.042	-0.043	-0.051	-0.040	-0.016
wPt.8006	2	0.021	-0.092	-0.083	0.000	0.068	-0.063	-0.052	-0.084	-0.051	-0.072
wPt.9369	2	-0.051	-0.045	-0.037	-0.084	0.017	-0.048	-0.059	-0.046	-0.046	-0.019
wPt.9666	2	0.061	-0.071	-0.100	0.025	0.057	-0.040	-0.070	-0.088	-0.026	-0.070
wPt.4223	2,3	0.220	-0.075	-0.055	0.031	0.044	-0.020	-0.025	-0.051	-0.017	-0.046
wPt.4555	3	0.093	-0.005	-0.019	0.076	0.009	0.000	0.023	0.002	0.010	-0.007
wPt.0695	3	0.121	0.001	-0.006	0.063	0.007	-0.015	0.018	0.006	0.042	-0.013
wPt.5316	3	0.080	0.034	0.059	0.062	-0.024	0.055	0.073	0.057	0.069	0.011
wPt.9350	3	0.110	0.009	0.032	0.075	-0.005	-0.023	0.033	0.031	0.032	0.026
wPt.7662	3	0.068	0.055	0.039	0.104	-0.057	0.027	0.033	0.046	0.029	0.016
wPt.6997	--	-0.154	0.014	0.007	-0.040	-0.018	0.002	-0.001	-0.004	-0.015	0.020
wPt.8096	--	-0.007	0.107	0.058	-0.125	-0.040	0.004	0.013	0.012	0.028	0.059

* Group of markers delineated in Fig. 1.

Patterns of marker effects in three environments for *Exserohilum turcicum* (NCLB) of the maize data

Principal component analysis on the estimated effects of the 1152 single nucleotide polymorphisms (SNPs) computed from the KM-BL model for the maize data is depicted in the biplot in Fig. 2. The first two principal components explained 83.58% of the total variability in estimated SNP effects. The correlation among the three environments (based on phenotypic data or based on estimates of marker effects) was 0.50 between Mpongwe and Harare and 0.26 between El Batán and Mpongwe and between El Batán and Harare. This pattern of correlations (depicted in Fig. 2) may indicate the presence of different races of NCLB in these environments, i.e., a pathogen population in El Batán (México) that might be different from populations in Harare and Mpongwe (Southern Africa).

The estimated effect of the allele coded as one in each SNPs and in a given environment is obtained by orthogonal projection of the estimated marker effect on the vector of the corresponding environment. Here, SNPs whose alleles coded with 1 had positive effects (i.e., those with estimated effects pointing in the same direction as the environment) are expected to produce an increase of the disease on the genotypes in that environment. Therefore, selection should try to decrease the frequency of the allele coded with 1. The opposite is true for SNPs whose estimated effects were negative.

The 21 SNPs with the largest effects (positive or negative) on NCLB (Table 8) are located farthest on the biplot in Fig. 2. Some SNPs whose alleles coded with 1 had negative effects for the three sites (e.g., PZA03153.3, see Fig. 2 and Table 8) and are located in the opposite direction of these sites were marked as Group 1 in Fig. 2; the presence of the allele coded with 1 of these SNPs should confer some degree of resistance to NCLB across environments. The opposite is true for SNPs in Group 2 and 3 (Fig. 2); here the presence of the allele coded with 1 is expected to favor the disease in El Batán (Group 2) and Harare and Mpongwe (Group 3); selection should thus aim at decreasing this allele.

Finally, caution must be exercised when interpreting the above results since they were obtained from one-year data, and the prevalence of NCLB races can change from year to year, depending on changes in environmental conditions (temperature, rainfall, relative humidity, etc).

Figure 2. Biplot of the first and second principal component axes (Comp. 1 and Comp. 2) of the *Exserohilum turcicum* (NCLB) disease effect of 1152 SNPs estimated from the full-data KM-BL model for the maize data set in each of three environments: El Batán (México), Harare (Zimbabwe), and Mpongwe (Zambia). Only the effects of 24 SNPs located far from the center of the biplot were identified with their corresponding SNPs name (filled-in circles). Three groups of environments and molecular markers are delineated as Groups 1, 2, and 3.

Fig. 2

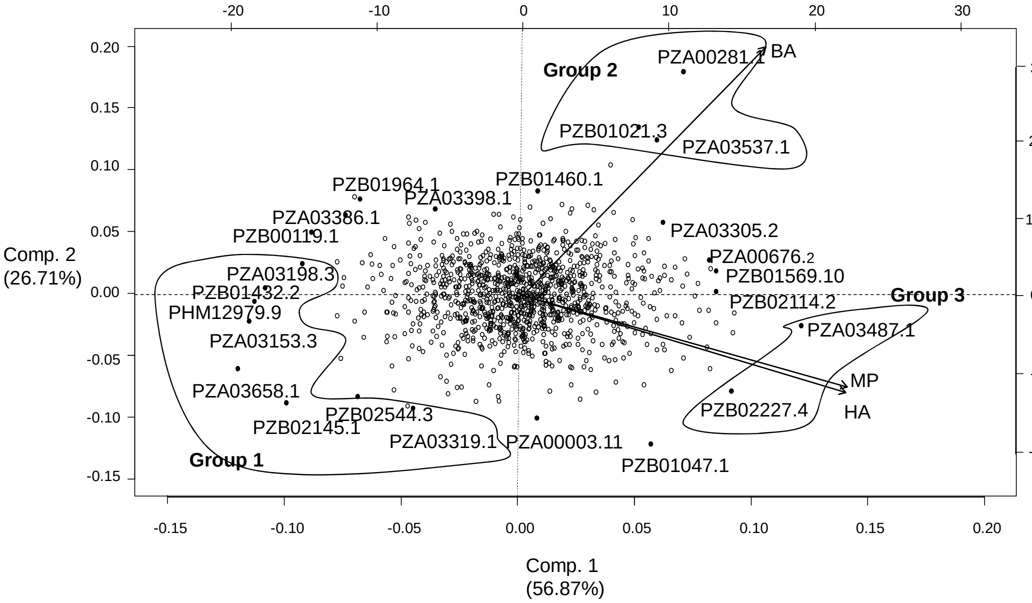


Table 8. Estimated effects of 24 SNP molecular markers located farthest from the center of the biplot (Fig. 2) of principal component analysis of the marker effects in each of three international environments: El Batán (Mexico), Mpongwe (Zambia), and Harare (Zimbabwe) for the maize *Exserohilum turcicum* (NCLB) disease. The 5 SNP molecular markers with the largest positive estimated effects and the 5 SNP molecular marker with the largest negative estimated effects in each of the two international environments: Pereira (Colombia) and San Pedro Lagunilla (México) for the maize *Cercospora zeae-maydis* (GLS) disease (also identified with their names in Fig. 3).

	Exserohilum turcicum (NCLB)		
SNP*	El Batán	Mpongwe	Harare
PHM12979.9	-0.021	-0.047	-0.018
PZB00119.1	-0.005	0.004	-0.056
PZB02145.1	-0.038	-0.007	-0.028
PZA03658.1	-0.035	-0.043	-0.015
PZB01047.1	-0.017	0.012	0.038
PZA00003.11	-0.021	0.013	0.011
PZA03487.1	0.015	0.041	0.030
PZB01021.3	0.039	-0.005	0.004
PZB02544.3	-0.031	-0.014	-0.009
PZB01460.1	0.020	-0.005	-0.008
PZA00676.2	0.021	0.014	0.023
PZA03305.2	0.024	0.011	0.010
PZB02114.2	0.015	0.017	0.026
PZA00281.1	0.053	0.004	-0.001
PZB02227.4	-0.002	0.045	0.022
PZA03198.3	-0.011	-0.037	-0.021
PZA03153.3	-0.025	-0.039	-0.023
PZA03398.1	0.009	-0.009	-0.023
PZB01432.2	-0.018	-0.044	-0.020
PZA03386.1	0.001	-0.027	-0.027
PZA03537.1	0.039	-0.006	0.011
PZB01569.10	0.019	0.018	0.023
PZA03319.1	-0.029	-0.002	-0.005

PZB01964.1	0.005	-0.021	-0.031
	Cercospora zeae-maydis (GLS)		
SNP[#]	Pereira	SNP[#]	San Pedro Lagunillas
PZA03575.2	0.0548	PZD00030.2	0.0309
PZA03154.2	0.0495	PZA00069.4	0.0275
PZB01977.1	0.0465	PZA03676.1	0.0261
PZA02982.5	0.0465	PZA00289.11	0.0250
PZA03284.3	0.0440	PZA03404.1	0.0240
PZB01944.1	-0.0516	PZA03559.1	-0.0325
PZB01482.3	-0.0485	PZA03651.2	-0.0275
PZB01111.6	-0.0466	PZB00857.2	-0.0254
PZB01487.1	-0.0451	PZB01471.3	-0.0238
PZA03570.1	-0.0437	PZA00615.3	-0.0232

* Name of the markers given in Fig. 2.

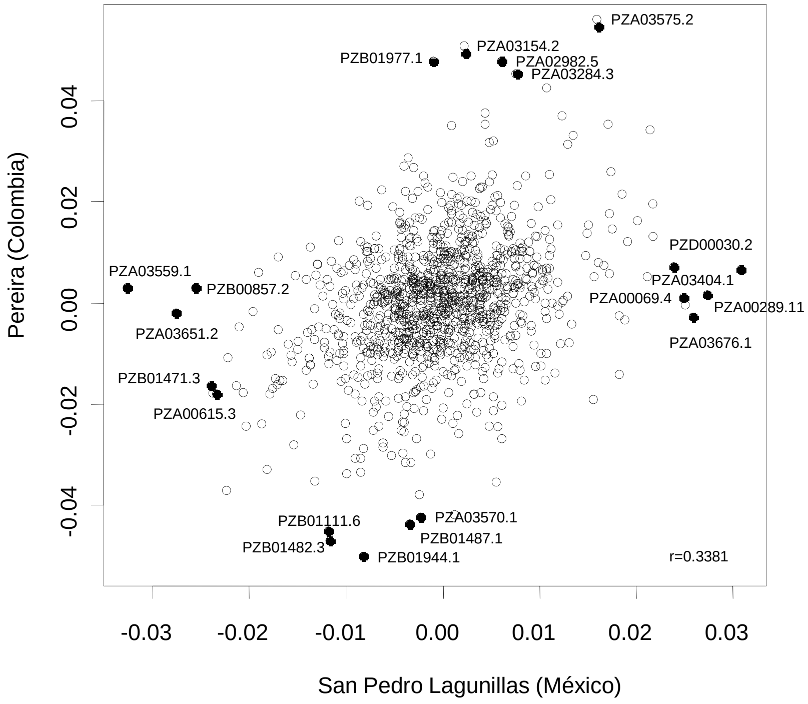
[#] Name of the markers given in Fig. 3.

Patterns of marker effects in two environments for *Cercospora zeae-maydis* (GLS) of the maize data

The effects of the SNPs molecular markers for GLS in two environments, Pereira and San Pedro Lagunilla are depicted in the scatter-plot (Fig. 3). Based on the molecular marker effect, the two sites had a correlation of 0.3381. The five SNPs with the largest positive effects and the five SNPs with the largest negative effect on GLS (Table 8) are identified by their names in Fig. 3. Markers located toward the lower left corner of the scatter-plot of Fig. 3 have a negative effect on GLS (i.e., PZB01482.3, PZA00615.3) and provide some degree of resistance to GLS disease in both locations. The opposite is true for markers located on the upper right corner of Fig. 3.

Figure 3. Scatter-plot of *Cercospora zeae-maydis* (GLS) disease effect of 1152 SNPs estimated from the full-data KM-BL model for the maize data set in each of two environments: Pereira (Colombia), and San Pedro Lagunilla (México). Only the effects of 5 SNPs with the largest positive and negative effects in both environments were identified with their corresponding SNPs name (filled-in circles).

Fig. 3



CONCLUDING REMARKS

The results of this study are encouraging; they indicate that, even with a modest number of molecular markers, models for GS can attain high predictive ability for genetic values of traits of economic interest and under contrasting environmental conditions.

Also, marker-based models gave important gains in predictive ability relative to pedigree-based models. 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 and in various environments have yet to be observed.

In general, M-RHKS and PM-RHKS had similar predictive abilities in the wheat data set, and in most cases they outperformed the parametric counterparts, M-BL and PM-BL. The increase in predictive ability of the PM models as compared to the M models was either nonexistent or only marginal. Similar results were found by de los Campos et al. (2009a), perhaps because there is some redundancy between regression on the pedigree and regression on the markers (e.g., Habier et al., 2009). The number of markers evaluated in the wheat data set was small, and it is expected that larger gains in predictive ability can be achieved with high-density markers.

For the maize data set, the difference in predictive ability between models was not as marked as in the wheat data set. This occurred because all models in the maize data set were marker-based; they only differed in how marker data were included in the models. These results also illustrate certain robustness of models for GS with respect to the choice of how marker data are included in the model.

An advantage of models including a parametric regression on marker covariates, such as M-BL and PM-BL, is that, in addition to estimating genetic values, they also provide information on ‘marker effects’. This information can be used to attain a better understanding of the genetic architecture of the traits under study.

In our study separate models were fitted to each trait/environment. An alternative to these single-environment models for GS is to use multiple-environment (or, equivalently, multiple-trait) models where genetic values and marker effects on several traits/environments are jointly estimated. Multiple-environment models allow borrowing information between correlated environments; thus it can be speculated that multiple-environment GS models can yield similar or even better predictions for individual environments. The literature on GS has focused on single-trait models only; the development of multiple-trait models for GS appears as a natural next step.

The results of this study can also be used to generate a better design for field evaluations. For example, they show that in CIMMYT’s Global Wheat Breeding Program, prediction of unobserved wheat lines in any of the correlated environments in Group 1 (Fig. 1) should be relatively accurate, and the scheme for testing wheat lines in any of those environments should be planned accordingly. It can be speculated that only one of the environments in Group 1 should be included in the trial, since information lacking on the other environments can be borrowed from the one in use. However, unobserved lines in E1 and E4 (which have low correlations with environments in Group 1) are expected to be poorly predicted when using observed data from environments in Group 1. Concerning the analysis of NCLB in the maize data, the results show that estimated markers effects in the two African sites (Mpongwe and Harare) (Fig. 2) were positively correlated and they were negatively correlated with estimates of marker effects obtained for the same trait in in Mexico (El Batán), which is suggestive of marker effect × environment interaction. The scatter-plot of Fig. 3 suggested marker effect × environment interaction and confirmed the low correlation between environments (0.3381).

ACKNOWLEDGMENTS

The maize data sets used in this study come 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 is acknowledged.

REFERENCES

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

Box, G.E.P., and Cox, D.R. 1964. An analysis of transformation (with discussion). Journal of the Royal Statistical Society B, 26: 211-252.

Burgueño, J., J. Crossa, P.L. Cornelius, R. Trethowan, G. McLaren, and A. Krishnamachari. 2007. Modeling additive × environment and additive × additive × environment using genetic covariances of relatives of wheat genotypes. Crop Science 43: 311-320.

Cornelius, P.L., Crossa, J., Seyedsadr, M.S., Liu, G., and Viele, K. 2001. Contributions to multiplicative model analysis of genotype-environment data. Statistical Consulting Section, American Statistical Association, Joint Statistical Meetings, Proceedings, American Statistical Association, CDROM.

Crossa, J., van Eeuwijk, F.,Vargas, M., and Cornelius, P.L. 2001. Linear, bilinear and linear-bilinear models for analyzing genotype × environment interaction. Statistical Consulting Section, American Statistical Association, Joint Statistical Meetings, Proceedings, American Statistical Association, CDROM.

Crossa, J., R-C, Yang, and P.L. Cornelius. 2004. Studying crossover genotype × environment interaction using linear-bilinear models and mixed models. Journal of Agricultural, Biological and Environmental Statistics (JABES) 9(3): 362-380.

Crossa, J., J. Burgueño, P.L. Cornelius, G. McLaren, R. Trethowan, and A. Krishnamachari. 2006. Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Science 46: 1722-1733.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 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. DOI.10.2527/as.2008-1259.

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

Gabriel, K.R. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58: 453-467.

Gabriel, K.R. 1978. Least squares approximation of matrices by additive and multiplicative models. Journal of the Royal Statistical Society, Series B 40: 186-196.

Gianola, D., Fernando, R.L., and Stella, A. 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, and S. Avendano. 2008. Non parametric methods for incorporating genomic information into genetic evaluation: An application to mortality in boilers. 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.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Ed. Springer-New York.

Henderson, C.R. 1975. Best inear unbiased estimation and prediction under a selection model. Biometrics, 31:423-447.

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

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. Bǔchse. 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.

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

Sorensen, D., and Gianola, D. 2002. Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. Springer.

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

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Reliability of genomic predictions for North American Holstein bulls (invited review). J. of Dairy Science 92: 16-24.

Ward, J.M.J., Stromberg, E.L., Nowell, D.C., and Nutter jr F.W. 1999. Gray leaf spot: a disease of global importance in maize production. Plant Disease 83:884-895.

Welz, H.G., and Geiger, H.H. 2000. Genes for resistance to northern corn leaf blight in diverse maize populations. Plant Breeding 119: 1-14

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68(1): 49-67.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67(2): 301-320.

Appendix

As stated, a slight extension of the algorithm used to draw samples from the posterior distribution of the pedigree model, given by [2b], can be used to obtain samples from the posterior distribution of PM-RKHS, [7b],

$$p\left(\mu,\mathbf{a},\mathbf{f},\sigma_\varepsilon^2,\sigma_a^2,\sigma_f^2\middle|\bar{\mathbf{y}},H\right)\propto\left\{\prod_{i=1}^{n}N\left(\bar{y}_i\middle|\mu+a_i+f_i,\frac{\sigma_\varepsilon^2}{n_i}\right)\right\}N\left(\mathbf{a}\middle|\mathbf{0},\mathbf{A}\sigma_a^2\right)N\left(\mathbf{f}\middle|\mathbf{0},\mathbf{K}_{\mathrm{RKHS}}\sigma_f^2\right)$$

$$\times\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_a^2\middle|df_a,S_a\right)\chi^{-2}\left(\sigma_f^2\middle|df_f,S_f\right).$$

To see this, note that, from [7b], the fully conditional distribution of $\{\mu,\mathbf{a},\sigma_a^2,\sigma_\varepsilon^2\}$ is proportional to:

$$p\left(\mu,\mathbf{a},\sigma_a^2,\sigma_\varepsilon^2\middle|\bar{\mathbf{y}},\mathbf{f}\right)\propto\left\{\prod_{i=1}^{n}N\left(\bar{y}_i\,\middle|\,\mu+a_i+f_i,\frac{\sigma_\varepsilon^2}{n_i}\right)\right\}N\left(\mathbf{a}\middle|\mathbf{0},\mathbf{A}\sigma_a^2\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_a^2\middle|df_a,S_a\right)$$

$$\propto\left\{\prod_{i=1}^{n}N\left(\bar{y}_i^*\,\middle|\,\mu+a_i,\frac{\sigma_\varepsilon^2}{n_i}\right)\right\}N\left(\mathbf{a}\middle|\mathbf{0},\mathbf{A}\sigma_a^2\right)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma_a^2\middle|df_a,S_a\right),$$

where $\bar{y}_i^*=\bar{y}_i-f_i$ is an off-set obtain by subtracting from the data the contribution of f_i.

The above joint conditional distribution is the same as that of the pedigree model with the original data, $\{\bar{y}_i\}$, replaced by $\bar{y}_i^*$. Therefore, algorithms used to draw samples from the posterior distribution of a pedigree model (e.g., that described in Sorensen and Gianola, 2002) can be used to draw samples from the posterior distribution of $\{\mu,\mathbf{a},\sigma_a^2,\sigma_\varepsilon^2\}$.

Once these unknowns have been updated, the Gibbs sampler draws samples of $\{\mathbf{f},\sigma_f^2\}$ from the corresponding fully-conditional distribution, which from [7b] is:

$$p\left(\mathbf{f},\sigma_f^2\middle|\bar{\mathbf{y}},\mathbf{a},\mu\right)\propto\left\{\prod_{i=1}^{n}N\left(\bar{y}_i\middle|\,\mu+a_i+f_i,\frac{\sigma_\varepsilon^2}{n_i}\right)\right\}N\left(\mathbf{f}\middle|\mathbf{0},\mathbf{K}_{\mathrm{RKHS}}\sigma_f^2\right)\chi^{-2}\left(\sigma_f^2\middle|df_f,S_f\right).$$

$$\propto\left\{\prod_{i=1}^{n}N\left(\bar{y}_i^{**}\middle|\,f_i,\frac{\sigma_\varepsilon^2}{n_i}\right)\right\}N\left(\mathbf{f}\middle|\mathbf{0},\mathbf{K}_{\mathrm{RKHS}}\sigma_f^2\right)\chi^{-2}\left(\sigma_f^2\middle|df_f,S_f\right),$$

where $\bar{y}_i^{**} = \bar{y}_i - \mu - a_i$. The above conditional distribution has the same structure as that of $\left(\mathbf{a}, \sigma_a^2 \middle| \bar{\mathbf{y}}, \mu, \sigma_\varepsilon^2\right)$ in the pedigree model, with: (a) the original data, $\bar{y}_i$ replaced by $\bar{y}_i^{**}$, and (b) $\{\mathbf{a}, \mathbf{A}, df_a, S_a\}$ replaced with $\{\mathbf{f}, \mathbf{K}_{\text{RKHS}}, df_f, S_f\}$ .Therefore, the same algorithms used to draw samples of $p\left(\mathbf{a}, \sigma_a^2 \middle| \text{else}\right)$ in the pedigree model can be used to draw samples of $p\left(\mathbf{f}, \sigma_f^2 \middle| \text{else}\right)$.