DOI: 10.1534/genetics.110.118521

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,§ Juan Burgueño,*,‡ José Luis Araus,* Dan Makumbi,* Ravi P. Singh,* Susanne Dreisigacker,* Jianbing Yan,* Vivi Arief, Marianne Banziger* and Hans-Joachim Braun***

*International Maize and Wheat Improvement Center (CIMMYT), 06600, México DF, México, †Department of Biostatistics, University of Alabama, Birmingham, Alabama 35216, §Departments of Animal Science, Dairy Science, and Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53706, ‡Colegio de Postgraduados, 50230, Montecillo, Edo. de Mexico Montecillos, México and **School of Land Crop and Food Sciences, University of Queensland, 4072, Sta. Lucia, Queensland, Australia

Manuscript received May 5, 2010
Accepted for publication July 28, 2010

ABSTRACT

The availability of dense molecular markers has made possible the use of genomic selection (GS) for plant breeding. However, the evaluation of models for GS in real plant populations is very limited. This article evaluates the performance of parametric and semiparametric models for GS using wheat (*Triticum aestivum* L.) and maize (*Zea mays*) data in which different traits were measured in several environmental conditions. The findings, based on extensive cross-validations, indicate that models including marker information had higher predictive ability than pedigree-based models. In the wheat data set, and relative to a pedigree model, gains in predictive ability due to inclusion of markers ranged from 7.7 to 35.7%. Correlation between observed and predictive values in the maize data set achieved values up to 0.79. 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.

PEDIGREE-BASED prediction of genetic values based on the additive infinitesimal model (Fisher 1918) has played a central role in genetic improvement of complex traits in plants and animals. Animal breeders have used this model for predicting breeding values either in a mixed model (best linear unbiased prediction, BLUP) (Henderson 1984) or in a Bayesian framework (Gianola and Fernando 1986). More recently, plant breeders have incorporated pedigree information into linear mixed models for predicting breeding values (Crossa *et al.* 2006, 2007; Oakey *et al.* 2006; Burgueño *et al.* 2007; Piepho *et al.* 2007).

The availability of thousands of genome-wide molecular markers has made possible the use of genomic selection (GS) for prediction of genetic values (Meuwissen *et al.* 2001) in plants (*e.g.*, Bernardo and Yu 2007; Piepho 2009; Jannink *et al.* 2010) and animals (Gonzalez-Recio *et al.* 2008; VanRaden *et al.* 2008; Hayes *et al.* 2009; de los Campos *et al.* 2009a). Implementing GS poses several statistical and computational challenges, such as how models can cope with the curse of dimensionality, collinearity between markers, or the complexity of quantitative traits. Parametric (*e.g.*, Meuwissen *et al.* 2001) and semiparametric (*e.g.*, Gianola *et al.* 2006; Gianola and van Kaam 2008) methods address these problems differently.

In standard genetic models, phenotypic outcomes, y_i ($i = 1, \ldots, n$), are viewed as the sum of a genetic value, g_i, and a model residual, ε_i; that is, $y_i = g_i + \varepsilon_i$. In parametric models for GS, g_i is described as a regression on marker covariates x_{ij} ($j = 1, \ldots, p$ molecular markers) of the form $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$, such that

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

(or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, in matrix notation), where β_j is the regression of y_i on the jth marker covariate x_{ij}.

Estimation of $\boldsymbol{\beta}$ via multiple regression by ordinary least squares (OLS) is not feasible when $p > n$. 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

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.118521/DC1.

[1]*Corresponding author:* Biometrics and Statistics Unit, Crop Research Informatics Laboratory, CIMMYT, Apdo. Postal 6-641, 06600 México, D. F., México. E-mail: j.crossa@cgiar.org

[2]These authors contributed equally to this work.

Shrinkage and Selection Operator (LASSO) (Tibshirani 1996) or their Bayesian counterpart. This approach yields greater accuracy of estimated genetic values and can be coupled with geostatistical techniques commonly used in plant breeding to model multienvironments trials (Piepho 2009).

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 QTL (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 using priors of marker effects that are mixtures of scaled-normal densities. Examples of this are methods Bayes A and Bayes B of Meuwissen *et al.* (2001) and the Bayesian LASSO of Park and Casella (2008).

An alternative to parametric regressions is to use semiparametric methods such as reproducing kernel Hilbert spaces (RKHS) regression (Gianola and van Kaam 2008). The Bayesian RKHS regression regards genetic values as random variables coming from a Gaussian process centered at zero and with a (co)variance structure that is proportional to a kernel matrix **K** (de los Campos *et al.* 2009b); that is, $\text{Cov}(g_i, g_j) \propto K(\mathbf{x}_i, \mathbf{x}_j)$, 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. In a finite-dimensional setting this amounts to modeling the vector of genetic values, $\mathbf{g} = \{g_i\}$, as multivariate normal; that is, $\mathbf{g} \sim N(\mathbf{0}, \mathbf{K}\sigma_g^2)$ where σ_g^2 is a variance parameter. 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). 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 some parametric methods, especially when $p \gg n$.

This study presents an evaluation of several methods for GS, using two extensive data sets. One contains phenotypic records of a series of wheat trials and recently generated genomic data. The other data set 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 data sets were used: the first one comprises information from a collection of 599 historical CIMMYT wheat lines, and the second one includes information on 300 CIMMYT maize lines.

Wheat data set: This data set includes 599 wheat lines developed by the CIMMYT Global Wheat Breeding program. Environments were grouped into four target sets of environments (E1–E4). The trait was grain yield (GY). Hereinafter we refer to this data set as wheat-grain yield (W-GY). A pedigree was used for deriving the additive relationship matrix **A** among the 599 lines, as described in http://cropwiki.irri.org/icis/index.php/TDM_GMS_Browse (McLaren *et al.* 2005). The entries of this matrix equal twice the kinship coefficient (or coefficient of parentage) between pairs of lines.

Wheat lines were genotyped using 1447 Diversity Array Technology markers (hereinafter generically referred to as markers) generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au). These markers may take on two values, denoted by their presence (1) or absence (0). In this data set, the overall mean frequency of the allele coded as 1 was 0.561, with a minimum of 0.008 and a maximum of 0.987. Markers with allele frequency <0.05 or >0.95 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 nonmissing genotypes. After edition, 1279 markers were retained.

Maize data set: The maize data set is from the Drought Tolerance Maize for Africa project of CIMMYT's Global Maize Program. The original data set included 300 tropical lines genotyped with 1148 single-nucleotide polymorphisms (hereinafter generically referred to as markers). For each marker, the allele with lowest frequency was coded as one.

No pedigree was available for these data. Traits analyzed for this study were GY, female flowering (FFL) (or days to silking), male flowering (MFL) (or days to anthesis), and the anthesis-silking interval (ASI), each evaluated under severe drought stress (SS) and well-watered (WW) conditions. Hereinafter we refer to these data sets as maize-grain yield (M-GY) and maize-flowering (M-F), respectively. The number of lines in the M-F data set was 284, whereas 264 lines were available in M-GY. The average minor allele frequency in these data sets was 0.20. After editing (with the same procedures as those described above), the numbers of markers available for analysis were 1148 and 1135 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.* semiparametric). All the unknowns in the model were trait–environment specific. 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 $y_i = n_i^{-1} \sum_k y_{ik}$ is the average performance of the *i*th line, n_i is the number of replicates used for computing the mean value of the *i*th genotype, μ is an intercept, g_i is the genetic value of the *i*th genotype, 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 $p(\boldsymbol{\varepsilon}) = \prod_{i=1}^{n} N(\varepsilon_i \,|\, 0, \sigma_\varepsilon^2/n_i)$. With this assumption, the likelihood function becomes

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

Models differed on how pedigree and molecular marker information was included in g_i.

Standard infinitesimal model: In this model, denoted as pedigree (P), $g_i = u_i$ and $p(\mathbf{u}\,|\sigma_u^2) = N(\mathbf{u}\,|\mathbf{0}, \mathbf{A}\sigma_u^2)$, where **A** is the additive relationship matrix computed from the pedi-

gree and σ^2_u is the infinitesimal additive genetic variance. Following standard assumptions, the joint prior of model unknowns in P was

$$p(\mu, \mathbf{u}, \sigma^2_\varepsilon, \sigma^2_u \,|\, \text{d.f.}_\varepsilon, S_\varepsilon, \text{d.f.}_u, S_u) \propto N(\mathbf{u} \,|\, \mathbf{0}, \mathbf{A}\sigma^2_u)\chi^{-2}(\sigma^2_\varepsilon \,|\, \text{d.f.}_\varepsilon, S_\varepsilon)\chi^{-2}(\sigma^2_u \,|\, \text{d.f.}_u, S_u), \quad (2a)$$

where $\chi^{-2}(\sigma_\cdot^2|\text{d.f.}_\cdot, S_\cdot)$ are scaled inverse chi-square priors assigned to the variance parameters. The prior scale and degrees of freedom parameters were set to $S_\cdot = 1$ and $\text{d.f.}_\cdot = 4$, respectively. This prior has finite variance and an expectation of 0.5. Combining (1) and (2a), the joint posterior density of P is

$$p(\mu, \mathbf{u}, \sigma^2_\varepsilon, \sigma^2_u \,|\, \mathbf{y}, H) \propto \prod_{i=1}^{n} N\left(y_i \,|\, \mu + u_i, \frac{\sigma^2_\varepsilon}{n_i}\right) \times N(\mathbf{u} \,|\, \mathbf{0}, \mathbf{A}\sigma^2_u)\chi^{-2}(\sigma^2_\varepsilon \,|\, \text{d.f.}_\varepsilon, S_\varepsilon)\chi^{-2}(\sigma^2_u \,|\, \text{d.f.}_u, S_u). \quad (2b)$$

Above, H denotes all hyperparameters indexing the prior distribution. This posterior distribution does not have a closed form; however, samples from the above model can be obtained from a Gibbs sampler, as described, for example, in Sorensen and Gianola (2002). No pedigree data were available for the maize data set; therefore, this model was only in the wheat data set.

Parametric genomic models: For parametric regression, we use the Bayesian LASSO (BL) (Park and Casella 2008), extended by inclusion of an infinitesimal effect, as described in de los Campos *et al.* (2009a). In this model,

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + u_i + \varepsilon_i,$$

and the joint prior density of the model unknowns (upon assigning a flat prior to μ) is

$$p(\mu, \mathbf{u}, \boldsymbol{\beta}, \lambda^2, \sigma^2_\varepsilon, \sigma^2_u \,|\, r, \delta, \text{d.f.}_\varepsilon, S_\varepsilon, \text{d.f.}_u, S_u) \propto N(\mathbf{u} \,|\, \mathbf{0}, \mathbf{A}\sigma^2_u)\left\{\prod_{j=1}^{p} N(\beta_j \,|\, 0, \sigma^2_\varepsilon\tau^2_j)\right\} \times \left\{\prod_{j=1}^{p} \text{Exp}(\tau^2_j \,|\, \lambda^2)\right\} G(\lambda^2 \,|\, r, \delta)\chi^{-2}(\sigma^2_\varepsilon \,|\, \text{d.f.}_\varepsilon, S_\varepsilon) \times \chi^{-2}(\sigma^2_u \,|\, \text{d.f.}_u, S_u). \quad (3a)$$

Above, marker effects are assigned independent Gaussian priors with marker-specific variances $(\sigma^2_\varepsilon\tau^2_j)$. At the next level of the hierarchical model, the τ^2_j's are assigned iid exponential priors $(\text{Exp}[\tau^2_j \,|\, \lambda^2])$. At a deeper level of the hierarchy λ^2 is assigned a Gamma prior with rate (δ) and shape (r), which in this study were set to $\delta = 1 \times 10^{-4}$ and $r = 0.6$, respectively. Finally, independent scaled inverse chi-square priors were assigned to the variance parameters, and the scale and degree of freedom parameters were set to $S_u = S_\varepsilon = 1$ and $\text{d.f.}_\varepsilon = \text{d.f.}_u = 4$, respectively. The above model is referred to as pedigrees plus markers BL (PM)-BL.

The effect of the prior choice for λ^2 in the BL has been addressed in de los Campos *et al.* (2009a). These authors studied the influence of the choice of hyperparameters for λ^2 on inference of several items and concluded that, even when the prior for λ^2 had influence on inferences about this unknown, model goodness-of-fit and estimates of genetic values were robust with respect to the choice of $p(\lambda^2)$. Figure A1 (appendix a) depicts the prior density of λ, $p(\lambda \,|\, r, \delta) = 2G(\lambda^2 | r, \delta)\lambda$, corresponding to the hyperparameter values used in this study; this prior gave a high density over a wide range of values of λ. Also, as shown later, the posterior mean of λ changed between traits and data sets, indicating that Bayesian learning took place.

Combining the assumptions of the likelihood (1) and the prior described in (3a), the joint posterior density is

$$p(\mu, \mathbf{u}, \boldsymbol{\beta}, \lambda^2, \sigma^2_\varepsilon, \sigma^2_u \,|\, \mathbf{y}, H) \propto \left\{\prod_{i=1}^{n} N\left(y_i \,|\, \mu + \sum_{j=1}^{p} x_{ij}\beta_j + u_i, \frac{\sigma^2_\varepsilon}{n_i}\right)\right\} N(\mathbf{u} \,|\, \mathbf{0}, \mathbf{A}\sigma^2_u) \times \left\{\prod_{j=1}^{p} N(\beta_j \,|\, 0, \sigma^2_\varepsilon\tau^2_j)\right\}\left\{\prod_{j=1}^{p} \text{Exp}(\tau^2_j \,|\, \lambda^2)\right\} \times G(\lambda^2 \,|\, r, \delta)\chi^{-2}(\sigma^2_\varepsilon \,|\, \text{d.f.}_\varepsilon, S_\varepsilon)\chi^{-2}(\sigma^2_u \,|\, \text{d.f.}_u, S_u). \quad (3b)$$

This density does not have a closed form; however, samples from the above model can be obtained from a Gibbs sampler, as described in de los Campos *et al.* (2009a). Inferences for the regularization parameter are presented in terms of λ, which were obtained by taking the positive square root of samples from the posterior distribution of λ^2.

A marker-based model, M-BL, can be obtained from (3b) by setting $\mathbf{u} = \mathbf{0}$, which implies $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$.

BLUP using marker genotypes: Prediction of genetic values using BLUP (*e.g.*, Robinson 1991) of marker effects is commonly used in GS (*e.g.*, Meuwissen *et al.* 2001; Bernardo and Yu 2007). We include this method as a reference. BLUP estimates are derived from the model

$$\mathbf{y} = \mu + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$p(\boldsymbol{\varepsilon}, \boldsymbol{\beta} \,|\, \sigma^2_\varepsilon, \sigma^2_\beta) = N(\boldsymbol{\varepsilon} \,|\, \mathbf{0}, \mathbf{D})N(\boldsymbol{\beta} \,|\, \mathbf{0}, \mathbf{I}\sigma^2_\beta),$$

where $\mathbf{D} = \sigma^2_\varepsilon \text{Diag}\{n_1^{-1}, \ldots, n_n^{-1}\}$. From these assumptions, the BLUP estimates of marker effects are

$$\begin{aligned} E(\boldsymbol{\beta} \,|\, \mathbf{y}, \mu, \sigma^2_\varepsilon, \sigma^2_\beta) &= \text{Cov}(\boldsymbol{\beta}, \mathbf{y}')\text{Var}(\mathbf{y})^{-1}(\mathbf{y} - \mathbf{1}\mu) \\ &= \text{Cov}(\boldsymbol{\beta}, \mathbf{1}'\mu + \boldsymbol{\beta}'\mathbf{X}' + \boldsymbol{\varepsilon}')\text{Var}(\mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})^{-1}(\mathbf{y} - \mathbf{1}\mu) \\ &= \sigma^2_\beta\mathbf{X}'\left[\sigma^2_\beta\mathbf{X}\mathbf{X}' + \sigma^2_\varepsilon\mathbf{D}\right]^{-1}(\mathbf{y} - \mathbf{1}\mu). \end{aligned}$$

Computation of BLUPs requires knowledge of $\left\{\mu, \sigma^2_\varepsilon, \sigma^2_\beta\right\}$. 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 kth replicate of the ith genotype $(i = 1, \ldots, n;\ k = 1, \ldots, n_i)$, $g_i \overset{\text{iid}}{\sim} N(0, \sigma^2_g)$ and $\varepsilon_{ik} \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon)$. This model yields estimates of $\left\{\mu, \sigma^2_\varepsilon, \sigma^2_g\right\}$, where $\text{Var}(g_i) = \sigma^2_g$. An estimate of σ^2_β was obtained by plugging the estimate of σ^2_g in $\sigma^2_\beta = \sigma^2_g / \sum_j^p 2\theta_j(1 - \theta_j) \approx \sigma^2_g / 2p\bar{\theta}(1 - \bar{\theta})$ (*e.g.*, Meuwissen *et al.* 2001; VanRaden 2007), where θ_j is the estimated allelic frequency of the jth marker, and $\bar{\theta}$ is the average (across markers) allele frequency, which in our case was estimated from the marker data.

Semiparametric models (RKHS): In RKHS, 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, $K(\mathbf{x}_i, \mathbf{x}_j)$, evaluated in marker genotypes; here $\mathbf{x}_i$ and $\mathbf{x}_j$ are vectors of marker genotype codes for the ith and jth individuals, respectively. The joint prior distribution of $\mathbf{u} = \{u_i\}$, $\mathbf{f} = \{f_i\}$, and the associated variance parameters μ, σ^2_ε, σ^2_u, and σ^2_f, are as follows:

$$\begin{aligned} &p(\mu, \mathbf{u}, \mathbf{f}, \sigma^2_\varepsilon, \sigma^2_u, \sigma^2_f \mid \text{d.f.}_\varepsilon, S_\varepsilon, \text{d.f.}_u, S_u, \text{d.f.}_f, S_f) \\ &\quad \propto N(\mathbf{u} \mid \mathbf{0}, \mathbf{A}\sigma^2_u) N(\mathbf{f} \mid \mathbf{0}, \mathbf{K}\sigma^2_f) \\ &\quad \times \chi^{-2}(\sigma^2_\varepsilon \mid \text{d.f.}_\varepsilon, S_\varepsilon) \chi^{-2}(\sigma^2_u \mid \text{d.f.}_u, S_u) \chi^{-2}(\sigma^2_f \mid \text{d.f.}_f, S_f). \end{aligned} \tag{4a}$$

Above, **K** is a kernel matrix, which is symmetric and positive definite. In this study, the entries of these matrices were the evaluations of a Gaussian kernel, $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\varphi \times d_{ij}\}$, where $d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$ is a squared-Euclidean distance, and φ is a bandwidth parameter that controls how fast the prior correlation drops as lines get farther apart in the sense of d_{ij}. The values of the distance function depend on p, on allele frequencies, and on how related the lines are. The choice of the bandwidth parameter should consider the observed distribution of d_{ij} to avoid situations where **K** is either a matrix full of ones or an identity matrix. In this study we chose $\varphi = 2q^{-1}_{0.5}$, where $q_{0.5}$ is the sample median of d_{ij}. This choice yields $\exp(-2) \approx 0.13$ at the median distance. Higher (lower) prior correlation is assigned to pairs of lines that are closer (farther apart) than $q_{0.5}$, as measured by d_{ij}. Addressing the optimal choice of bandwidth parameter is not within the scope of this study; see de los Campos *et al.* (2010). The scale and degree of freedom parameters of the prior described in (4a) were $S_\varepsilon = S_u = S_f = 1$ and $\text{d.f.}_\varepsilon = \text{d.f.}_u = \text{d.f.}_f = 4$.

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

$$\begin{aligned} &p(\mu, \mathbf{u}, \mathbf{f}, \sigma^2_\varepsilon, \sigma^2_u, \sigma^2_f \mid \mathbf{y}, H) \\ &\quad \propto \left\{ \prod_{i=1}^{n} N\left(y_i \mid \mu + f_i + u_i, \frac{\sigma^2_\varepsilon}{n_i}\right) \right\} \\ &\quad \times N(\mathbf{u} \mid \mathbf{0}, \mathbf{A}\sigma^2_u) N(\mathbf{f} \mid \mathbf{0}, \mathbf{K}\sigma^2_f) \\ &\quad \times \chi^{-2}(\sigma^2_\varepsilon \mid \text{d.f.}_\varepsilon, S_\varepsilon) \chi^{-2}(\sigma^2_u \mid \text{d.f.}_u, S_u) \chi^{-2}(\sigma^2_f \mid \text{d.f.}_f, S_f). \end{aligned} \tag{4b}$$

This density does not possess a closed form; however, samples from this posterior distribution can be obtained using a slightly modified version of the Gibbs sampler that implements the pedigree model in (2a).

In the RKHS regression of (4b), the variances of u_i and f_i can gauge the relative contribution of each of these components to the conditional expectation function. From (4a), $\text{Var}(u_i) = a(i,i)\sigma^2_u$, where $a(i,i)$ is the ith diagonal element of matrix **A**, and $\text{Var}(f_i) = K(\mathbf{x}_i, \mathbf{x}_i)\sigma^2_f$. Here, $K(\mathbf{x}_i, \mathbf{x}_i)$ is a standardized kernel, with $K(\mathbf{x}_i, \mathbf{x}_i) = 1$. This does not occur in $a(i,i)$; here $a(i,i) = 1 + F_i$, where F_i is the coefficient of inbreeding of the ith individual. In the wheat population, the average value of $a(i,i)$ was 1.98.

As with parametric methods, a marker-based model, M-RKHS, can be obtained as a particular case of (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 data set, and inferences for each fit were based on 30,000 samples (obtained after discarding 5000 samples as burn-in). Convergence was checked by inspecting trace plots of variance parameters.

Cross-validation: Prediction of performance of lines whose phenotypes are yet to be observed is a central problem in plant breeding. Such prediction can be used, for example, to decide which of the newly generated lines will be evaluated in field trials. Cross-validation (CV) methods were used to evaluate the ability of a model to predict future outcomes. To this end, data were divided into 10 folds; this was done by using an index variable, $I_i \in \{1, \ldots, 10\}$, $i = 1, \ldots, n$, that randomly assigns observations to 10 disjoint folds, $F_j = \{i : I_i = j\}$, $j = 1, \ldots, 10$. CV predictions of the observations in the first fold, $F_1 = \{i : I_i = 1\}$, are obtained by omitting phenotypic data on all lines in the first fold. This yields CV predictions of lines in the first fold, that is, $\{\hat{y}_i : I_i = 1\}$. Repeating this exercise for the second, third, ..., 10th 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.

Principal component analysis of estimated marker effects: Parametric models such as the BL yield estimates of marker effects, which, in our case, are environment specific. These estimates can be used to assess and visualize genetic effect × environment interaction. Biplots from principal component analysis of the matrix of estimated marker effects in each trait–environment combination were obtained. The methodology is briefly explained in appendix b. Use of biplots to assess genetic effect × environment interaction is further described in Cornelius *et al.* (2001).

RESULTS

This section begins by presenting estimates of variance parameters and of the regularization parameters of BL and RKHS that were obtained when models were fitted using all available records (*i.e.*, full data analysis). Next, results from the principal components analysis of estimated marker effects (also obtained from the full data analysis) for the W-GY data set are given (results for the maize data set are provided in appendix c). Subsequently, estimates of measures of predictive ability obtained from cross-validation are presented.

Variance and regularization parameters: Tables 1 and 2 give the estimates of posterior means of variance parameters and of λ in the BL. 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 σ^2_ε gives an indication of the fraction of the phenotypic variance that can be attributable to model residuals. In the GY-W data set (Table 1), RKHS models fitted data markedly better (smaller σ^2_ε) than P, M-BL, or PM-BL. Model M-BL had a posterior mean of residual variance that was either similar to or slightly larger than that of P, while PM-BL fitted the data better than P. Results from the maize data sets (Table 2) were mixed: M-BL fitted the data much better than M-RKHS for FFL and MFL, regardless of environmental conditions, but the opposite was observed (*i.e.*, M-RKHS fitted data better than M-BL) for ASI and GY (Table 2).

For the W-GY data set, the posterior means of σ^2_u in PM-BL and PM-RKHS were smaller than that obtained in P (Table 1). This indicates that the inclusion of markers reduces the relative contribution of the regression on

the pedigree, u_i. In PM-RKHS, the ratio $\sigma_f^2/a(i,i)\sigma_u^2$, evaluated at $a(i,i) = 1.98$ and at the posterior mean of σ_f^2 and σ_u^2, was always >2 (Table 1), indicating that in PM-RKHS models, the regression on the markers made a much more important contribution to the conditional expectation than the regression on the pedigree.

Marker effects: Estimated marker effects obtained from PM-BL are provided in supporting information, Table S1, Table S2, and Table S3.

The multivariate analysis of estimated marker effects for the W-GY data set indicated that the first two principal components explained 74% of the total variability in estimated marker effects (Figure 1). Sample correlations between phenotypes in the four environments (E) 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. The correlation patterns of estimated marker effects were similar, but the strength of the association was slightly weaker. For instance, the correlations between estimates of marker effects were 0.633 (E2–E3), 0.388 (E2–E4), and 0.384 (E3–E4). Correlations between E1 and the other environments were low and negative for phenotypic and estimated marker effect data.

The variance of estimated marker effects was slightly smaller in E4; this can be inferred by the length of the corresponding vector in Figure 1. The vast majority of the estimated effects are located around the center of Figure 1 (*i.e.*, estimated effects were small, in absolute value), which reflects shrinkage of the BL model. However, some markers had estimated effects that were large in absolute value; some of those markers are identified by their name in Figure 1, and the estimated effects are given in Table S1. An approximation to the estimated effect of the presence of a marker in GY for a given environment can be obtained by orthogonal projection of the marker effect displayed in Figure 1 on the vector of the corresponding environment. To illustrate this, consider E1, where the presence of markers wPt.9256, wPt.6047, and wPt.3904 is expected to increase GY (Figure 1); in contrast, the presence of markers wPt.3462, wPt.3922, and wPt.4988 (located in the opposite direction of E1) is expected to reduce GY.

The multivariate analysis of estimated marker effects allows identifying which markers contribute to positive/negative genetic correlation between environments. Markers whose presence is expected to increase or decrease GY across environments can be viewed as

TABLE 1

Estimates of posterior mean of parameters σ_ε^2, σ_u^2, σ_f^2, and λ from the full-data analysis of grain yield (GY) of 599 wheat lines genotyped with 1279 molecular markers

Trait–environment	Model[a]	Parameter			
		σ_ε^{2b}	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

Five models were fitted to each trait (GY) and environment (E1, E2, E3, and E4) combination.

[a] Models were pedigree model (P), molecular marker model using reproducing kernel Hilbert space (M-RKHS) regression, pedigree plus molecular marker model using reproducing kernel Hilbert space regression (PM-RKHS), molecular marker regression model using the Bayesian LASSO (M-BL), and pedigree plus the molecular marker model regression using the Bayesian LASSO (PM-BL). 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.

[b] Phenotypes were standardized to a unit variance within environment.

TABLE 2

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 1148 markers, and grain yield (GY) of 264 genotypes and 1135 markers

		Parameter		
Trait–environment	Model[a]	$\sigma^{2\,b}_\varepsilon$	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

Two models were fitted to each of the trait (FFL, MFL, ASI, and GY) and environment (SS, severe stress; WW, well watered) combinations.

[a] Models were molecular marker (M) using reproducing kernel Hilbert space (M-RKHS) regression and molecular marker (M) regression model using the Bayesian LASSO (M-BL). 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.

[b] Phenotypes were standardized to a unit variance within trait and environment.

contributing to positive genetic correlations in GY between environments. Examples of this group are markers 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 markers act in an "antagonistic" fashion; that is, the presence of a marker increases (decreases) GY in some environments and decreases (increases) GY in others.

Results from the multivariate analysis of marker effects in the maize data sets (M-F and M-GY) were similar to those observed in the wheat data set in regard to the following: (1) the first two principal components explained a large proportion (85.8%) of the observed variability of estimated marker effects; (2) due to shrinkage, most estimated marker effects clustered around zero; and (3) although the overall correlation patterns between estimated marker effects reflected the type of association observed between phenotypes, it was possible to identify subsets of markers that contributed to positive genetic correlation and others that induced negative genetic associations. A detailed discussion of these results is given in APPENDIX C.

Predictive ability: Tables 3 and 4 show the estimated correlations between phenotypic outcomes and CV predictions for W-GY, M-F, and M-GY data sets. Overall, the values of these correlations, especially those obtained with BL or RKHS methods, were large for all models, data sets, and traits, indicating that genomic selection can be effective for predicting the performance of lines with yet-to-be observed phenotypes. Predictive ability was different between models and data sets: for W-GY correlations ranged from 0.355 to 0.608, for M-F correlations varied from 0.464 to 0.79, and for M-GY they ranged from 0.415 to 0.514.

Wheat data set: In the W-GY, correlations ranged from 0.355 (BLUP in E3) to 0.608 (PM-RKHS in E1) (Table 3), and relative to the P model, the PM-RKHS model produced the highest relative gain in CV correlation in three of four environments. BLUP was outperformed by BL and RKHS methods across environments. In 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 3). In general, RKHS outperformed BL both in M and PM, and BLUP outperformed P models in three of four environments (all but E3); however, as stated, BLUP was outperformed by BL and RKHS.

Maize flowering: In the M-F, correlations ranged from 0.464 (BLUP for MFL-SS) to 0.790 (M-BL for MFL-WW) (Table 4). For these traits, BLUP was systematically outperformed by BL and RKHS. Also for these traits,

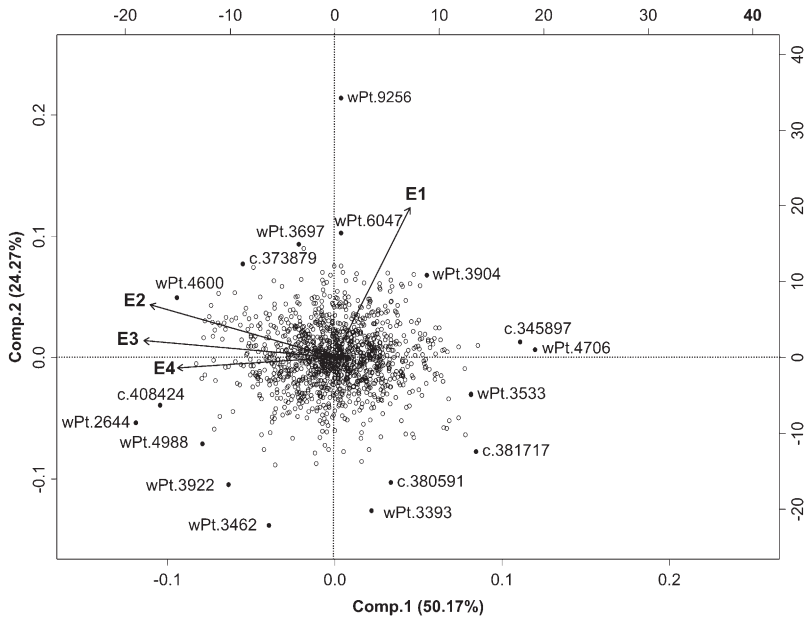

FIGURE 1.—Biplot of the first two principal components (Comp. 1 and Comp. 2) of estimates of marker effects on grain yield (GY) in wheat evaluated in four environments (E1–E4). Marker effects were obtained from a full-data analysis and using a pedigree plus marker model (PM-BL). Only the effects of 17 markers that are located farthest from the center of the biplot were identified with their corresponding marker's name (solid circles).

M-BL yielded better predictions than M-RKHS, with relatively high correlation values that ranged 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 4).

Maize grain yield: Predictive correlations in M-GY (Table 4) were smaller than those obtained in flowering traits, and the differences between methods were not clear as in the M-F data set. 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 models were not large.

DISCUSSION

Several simulation studies (BERNARDO and YU 2007; WONG and BERNARDO 2008; MAYOR and BERNARDO 2009; ZHONG *et al.* 2009) have reported important gains in genetic progress associated with the use of GS in plant breeding. Recently, HEFFNER *et al.* (2009) concluded that the high correlation between true breeding values and the genomic estimated breeding values found in several simulation studies is sufficient for considering selection based on molecular markers alone; however, evaluation of these methods with real plant data is still very limited.

Empirical evaluation of GS: The results of this study 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. These findings are in agreement with simulation-based studies such as those mentioned above and with empirical evidence reported in animal breeding (*e.g.*, GONZALEZ-RECIO *et al.* 2008; VANRADEN *et al.* 2008; HAYES *et al.* 2009; WEIGEL *et al.* 2009).

Evaluation of predictive ability indicated that models using marker and pedigree data jointly (PM) outperformed pedigree models (P) across traits and environments, regardless of the choice of model (BL, RKHS). These results are consistent with those reported by CROSSA *et al.* (2010), who evaluated P, M, and PM models using the BL and RKHS for grain yield in wheat ($n = 170$) and several disease traits in maize.

Despite the gains in predictive ability obtained with PM models, our results suggest that there is room for improving predictive ability even further. To illustrate this, and as an exercise, let us assume that the model $y_i = g_i + \varepsilon_i$ holds, and consider as the best (unlikely) scenario that CV predictions, $\hat{g}_{i,\mathrm{CV}}$, are such that $\hat{g}_{i,\mathrm{CV}} = g_i$. If so, the maximum attainable correlation is $\mathrm{Cor}(g_i, y_i) = (\sigma_g^2 + \sigma_\varepsilon^2)^{-(1/2)}\sigma_g = h$, 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. This will hold if only one replicate is available; for data involving repeated measures, as was the case in this study, the maximum correlation is $\mathrm{Cor}(g_i, y_i) = (\sigma_g^2 + n_i^{-1}\sigma_\varepsilon^2)^{-(1/2)}\ \sigma_g > h$. CV correlations in this study ranged from 0.40 to 0.79; these values are well below the theoretical maxima given the heritability of the traits and the number of replicates available. We therefore conclude that larger gains in predictive ability can be expected (1) when more markers are available or (2) by improving upon the methods used to implement GS.

Choice of model: There are different ways of incorporating markers into models for GS. Here we evaluated the BL, BLUP, and RKHS methods. BLUP

TABLE 3

Cross-validation (CV) correlation between predicted and observed phenotypes, obtained in a 10-fold CV conducted for grain yield (GY) records of 599 wheat lines genotyped with 1279 molecular markers

	Model[a]					
Trait–environment	P	M-RKHS	PM-RKHS	M-BL	PM-BL	BLUP[b]
			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

Six models were fitted to GY measured in four environments (E1, E2, E3, and E4).

[a] Models were pedigree model (P), molecular marker model using reproducing kernel Hilbert space (M-RKHS) regression, pedigree plus molecular marker model using reproducing kernel Hilbert space regression (PM-RKHS), molecular marker regression model using the Bayesian LASSO (M-BL), pedigree plus molecular marker model regression using the Bayesian LASSO (PM-BL), and best linear unbiased prediction (BLUP) using marker genotypes.

[b] Values of genetic variances used to compute BLUP ranged from 0.8065 to 0.9141.

and BL use parametric regression on marker covariates, whereas RKHS is a semiparametric method. In general, BL outperformed BLUP, which may be attributed to at least two reasons: (1) similar to other methods for GS such as methods Bayes A and Bayes B of Meuwissen *et al.* (2001), BL performs marker-specific shrinkage of effects, whereas BLUP penalizes all marker effects equally; and (2) in BL, variance parameters and marker effects are inferred jointly, whereas BLUP typically involves two steps (a first one in which variance parameters are inferred and a second one in which marker effects are estimated).

The comparison between BL and RKHS yielded mixed results; this finding is in agreement with those of Zhong *et al.* (2009), who evaluated different models in different scenarios (mating systems) and did not find one method that performed best across scenarios. For grain yield and anthesis-silking interval, RKHS methods performed either similarly or better than the BL; however, for female and male flowering traits in maize, BL outperformed RKHS markedly. The BL is an additive model, whereas RKHS may be able to capture complex epistatic interactions better (*e.g.*, Gianola and van Kaam 2008). Therefore, one could expect the BL to perform well in traits where additive effects play a central role and RKHS to perform better in traits where epitasis is more relevant. Buckler *et al.* (2009) provide evidence suggesting that female and male flowering traits in maize are, for the most part, additive traits. The good performance of the BL observed in this study for those traits is consistent with this finding.

Marker *vs.* pedigree plus marker models: In general, PM models in W-GY had a slight but consistent superiority in all four environments for predictive ability as compared to the M model; this is in agreement with previous findings (*e.g.*, de los Campos *et al.* 2009a). The advantage of considering pedigree and markers jointly is small

TABLE 4

Cross-validation (CV) correlation between predicted and observed phenotypes, obtained in a 10-fold CV conducted for female flowering (FFL), male flowering (MFL), the MFL to FFL interval (ASI) of 284 maize lines genotyped for 1148 markers, and grain yield (GY) of 264 maize lines genotyped for 1135 markers

	Model[a]		
Trait–environment	M-RKHS	M-BL	BLUP[b]
MFL-WW	0.607	0.790	—[c]
MFL-SS	0.674	0.778	0.464
FFL-WW	0.588	0.781	—[c]
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

Three models were fitted to each trait (FFL, MFL, ASI, and GY) and environment (SS, severe drought stress; WW, well watered) combination.

[a] Models were molecular marker (M) using reproducing kernel Hilbert space (M-RKHS) regression, molecular marker (M) regression model using the Bayesian LASSO (M-BL), and best linear unbiased predictor (BLUP) using marker genotypes.

[b] Values of genetic variances used to compute BLUP ranged from 0.000 to 0.319 for flowering, and from 0.017 to 0.206 for grain yield.

[c] BLUPs were not computed because the estimated genetic variances were negligible.

because there is some redundancy between regression on the pedigree and regression on markers (*e.g.*, Habier *et al.* 2009). It is reasonable to expect that as the number of molecular markers increases, the relative contribution of pedigree information will decrease.

Assessment of genetic effect × environment interaction with estimates of marker effects: Parametric methods such as M-BL, PM-BL, or BLUP provide estimates of "marker effects" that may be used to gain a better understanding of the underlying architecture of the traits. The results obtained here with W-GY are consistent with those reported by Crossa *et al.* (2007) and indicate that markers such as wPt.6047, wPt.3393, wPt3462, and wPt.3904 (located in chromosome 3B, the long arm of chromosome 7A, chromosome 1A, and the short arm of chromosome 1A, respectively) are indeed associated with GY in wheat.

Estimates of marker effects can be also used to gain insights on the sources of genetic effect × environment interaction. Here, we used principal component analysis of estimates of marker effects as a way of assessing sources of marker effect × environment interaction. Overall, the correlation patterns of estimated marker effects were similar to those observed at the phenotypic level; however, in all trait–environment combinations it was possible to detect markers that made contributions to positive or negative genetic correlation. For example, for the M-F data set, results indicate important molecular marker effect × environment interactions, which translate into genotype × environment interaction. In this respect, our results are different from those of Buckler *et al.* (2009), who reported low levels of genotype × environment interaction for the same traits.

Conclusion: Results of this study showed that models including markers or markers and pedigrees yield relatively high correlations between predicted and observed phenotypic outcomes. The superiority of models using markers or markers and pedigree was clear regardless of the choice of method (BL, RKHS). Moreover, we did not find a method (BL or RKHS) that was consistently superior across environments and traits. Differences in the underlying genetic architecture of the traits may well explain these results.

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

The Bayesian LASSO was fitted using the BLR package which is available in R (R Development Core Team 2010; G. de los Campos and P. Pérez) and described in Pérez *et al.* (2010). The wheat and maize experimental data, and other computer programs written in R for fitting the RKHS models using the Gibbs sampler described in this article, are available in File S1.

This article benefited from valuable comments from two associate editors and two anonymous reviewers. The maize data set used in this study comes from the Drought Tolerance Maize for Africa project financed by the Bill and Melinda Gates Foundation. We thank the numerous cooperators in national agricultural research institutes who carried out the maize trials in Africa and the Elite Spring Wheat Yield Trials and provided the phenotypic data analyzed in this article. We also thank the International Nursery and Seed Distribution Units in the International Maize and Wheat Improvement Center (CIMMYT, Mexico), for preparing and distributing the seed and digitalizing the data. Gustavo de los Campos and Daniel Gianola acknowledge support by the Wisconsin Agriculture Experiment Station and from grant DMS-11044371 made by the Division of Mathematical Sciences of the National Science Foundation.

LITERATURE CITED

Bernardo, R., and J. Yu, 2007 Prospects for genome-wide selection for quantitative traits in maize. Crop Sci. **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 Sci. **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. Statistical Consulting Section, American Statistical Association, Joint Statistical Meetings, August 7, Atlanta, GA.

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 Sci. **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.

Crossa, J., P. Perez, G. de los Campos, G. Mahuku, S. Dreisigacker *et al.*, 2010 Genomic selection and prediction in plant breeding. *Quantitative Genetics, Genomics, and Plant Breeding*, Ed. 2, edited by M. S. Kang. CABI Publishing, New York (in press) http://genomics.cimmyt.org/.

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.

de los Campos, G., D. Gianola, G. J. M. Rosa, K. A. Wiegel and J. Crossa, 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. (in press). http://genomics.cimmyt.org/.

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., and J. B. C. H. M. van Kaam, 2008 Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics **178:** 2289–2303.

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

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. Dairy Sci. **92:** 433–443.
Heffner, E. L., M. R. Sorrels and J.-L. Jannink, 2009 Genomic selection for crop improvement. Crop Sci. **49:** 1–12.
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.
Jannink, J.-L., A. J. Lorenz and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics. **9**(2): 166–177.
Mayor, P. J., and R. Bernardo, 2009 Genome-wide selection and marker-assisted recurrent selection in double haploid versus F_2 population. Crop Sci. **49:** 1719–1725.
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 Physiol. **139:** 637–642.
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.
Oakey, H., A. Verbyla, W. Pitchford, B. Cullis and H. Kuchel, 2006 Joint modeling of additive and non-additive genetic line effects in single field trials. Theor. Appl. Genet. **113:** 809–819.
Park, T., and G. Casella, 2008 The Bayesian LASSO. J. Am. Stat. Assoc. **103:** 681–686.
Pérez, P., G. de los Campos, J. Crossa and D. Gianola, 2010 Genomic-enabled prediction based on molecular markers and pedigree using the BLR package in R. Plant Genome (in press). http://genomics.cimmyt.org/.
Piepho, H. P., 2009 Ridge regression and extensions for genome-wide selection in maize. Crop Sci. **49:** 1165–1176.
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.
Robinson, G. K., 1991 That BLUP is a good thing: the estimation of random effects. Stat. Sci. **6**(1): 15–51.
R Development Core Team, 2010 *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna.http://www.R-project.org.
Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics.* Springer-Verlag, New York.
Tibshirani, R., 1996 Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. B **58:** 267–288.
vanRaden, P. M., 2007 Genomic measures of relationship and inbreeding. *Interbull Annual Meeting Proceedings, Interbull Bulletin*, Vol. 37, pp. 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. Dairy Sci. **92:** 16–24.
Weigel, K. A., G. de los Campos, O. González-Recio, H. Naya, X. L. Wu *et al.*, 2009 Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. **92:** 5248–5257.
Wong, C., and R. Bernardo, 2008 Genome-wide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. **116:** 815–824.
Zhong, S., J. C. M. Dekker, R. L. Fernando and J.-L. Jannink, 2009 Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics **182:** 355–364.

Communicating editor: M. Kirst

APPENDIX A:

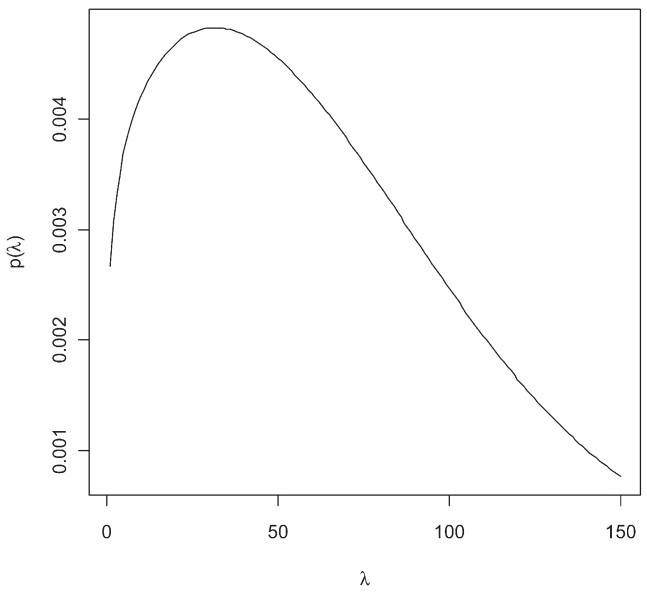

Figure A1.—Prior density of the regularization parameter, $p(\lambda)$, used to fit the Bayesian LASSO.

APPENDIX B: MULTIVARIATE ANALYSIS OF ESTIMATED MARKER EFFECTS

Consider a matrix of estimated molecular marker effects, $\hat{\mathbf{B}}_{p\times q} = \left[\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_q\right] = \left\{\hat{\beta}_{jk}\right\}$, whose columns, $\hat{\boldsymbol{\beta}}_k, k = 1, \ldots, q$, 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 $\mathbf{U}_{p\times q} = \left[\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q\right] = \left\{\alpha_{jk}\right\}$ and $\mathbf{V}_{q\times q} = \left[\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q\right] = \{\gamma_{kl}\}$ 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 nonnull entries are the singular values of $\hat{\mathbf{B}}$; that is, $\mathbf{D} = \text{Diag}\{\lambda_k\}$.

The biplot is constructed using the first two principal components axis of $\hat{\mathbf{B}}$ ($\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$). 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 to the environmental vectors had negative effects on those environments. The cosine of the angle between the

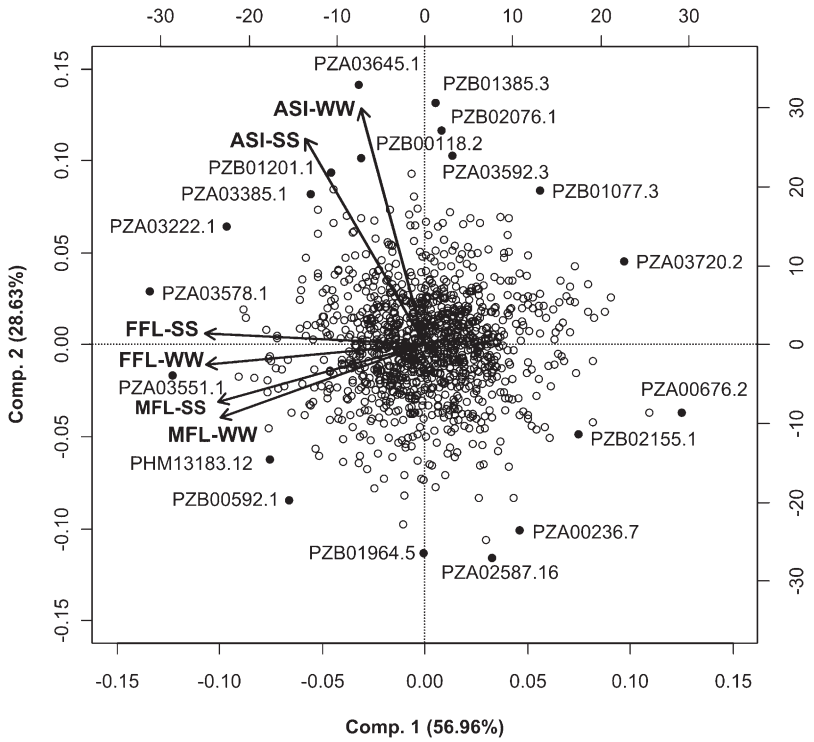

FIGURE C1.—Biplot of the first two principal components (Comp. 1 and Comp. 2) of estimates of marker effects for female flowering (FFL), male flowering (MFL), and the FFL-MFL interval (ASI) evaluated under well-watered (WW) and drought-stress (SS) conditions. Estimates of marker effects were obtained from a full-data analysis and using a pedigree plus marker model (PM-BL). Only the effects of the 19 markers that are located farthest from the center of the biplot were identified with their corresponding marker's name (solid circles).

vectors representing a pair of 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 marker effects) on estimated effects of the markers in the six trait–environment combinations (MFL-SS, MFL-WW, FFL-SS, FFL-WW, ASI-SS, and ASI-WW) of the M-F data set obtained from the BL model is depicted in Figure C1. Clearly the two groups of trait–environment combinations are dominated more by the trait (ASI *vs.* FFL and MFL) and less by the environmental condition (SS and WW). Phenotypic outcomes and estimates of marker effects for ASI showed relatively small correlations with those of FFL and MFL. Phenotypic correlations between MFL in WW and SS, ASI in WW and SS, and FFL in SS and WW were positive and high, ranging from 0.686 to 0.728. Correlations ASI-MFL and ASI-FFL at the different water regimes (SS and WW) ranged from −0.123 to 0.446.

Interpretation of the estimated marker effect on these traits should be different from that for grain yield. For FFL and MFL, the favorable allele is the one whose estimated effect is negative (*i.e.*, it decreases FFL and MFL), whereas for ASI, selection seeks to set this trait as close to zero as possible. Alleles coded as 1 of markers whose estimated effects are located on the left side and in the top left corner of Figure C1 (*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–environment combinations), whereas those markers located on the opposite side of the biplot (bottom right corner) (*i.e.*, PZA02587.16, PZA00236.7, PZB0255.1, and PZA00676.2) decrease the value of FFL, MFL, and ASI. Those markers 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 Figure C1 shows markers that had very contrasting effects across environments. For example, the minor alleles of markers whose estimated effects are located in the top right corner of the biplot (PZA03592.3, PZB01077.3, and PZB02076.1) increase the anthesis-silking interval under drought and well-watered conditions, but decrease days to male and female flowering. In contrast, the minor alleles of markers whose estimated effects are located in the opposite quadrant of the biplot (bottom left corner) (PZB00592.1, PHM13183.12, and PZB01964.5) showed a complete rank reversal with respect to the effects of markers 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.

The estimated effects used to perform the multivariate analysis included in this section are provided in Table S2.

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 data set, no principal component analysis was performed. The phenotypic correlations between GY-WW and GY-SS (0.260), as well as the correlations between the estimated marker effects for grain yield (0.251), were low. Also, none of the 10 markers with the largest/smallest estimated effects in GY-WW was among those with the largest/smallest effects under GY-SS conditions. This indicates important context-dependent effects due to genotype × environment interaction. Estimates of marker effects for GY-WW and GY-SS are provided in Table S3.

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.118521/DC1

Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers

José Crossa, Gustavo de los Campos, Paulino Pérez, Daniel Gianola, Juan Burgueño, José Luis Araus, Dan Makumbi, Ravi P. Singh, Susanne Dreisigacker, Jianbing Yan, Vivi Arief, Marianne Banziger and Hans-Joachim Braun

DOI: 10.1534/genetics.110.118521

FILE S1

The wheat and maize experimental data, and other computer programs written in R for fitting the RKHS models using the Gibbs sampler

File S1 is available for download as a compressed file at http://www.genetics.org/cgi/content/full/genetics.110.118521/DC1.

TABLE S1

Effect of 1,279 DArT markers in three environments (E1-E4) for the WHEAT GRAIN YIELD DATA

TABLE S2

Effect of 1,148 SNP markers in six trait-environment combinations for the MAIZE FLOWERING DATA

TABLE S3

Effect of 1,135 SNP markers in two environments (SS and WW) for the MAIZE GRAIN YIELD DATA

Tables S1-S3 are available for download as Excel files at http://www.genetics.org/cgi/content/full/genetics.110.118521/DC1.