Genomic Prediction of Breeding Values when Modeling Genotype \( \times \) Environment Interaction using Pedigree and Dense Molecular Markers

Juan Burgueño, Gustavo de los Campos, Kent Weigel, and José Crossa*

ABSTRACT
Genomic selection (GS) has become an important aid in plant and animal breeding. Multi-environment (multitrait) models allow borrowing of information across environments (traits), which could enhance prediction accuracy. This study presents multi-environment (multitrait) models for GS and compares the predictive accuracy of these models with: (i) multi-environment analysis without pedigree and marker information, and (ii) multi-environment pedigree or/and marker-based models. A statistical framework for incorporating pedigree and molecular marker information in models for multi-environment data is described and applied to data that originate from wheat (Triticum aestivum L.) multi-environment trials. Two prediction problems relevant to plant breeders are considered: (CV1) predicting the performance of untested genotypes (“newly” developed lines), and (CV2) predicting the performance of genotypes that have been evaluated in some environments but not in others. Results confirmed the superiority of models using both marker and pedigree information over those based on pedigree information only. Models with pedigree and/or markers had better predictive accuracy than simple linear mixed models that do not include either of these two sources of information. We concluded that the evaluation of such trials can benefit greatly from using multi-environment GS models.

doi: 10.2135/cropsci2011.06.0299
Freely available online through the author-supported open-access option.
© Crop Science Society of America | 5885 Guilford Rd., Madison, WI 53711 USA
All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.
kinship coefficient between individuals $i$ and $i'$, describes a priori, the expected correlations between the genetic values of individuals $i$ and $i'$, given the information conveyed by the pedigree and under the assumption of an additive infinitesimal model. Expected and realized degrees of genetic similarity differ due to factors such as common ancestry not accounted for by the pedigree data and, more importantly, Mendelian segregation. In additive models and in absence of inbreeding and assortative mating, Mendelian segregation can explain one-half of the additive genetic variance. Dense molecular markers can be used to estimate the realized genetic similarity between individuals (e.g., Ritland, 1996; Lynch and Ritland, 1999; VanRaden, 2007; Habier et al., 2007; de los Campos et al., 2010), and such information can be incorporated into multienvironment models in the same way that pedigree-derived genetic relationships are incorporated into those models.

The availability of thousands of genome-wide molecular markers has made it possible to use genomic selection (GS) for the prediction of genetic values (Meuwissen et al., 2001) in plants (Bernardo and Yu, 2007; de los Campos et al., 2009, 2010; Crossa et al., 2010, 2011; Pérez et al., 2010) and animals (Hayes and Goddard, 2010; de los Campos et al., 2009). Most GS applications use a single-environment model. However, this approach cannot exploit information across environments (or traits), which may limit the predictive power of models for GS.

The importance of exploiting across-environment information has been demonstrated in plant breeding. For example, by using data from multienvironment trials and mixed linear models, Burgueño et al. (2011) showed that, relative to single-environment mixed models, the modeling of GE with a factor analytic (FA) structure can increase predictive power by up to 6%. This study was performed in a context in which pedigree information was not available. Similar results were found by So and Edwards (2011), who compared predictability of the FA structure with the compound symmetric model. On the other hand, empirical evidence with plant breeding data shows that GS can outperform the predictive power of pedigree-based methods by a sizable amount (de los Campos et al., 2009, 2010; Crossa et al., 2010, 2011; Pérez et al., 2010), using single-environment models that did not exploit information across environments (or traits). This body of evidence, together with the relevance of GE in plant breeding, suggests that the development of multienvironment models for GS can boost the predictive power of genomic prediction and, hence, the rate of genetic gain that can be attained with GS.

In this article, we incorporate into a single model elements for modeling GE using FA along with genetic relationship information based on pedigrees or markers for the purpose of genomic prediction. To our knowledge, this is the first demonstration of the combination of complex model components for both genetic relationships and GE interaction. The main objective of this study was to develop multienvironment models for GS and to compare their predictive power with: (i) multienvironment analysis without pedigree and marker information, and (ii) multienvironment pedigree or/and marker-based models. Two different prediction problems that are relevant to plant breeders were considered: (i) predicting performance of newly developed genotypes (i.e., genotypes for which no phenotypic records were available); and (ii) prediction of performance of genotypes across environments, when some genotypes were evaluated in some environments but not in others. A statistical framework for incorporating molecular marker information in models for multienvironment data is described and applied in a plant breeding context using a multienvironment wheat (*Triticum aestivum* L.) trial.

### MATERIALS AND METHODS

We begin this section by introducing a conceptual framework that allows modeling pedigree and molecular marker data using mixed models. Subsequently, we apply this framework to data from CIMMYT’s Global Wheat Program.

#### A Multienvironment Mixed Model for Regression using Molecular Marker and Pedigree Data

In a multienvironment linear mixed model, the phenotypic outcomes of $g$ individuals ($i = 1, 2,...,g$) evaluated in $J$ environments ($j = 1, 2,...,J$) are expressed as follows:

\[
y = X\beta + Z\gamma + \varepsilon
\]

where $y_i$ is the vector of the response variable of phenotypic records (or means) collected in the $j$th environment, and $X$ and $Z$ are incidence matrices vectors of systematic effects, $\beta$, and random genetic effects, $\gamma$, in the $j$th environment (with elements $g_{ij}$ for the $ij$th genotype). In reduced matrix notation, the covariance matrices for the error term $\varepsilon$ are expressed as follows:

\[
y = X\beta + Z\gamma + \varepsilon
\]

where $y = \begin{pmatrix} y_1' & y_2' & \ldots & y_J' \end{pmatrix}'$, $\beta = \begin{pmatrix} \beta_1' & \beta_2' & \ldots & \beta_J' \end{pmatrix}'$, $\gamma = \begin{pmatrix} \gamma_1' & \gamma_2' & \ldots & \gamma_J' \end{pmatrix}'$, and $\varepsilon = \begin{pmatrix} \varepsilon_1' & \varepsilon_2' & \ldots & \varepsilon_J' \end{pmatrix}'$.

In a standard multienvironment linear mixed model, vectors containing the random effects entering on the right-hand side of Eq. [2] are assumed to follow a multivariate normal density, centered at zero, and with covariance structure $\operatorname{Cov} (\varepsilon) = I_J \otimes \mathbf{R}_e$, and $\mathbf{R}_e = 0$. Where $I_J$ is the identity matrix of size $g$, $\mathbf{G}_0$ is a $J \times J$ covariance matrix of the genetic effect of genotypes in environments, $\otimes$ denotes the Kronecker product of matrices, and $\mathbf{G}_0 = \begin{pmatrix} a(i,j) \end{pmatrix}$ is a $g \times g$ numerator relationship matrix, whose entries are twice the kinship coefficients between pairs of lines. The $\operatorname{Cov} (\varepsilon) = \mathbf{G}_0 \otimes \mathbf{A}$ is usually represented as
where the $j$th diagonal element of the $J \times J$ matrix $G_j$ is the additive genetic variance $\sigma^2_a$ within the $j$th environment, and the $j'j$th element is the additive genetic covariance $\rho_{j'j} \sigma_s \sigma_s$ between sites $j$ and $j'$; thus, $\rho_{j'j}$ is the correlation of the additive genetic effects between sites $j$ and $j'$.

The $R_h = \{\text{Cov} \{e_i, e_j\}\}$ is a $J \times J$ covariance matrix of (within-genotype, across-environments) model residuals that could contain different variance–covariance structures for groups of observations and for field experiments in plant breeding evaluation, where spatial correlation for modeling the plot-to-plot variability is important. Model [2] can be extended to accommodate heterogeneous residual variances by replacing $I_g$ in $I_g \otimes R_h$ with a diagonal matrix, $D = \text{Diag}\{\sigma_i^2\}$.

With these assumptions, and following properties of the multivariate normal distribution (MVN), the marginal density of the data is multivariate normal

$$
[y|3, R_h, G_j] \sim \text{MVN}\left[0, Z'G_j \otimes A[Z' + I_g \otimes R_h]\right] \tag{3}
$$

Estimation of model unknowns in the above model can be performed using Bayesian, Likelihood, or Restricted Maximum Likelihood (REML) methods. A standard approach is to first estimate covariance matrices using REML and then obtain estimates of fixed effects and predictions of genetic values by solving the mixed-model equations associated with Eq. [3] with codispersion parameters replaced by REML estimates. The solution to the mixed-model equations are the empirical Best Linear Unbiased Estimates (BLUE) of fixed effects and the empirical Best Linear Unbiased Prediction (BLUP) of random effects (e.g., Henderson, 1975).

In this study, we performed a two-stage analysis consisting of: first, computing the mean of the genotypes in each mega-environment and standardized the data; and second, fitting the different models with the standardized data. For cases where heterogeneity of within-site error variances is evident, a weighted analysis in the models with the standardized data. For cases where heterogeneity of within-site error variances is evident, a weighted analysis in the models with the standardized data. For cases where heterogeneity of within-site error variances is evident, a weighted analysis in the models with the standardized data.

**Pedigree and Marker-based Prediction**

In the models described by [1–3], $A = \{a(i, i')\}$ represents a $g \times g$ matrix of additive relationships. Commonly, these are derived from a pedigree (hereinafter denoted as $A_p$). Alternatively, this matrix can be derived from molecular marker information (hereinafter denoted as $A_m$). Unlike standard pedigree-derived genetic relationships, marker-derived estimates of relationships can account for common ancestry not considered in the pedigree, and more importantly, it can account for departures of realized genetic relationships from their expected values given the pedigree due to Mendelian segregation.

There are multiple ways of mapping markers to estimates of genetic relationships (e.g., Ritland, 1996; Lynch and Ritland, 1999; Van Raden, 2007; Habier et al., 2007; de los Campos et al., 2010), and none is considered to be superior (Makowsky et al., 2011). Here, we consider $A_m = MM'$, where $M = \{m_{ij}\}$ is a matrix containing centered and standardized genotypes ($i = 1, \ldots, g$) at different markers ($l = 1, \ldots, L$), that is, $m_{ij} = \frac{m_{ij} - \bar{m}_j}{\sqrt{p_i(1 - p_i)}}$, where $m_{ij}$ counts the number of copies of the minor-frequency allele carried by the $i$th line at the $l$th locus, and $p_i$ is the estimated minor-frequency allele of the $l$th marker. In Appendix A we show the equivalence of using $A_m = MM'$ and a certain class of regression with marker covariates.

Finally, pedigree and marker information can be combined in a single model by simply extending Eq. [2] to a model with two random effects, one of which, $g_j - N(0, G_p \otimes A_p)$, represents a regression on a pedigree, and the other, a regression on markers, $g_j - N(0, G_m \otimes A_m)$, so Eq. [2] and [3] become

$$
y = X\beta + Zg_p + Zg_m + \epsilon
$$

and

$$
y|3, G_p, G_m, R_h] \sim \text{MVN}\left[0, Z'G_p \otimes A_p + G_m + A_m[Z' + I_g \otimes R_h]\right] \tag{5}
$$

**Modeling Covariance Structures**

In the multi-environment model of Eq. [2–5], the within-line across-environment covariance matrices $R_h$ and $G_j$ are completely unstructured (UN). Each of the matrices involves $J(J + 1)$ parameters. Therefore, the number of codispersion parameters grows quadratically with the number of environments. Also, when $J$ is large, and genetic and/or residual effects are highly correlated between environments, estimation of $G_{ij}$ or $R_h$ may be (close to) singular, making the convergence of algorithms slow. This problem may be overcome, and insights about correlations patterns may be gained, by using structured covariance matrices. This is done by representing covariance matrices as functions of a parameter vector. Factor analysis (e.g., Smith et al., 2001; Crossa et al., 2006; de los Campos and Gianola, 2007; Burgueño et al., 2007, 2008, 2011), random regression, principal components (Kirkpatrick and Meyer, 2004; Meyer and Kirkpatrick, 2005), and recursive and simultaneous structural equation models (Gianola and Sorensen, 2004) are alternative ways of structuring covariance matrices. Below we describe a few covariance structures applied to $R_h$, or $G_j$.

The most restrictive and simple covariance structure is obtained by setting $R_h = I \sigma^2$, where $\sigma^2$ is a variance parameter that is common across environments. This model, hereinafter denoted as $I$, imposes independence and homoscedasticity of model residuals across environments. A slightly less restrictive covariance structure is $R_h = \text{Diag}\{\sigma^2_i\}$, where $\sigma^2_i$, $(j$
and \((for \ FA)\) is equivalent to fitting \(J\) single-environments models separately.

The FA model is commonly used for modeling covariance matrices in quantitative genetic models in plants (e.g., Smith et al., 2001; Crossa et al., 2004, 2006; Burgueño et al., 2007, 2008, 2011) and animals (de los Campos and Gianola, 2007). The FA model for \(\text{Cov}(\mathbf{g}, \mathbf{g}') = \mathbf{G}_0 \otimes \mathbf{A}\) is \((\mathbf{AA}' + \Psi) \otimes \mathbf{A} = \mathbf{FA}(k) \otimes \mathbf{A}\) (for \(\mathbf{FA}(k) (k \leq j)\)), where \(\mathbf{A}\) is an \(J \times k\) matrix where the \(k\)th column contains the environmental loadings for the \(k\)th latent factor, and \(\Psi\) is an \(J \times J\) diagonal matrix, with different nonnegative parameters on the diagonal. When only one factor, \(k = 1\), is considered, the model is denoted as \(\text{FA}(1)\); for \(k = 2\), \(\text{FA}(2)\) has two multiplicative components, and so on. Thus, \(\mathbf{FA}\) can be interpreted as the linear regression of genotype and GE on environmental covariates (environmental loadings), with each genotype having a separate slope (genotypic scores) but a common intercept (if main effects of genotypes are not distinguished from GE). The slopes of genotypes measure the sensitivity of the genotypes to hypothetical environmental factors represented by the loadings of each environment (Smith et al., 2002).

Parameter identification requires imposing restrictions. Even after imposing these restrictions, loadings are identified up to an orthonormal rotation. Unique estimates can be obtained by either adopting a rotation (e.g., varimax) or by fixing some of the loadings. The number of loadings to fix for a \(k\)-common factor model, \(k > 1\), is equal to \(C_{kk}^{\frac{1}{2}}\). The number of parameters of a \(k\)-common factor model \((J \times k + J)\) must be equal to the number of unknowns in an unstructured model \((J \times (J + 1)/2)\); therefore, additional restrictions are required for identification, if \(k > (J - 1)/2\).

**Experimental Data**

We evaluated models such as those described in Eq. [2–3] and [4–5] using data from CIMMYT’s Global Wheat Program, which contains information on 599 wheat lines whose grain yield was evaluated in four environments (E1, low rainfall and irrigated; E2, high rainfall; E3, low rainfall and high temperature; and E4, low humidity and hot) (Braun et al., 1996) with two replicates in each environment. The combined analyses across four environments based on unstandardized data give a heritability for grain yield of 0.50 with a variance component for GE \((0.3132)\) larger than the variance component for entry \((0.0889)\). Individual environment analyses show some heterogeneity of residual errors ranging from 0.0335 for E1 to 0.1121 for E2. It should be pointed out that for pedigree and genomic prediction the data were environment standardized.

A pedigree tracing back many generations was available, and the Browse application of the International Crop Information System, as described in http://cropwiki.irri.org/icis/index.php/TDM_GMS_Browse (verified 20 Nov. 2011) (McLaren et al., 2005), was used for deriving the pedigree relationship matrix \((\mathbf{A}_p)\). Wheat lines were genotyped using Diversity Array Technology (DArt) markers generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au [verified 20 Nov. 2011]); after editing there were 1279 DArTs available for analysis. This data set is publicly available with the BLR package of R (de los Campos and Pérez, 2010) and was analyzed by Crossa et al. (2010), who used single-trait models for genomic prediction. Further details about the data set and editing procedures applied are described in Crossa et al. (2010).

**Models**

The main objective of this research was to evaluate the impact of modeling covariance structures in the context of pedigree, marker, and pedigree + marker models. To this end, we defined a sequence of 27 models, fitted them to the entire data set, and evaluated predictive ability using cross-validation (see below). Table 1 describes the models used for analysis. Models 1–9 (P) are pedigree-based models; Models 10–18 (M) are marker-based models. These models are as described by Eq. [2] and [3] with \(\mathbf{A} = \mathbf{A}_p\) and \(\mathbf{A} = \mathbf{A}_m\), respectively. Models 19–27 (PM) combine pedigree and marker information, as described by Eq. [4] and [5]. Therefore, the impact of incorporating marker information for prediction can be assessed by comparing models among these three groups.

Within each of these groups (P, M, and PM), we evaluated different strategies for structuring the covariance matrices of model residuals and marker effects. Comparing models using different covariance structures can be used to assess the importance of modeling heterogeneous variances (e.g., in P models by comparing Model 1 vs. Model 5) and covariances across environments (e.g., in P models by comparing Model 5 vs. Model 9). As stated earlier, using \(\mathbf{R}_0 = \mathbf{D}\) and \(\mathbf{G}_0 = \mathbf{D}\) (diagonal covariance structures) is equivalent to fitting \(J\) single-environment models separately; therefore, the impact of modeling covariances can be assessed by comparing the performance of models using \(\mathbf{R}_0 = \mathbf{D}\) and \(\mathbf{G}_0 = \mathbf{D}\) with those that consider genetic or residual covariances.

In models using FA(2), we imposed two restrictions: (i) we set the loading of the first trait on the second factor equal to zero, and (ii) we set one of the specific variances equal to zero. This yielded a model with 10 unknowns per covariance matrix.

**Model Validation**

Models were compared based on their predictive ability calculated as the correlation between the predictive and observed phenotypic values; the models were fitted to the training set and their predicted values were correlated to the observed values in the validation set. Additionally, we report an estimate of the probability of one change in the ranks produced by predicted and by observed performance. Details about this statistic are given in Appendix B.

Two distinct cross-validation (CV) schemes were used for generating the training and validation sets. A \(k\)-fold cross-validation, which consists of randomly dividing the \(n\) observations into \(k\) nonoverlapping subsets, say, \(S_1, S_2, \ldots, S_k\) was used. Cross-validation was then applied to each partition of the data, that is, \(k - 1\) groups were taken as the training set and the remaining group as the validation set. In this paper, we used \(k = 10\), a 10-fold cross-validation scheme.

The two cross-validation schemes (CV1 and CV2) mimic two possible real situations a breeder might face. The first CV scheme (CV1) was designed to evaluate the predictive ability of models when used for predicting the performance of genotypes that have not undergone field evaluation (i.e., newly developed
lines). When analyzing data for the kth fold of CV1, there was a total of 60 genotypes that were missing in all four environments; they were treated as unknown (Table C1 and Appendix C). Thus, predictions derived using CV1 were entirely based on phenotypic records of other lines. In the CV1 scheme, each validation set has 60 genotypes (except Fold 10, which has 59 genotypes).

The other cross-validation scheme (CV2) mimics a situation where lines are evaluated in two environments but missing in another two environments. Here information from relatives is used, and the prediction assessment can benefit from borrowing information between lines within an environment, between lines across environments, and among correlated environments (Table C1 and Appendix C). In the CV2 scheme, for each nonoverlapping validation set, we selected about 240 cells from the 599 × 4 GE matrix to confirm the validation set with the restriction that one genotype is missing only in two random environments and, for the other two environments, the same genotype is set to be missing in another fold. Randomness of the procedure produces a nonperfect balance pattern of environments and genotypes. The number of missing data in the different validation sets in CV2 ranged from 214 (Fold 7) to 268 (Fold 3). Appendix C shows a typical fold for CV1 and CV2.

Software
Models were fitted using the ASReml (2010) package. The REML estimates of variance parameters were obtained using the Average Information algorithm as implemented in ASReml (2010). Subsequently, estimates of fixed effects and predictions of random effects were obtained by solving the corresponding mixed-model equations with (co)dispersion parameters replaced by REML estimates.

RESULTS AND DISCUSSION
Patterns of Genetic and Residual Covariability
The estimates of residual and genetic covariance components from PMFA-UN (Model 9), MFA-UN (Model 18), and PMFA-UN (Model 27) in Table 2 show some patterns of heterogeneity of residuals and genetic variances in environments and varying trends of residual and genetic correlations across environments. Although the phenotypic data were standardized, results show that the estimates of residuals and genetic components are heterogeneous across environments for the three models due to differences in heritability. The extent of variance heterogeneity is expected to be even greater when data are not standardized.

The sample phenotypic correlation matrix indicated that there are two groups of environments, E1, which has a low and negative correlation with E2–E4, and the E2–E4 group of environments with moderate to high positive correlations among them. The residual and genetic correlation among environments (Table 2) confirmed these patterns of association among environments (i.e., E1 vs. E2–E4) and also indicated that the correlations induced by both effects are much smaller in R0 than in G0. Furthermore, the genomic-derived models (G0(UN)) produced higher association among environments than pedigree-derived models (G0).

Figure 1 displays the estimated loading (after varimax rotation) of each of the four environments on the first vs. the second common factor of G0 derived from models using pedigree- (PFA-UN) and marker-derived (MFA-UN) genetic relationship matrices. The two groups of environments, E1 and E2–E4, are clearly depicted. The results of the pedigree and genomic patterns of covariance among environments indicated some evidence of heterogeneous genetic and residual variances, as well as clear evidence of genetic association among some environments. The presence of such correlations can be exploited using multienvironment models to derive predictions that borrow information not only across lines, but also across environments. In the next section, we evaluate the impact of modeling across-environment covariances on prediction accuracy.
The residuals modeled using I, D, and UN had two effects that are confounded, the genetic variances that are represented by the additive × additive epistasis effects and the error variance. It is possible to partially overcome this problem by partitioning the total genotypic effects into additive and additive × additive and their interactions with environments (Burgueño et al., 2007). Fitting the additive × additive will increase the complexity of the models and it was not the main objective of the article. The authors are currently investigating modeling additive × additive epistasis and additive × additive × environment using pedigree and genomic relationship matrices in the context of genetic prediction in GS.

### Predictive Ability of Models with Pedigree and Genomic Information

The average correlations between predictions and phenotypes in CV1 and CV2 are presented in Table 3 (Models 1–18) and Table 4 (Models 19–27). Summaries of these
results are given in Fig. 2 and 3. The CV-correlations ranged from 0.517 to 0.646. In this range, the relationship between correlation and the estimated probability of having a rank change is close to linear (see Appendix B). A correlation of 0.317 (0.646) corresponds to a 0.40 (0.28) probability of having one change in rank.

**Pedigree vs. Marker-based Prediction**

The predictive ability of marker-based models was higher than that of pedigree-based models, both in CV1 and in CV2. Combining pedigree and marker information into the same model improved predictive ability as compared with that of pedigree- or marker-based models across covariance structures, environments, and CV schemes. Therefore, our results confirm the superiority of models for GS over pedigree-based predictions or marker-based predictions alone.

Crossa et al. (2010) evaluated the predictive ability of pedigree, marker, and pedigree plus marker single-environment models using the same data set used here. As expected, the estimates of CV-correlations obtained here in CV1 with models P_D_D (4), M_D_D (14), and P_M_D (23) are similar to those obtained by Crossa et al. (2010) using single-environment models for pedigree, markers (MBL = molecular marker regression using the Bayesian Least Absolute Shrinkage and Selection Operator [Bayesian LASSO]), and pedigree + markers combined (PM-BL = regression on markers and pedigree using the Bayesian LASSO). Differences are small (on the order of 0.4–4%) and the only exception was pedigree-based prediction in E3, where we obtained a predictive correlation that was 7.4% lower than that reported in Crossa et al. (2010). This difference could be due to many factors, such as the fact that Crossa et al. (2010) used a fully Bayesian approach, while we used REML followed by BLUE–BLUP, or differences could be due to the model used to incorporate marker information (we used a model based entirely on Gaussian assumptions, while the Bayesian LASSO uses a prior for marker effects which is a mixture of scaled normal densities). Finally, differences in estimates of CV-correlation may be partially due to variability in estimates originated by sampling of training and validation sets.

**Cross-validation Scheme**

Predicting the performance of newly developed lines (CVI) (not tested in the field) is more challenging than predicting the performance of lines that have been evaluated in different but correlated environments

---

Table 3. Mean correlations from 10-fold cross-validation between the predicted and the observed values of genotypes for all environments, for individual Environments CV1 and CV2, and for Environments 1, 2, 3, and 4 together for 18 different models (Models 1–18) for two different cross-validation schemes (CV1 and CV2), each with 10-fold. For Model 12, the predicted and observed values are from single-environment models using the same covariance structure as that of the model.

<table>
<thead>
<tr>
<th>Model code†</th>
<th>PI-I</th>
<th>PD-I</th>
<th>PFA-I</th>
<th>PI-D</th>
<th>PD-D</th>
<th>PFA-D</th>
<th>PI-UN</th>
<th>PD-UN</th>
<th>PFA-UN</th>
<th>MI-I</th>
<th>MD-I</th>
<th>MFA-I</th>
<th>MI-D</th>
<th>MD-D</th>
<th>MFA-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environment 1</td>
<td>0.444</td>
<td>0.443</td>
<td>0.437</td>
<td>0.444</td>
<td>0.443</td>
<td>0.433</td>
<td>0.428</td>
<td>0.444</td>
<td>0.444</td>
<td>0.449</td>
<td>0.449</td>
<td>0.449</td>
<td>0.449</td>
<td>0.449</td>
<td>0.449</td>
</tr>
<tr>
<td>Environment 2</td>
<td>0.416</td>
<td>0.416</td>
<td>0.410</td>
<td>0.416</td>
<td>0.416</td>
<td>0.406</td>
<td>0.365</td>
<td>0.350</td>
<td>0.422</td>
<td>0.422</td>
<td>0.623</td>
<td>0.422</td>
<td>0.422</td>
<td>0.623</td>
<td>0.422</td>
</tr>
<tr>
<td>Environment 3</td>
<td>0.387</td>
<td>0.386</td>
<td>0.389</td>
<td>0.387</td>
<td>0.386</td>
<td>0.388</td>
<td>0.331</td>
<td>0.317</td>
<td>0.424</td>
<td>0.424</td>
<td>0.633</td>
<td>0.424</td>
<td>0.424</td>
<td>0.633</td>
<td>0.424</td>
</tr>
<tr>
<td>Environments E2–E4</td>
<td>0.414</td>
<td>0.413</td>
<td>0.409</td>
<td>0.414</td>
<td>0.413</td>
<td>0.405</td>
<td>0.365</td>
<td>0.355</td>
<td>0.413</td>
<td>0.413</td>
<td>0.596</td>
<td>0.413</td>
<td>0.413</td>
<td>0.596</td>
<td>0.413</td>
</tr>
</tbody>
</table>

† Model code: Model 1: PI-I = identity-pedigree for R0P and identity for R0; Model 4: PI-D = identity-pedigree for R0P and diagonal for R0; Model 5: PFA-I = factor analytic-pedigree for R0P and identity for R0; Model 8: PD-UN = diagonal-pedigree for R0P and unstructured for R0; Model 11: MD-I = diagonal-genomic for G0M and identity for G0M; Model 14: MD-D = diagonal-genomic for G0M and diagonal for G0M; Model 17: MI-I = identity-genomic for G0M and identity for G0M; Model 20: MI-UN = identity-genomic for G0M and unstructured for G0M.
Table 4. Mean correlations from 10-fold cross-validation between the predicted and the observed values of genotypes for all environments, for individual Environments 1, 2, 3, and 4, and for Environments 2, 3, 4 together for nine different models including $A_p$ and $A_m$ simultaneously (Models 19–27) for two different cross-validation schemes (CV1 and CV2), each with 10-fold. For CV1 and CV2, the best predicted model among Models 19–27 are underlined.

<table>
<thead>
<tr>
<th>Model no.</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model code</strong></td>
<td>PM $I_I$</td>
<td>PM $P-I$</td>
<td>PM $FA-I$</td>
<td>PM $I_D$</td>
<td>PM $P-D$</td>
<td>PM $FA-D$</td>
<td>PM $I-UN$</td>
<td>PM $P-UN$</td>
<td>PM $FA-UN$</td>
</tr>
<tr>
<td>Environment 1</td>
<td>0.562</td>
<td>0.559</td>
<td>0.542</td>
<td>0.563</td>
<td>0.560</td>
<td>0.551</td>
<td>0.560</td>
<td>0.564</td>
<td>0.560</td>
</tr>
<tr>
<td>Environment 2</td>
<td>0.502</td>
<td>0.504</td>
<td>0.469</td>
<td>0.502</td>
<td>0.503</td>
<td>0.482</td>
<td>0.469</td>
<td>0.458</td>
<td>0.495</td>
</tr>
<tr>
<td>Environment 3</td>
<td>0.432</td>
<td>0.433</td>
<td>0.432</td>
<td>0.433</td>
<td>0.433</td>
<td>0.432</td>
<td>0.380</td>
<td>0.374</td>
<td>0.423</td>
</tr>
<tr>
<td>Environment 4</td>
<td>0.479</td>
<td>0.481</td>
<td>0.444</td>
<td>0.479</td>
<td>0.480</td>
<td>0.453</td>
<td>0.457</td>
<td>0.462</td>
<td>0.462</td>
</tr>
<tr>
<td>Environments E2–E4</td>
<td>0.471</td>
<td>0.474</td>
<td>0.445</td>
<td>0.471</td>
<td>0.472</td>
<td>0.454</td>
<td>0.437</td>
<td>0.431</td>
<td>0.459</td>
</tr>
<tr>
<td>Environments E1–E4</td>
<td>0.494</td>
<td>0.496</td>
<td>0.4722</td>
<td>0.494</td>
<td>0.495</td>
<td>0.480</td>
<td>0.469</td>
<td>0.464</td>
<td>0.486</td>
</tr>
</tbody>
</table>

1 Model code: Model 19: PM $I_I$ = identity-pedigree ($G_{Pi}$), identity-genomic ($G_{Pi}$) with identity for $R_i$; Model 20: PM $P-I$ = diagonal-pedigree ($G_{Pi}$), diagonal-genomic for ($G_{Pi}$) with identity for $R_i$; Model 21: PM $FA-I$ = factor analytic-pedigree ($G_{Pi}$), factor analytic-genomic for ($G_{Pi}$) with identity for $R_i$; Model 22: PM $I-D$ = identity-pedigree ($G_{Pi}$), identity-genomic ($G_{Pi}$) with diagonal for $R_i$; Model 23: PM $P-D$ = diagonal-pedigree ($G_{Pi}$), diagonal-genomic for ($G_{Pi}$) with diagonal for $R_i$; Model 24: PM $FA-D$ = factor analytic-pedigree ($G_{Pi}$), factor analytic-genomic for ($G_{Pi}$) with diagonal for $R_i$; Model 25: PM $I-UN$ = identity-pedigree ($G_{Pi}$), identity-genomic ($G_{Pi}$) with unstructured for $R_i$; Model 26: PM $P-UN$ = diagonal-pedigree ($G_{Pi}$), diagonal-genomic for ($G_{Pi}$) with unstructured for $R_i$; Model 27: PM $FA-UN$ = factor analytic-pedigree ($G_{Pi}$), factor analytic-genomic for ($G_{Pi}$) with unstructured for $R_i$.

(CV2). This is reflected in the results depicted in Fig. 2: averaged across environments, CV-correlations obtained from multi-environment models (PM $FA-UN$, PM $FA-UN^*$, and PM $FA-UN$) in CV2 were 31, 17.5, and 21.8% greater than those obtained in CV1 (Fig. 2). This highlights the value of having information from correlated environments when predicting performance. Selection of lines without field testing, as mimicked in CV1, allows shortening of the generation interval, but the predictability is poorer, which might compromise the annual rate of genetic progress in a GS breeding program. Ultimately, the decision of how the breeding scheme should be structured depends on the tradeoff between desired prediction accuracy and generation interval.

### Prediction Assessment

Modeling genetic and residual covariances improved the predictability in CV2 but not in CV1. The reason for this is clear; in CV2 modeling GE allows borrowing of information within line across environments. This is not possible in CV1 because in this scheme the lines being predicted have not been evaluated in any environment.

The impact of modeling GE in CV2 is marked in E2–E4, and genetic markers. In terms of probability of a rank change between observed and predicted genotypic values in E2, E3, and E4 using FA-D and FA-UN structures for $G_{UP}$ $G_{IM}$ and $R_0$ (for CV2 in Fig. 3c), results show a probability of a single rank change of about 0.145; this value is slightly higher than that estimated for E1. When GE is not modeled (D-D and D-UN), the probability of one rank change for genotypes in E2, E3, and E4 is slightly larger (0.17, Fig. 3c) and is the same for rank changes occurring in E1.

### Predictive Ability of Models without Pedigree and Genomic Information

It is interesting to examine the predictive accuracy of models without using pedigree and markers. However, if pedigree or markers are not included, not many models can be fitted for prediction; that is, models for CV1 will predict with the mean, and it is not possible to compute correlations for those genotypes that are completely missing in environments. For CV2, two simple linear mixed models (I and II) were fitted to the training sets of the 10-fold cross-validations without including the pedigree and the marker information. Both models can be described from Model [2]. Model I is $y = X\beta + Z_\epsilon + \epsilon$, where $X$ is the incidence matrix for the fixed effects of environments $\beta$, and $Z$ is the incidence matrix for the random effects of the genotypes $\epsilon$ with Cov$(\epsilon, \epsilon') = \sigma^2 I$, and the random residual ($\epsilon$) is Cov$(\epsilon, \epsilon') = \sigma^2 I$. In this model, the residual contains some of the nonadditive genetic effects, the GE effects, and parts of the error. Model II is $y = X\beta + Z_\epsilon + \epsilon$, with the fixed effects of environments $\beta$ and the random
effects $g$ that contain the genetic effects, the GE effects, and the error term, $e$; the random $g$ effects are modeled using the FA such that $\text{Cov}(g, g') = (A \Lambda' + \Psi) \otimes I_g = FA(k) \otimes I_g$.

The prediction accuracy of Model I for individual environments and for E2–E4 and E1–E4 ranged from $-0.21$ for E1 to 0.50 for E2; it was always lower than correlations obtained from Model II. The prediction accuracy of Model II was always better than that of Model I and ranged from 0.22 for E1 to 0.607 for E2. As for the previous cases, E2 and E3 are predicted with more
accuracy than the other environments because they have higher genetic correlations. These results are in agreement with Burgueño et al. (2011), who demonstrated that modeling GE is a good thing because it always gives better predictability than using simple linear mixed models.

When comparing the correlations of Model II with those of Models 3, 6, and 9 (pedigree with FA) and Models 12, 15, and 18 (markers with FA) (Table 3), it is clear that all the models that use pedigree or markers or both gave better predictive accuracy for individual environments and for E2–E4 and E1–E4 than Model II (which modeled the GE but did not include additional pedigree and marker information). These results show that for predicting the environments that cause a great deal of the GE, such as environments E1 (and E4), modeling GE using information on molecular markers and/or pedigree gives better prediction accuracy than not modeling GE and not using molecular markers and pedigree information.

CONCLUSIONS
This is the first research that incorporates FA models for GE interaction using pedigree or/and marker-based information simultaneously for genetic prediction in the context of GS. The results show that combining pedigree and marker data can yield substantial increases in prediction accuracy relative (i) to traditional pedigree-based prediction and (ii) to single-environment pedigree and genomics prediction models. Furthermore, for predicting the performance of newly developed lines (CV1), single-environment models are expected to perform similarly to multienvironment models. However, multienvironment models can boost predictive power in across-environment prediction, a problem of great interest in most plant breeding programs. Our results suggest that for the germplasm and environmental conditions used here, most of the benefit of the multienvironment model comes from modeling genetic correlations between environments. We also showed that modeling GE using information on molecular markers and/or pedigree gives better prediction accuracy than not using molecular markers and pedigree information. Further research is required to examine the prediction assessment of modeling the nonadditive genetic variances such as dominance and epistasis and their interactions with environments.

APPENDIX A
Equivalence of Regression Models for Genomic Selection
In the model of Eq. [1], consider replacing genetic values, $g_{ij}$, with a regression on marker genotypes, that is, $g_{ij} = \sum_{i=1}^{L} m_{ij}b_{ij}$, where $m_{ij}$ counts the number of copies of the minor-frequency allele carried by the $i$th line at the $j$th locus, and $b_{ij}$ represents the effect of the $i$th marker on phenotypes in the $j$th environment. In matrix notation this can be expressed as $\mathbf{g}_j = \mathbf{M}\mathbf{b}_j$, where $\mathbf{M} = \{m_{ij}\}$ is a matrix of marker genotypes, and $\mathbf{b}_j = (b_{ij},..., b_{ij})$ is a vector of marker effects. By stacking the genetic values of all lines in all environments into a single vector, we obtained $\mathbf{g} = [\mathbf{I}_j \otimes \mathbf{M}]\mathbf{b}_j$, where $\mathbf{g} = (\mathbf{g}_1',...\mathbf{g}_7')'$ and $\mathbf{b} = (\mathbf{b}_1',...\mathbf{b}_7')$.

Consider assigning a multivariate normal before the joint vector of marker effects, centered at zero and with covariance matrix

$$\text{Cov}(\mathbf{b}_i, \mathbf{b}_j) = \begin{cases} G_{ij} & \text{if } l = l' \\ 0 & \text{otherwise} \end{cases}$$

In matrix notation this is expressed as $\text{Cov} (\mathbf{b}) = \mathbf{G}_b \otimes I_{L}$, where $\mathbf{G}_b = \{g_{ij}, g_{ij}\}$ is a within-locus, across-environments covariance matrix of marker effects. Therefore,

$$\mathbf{b} - \text{MVN}(\mathbf{0}, \mathbf{G}_b \otimes I_{L})$$

The vector of genetic values, $\mathbf{g} = [\mathbf{I}_j \otimes \mathbf{M}]\mathbf{b}_j$, is a linear combination of the vector of marker effects. Therefore, using Eq. [1A] and from properties of the multivariate normal density, it follows that $\mathbf{g}$ is also multivariate normal; further, $E[\mathbf{g}] = [\mathbf{I}_j \otimes \mathbf{M}]E[\mathbf{b}] = 0$ and

$$\text{Cov}(\mathbf{g},\mathbf{g}') = [\mathbf{I}_j \otimes \mathbf{M}]\text{Cov}(\mathbf{b},\mathbf{b}')[\mathbf{M}' \otimes \mathbf{I}_j] = [\mathbf{I}_j \otimes \mathbf{M}][\mathbf{G}_b \otimes \mathbf{I}_L][\mathbf{M}' \otimes \mathbf{I}_j] = [\mathbf{G}_b \otimes \mathbf{M}][\mathbf{M}' \otimes \mathbf{I}_j] = \mathbf{G}_b \otimes [\mathbf{M}\mathbf{M}']$$

Therefore, using $\mathbf{A} = \mathbf{MM}'$ in the multi-environment linear mixed model of Eq. [1–3] is equivalent to a linear regression model for genomic selection in which marker effects are environment specific and are assigned a multivariate normal prior centered at zero and with covariance structure $\text{Cov}(\mathbf{b}) = \mathbf{G}_b \otimes I_{L}$.

APPENDIX B
Probability of Rank Change of Two Correlated Normal Bivariate Random Variables
We derived the probability of a rank change between the observed phenotypes and the predicted genetic values by assuming a bivariate normal distribution of a pair of random variables with a given correlation. Perfectly correlated variables have zero probability of rank change, whereas variables with zero correlations have a rank change probability of 0.50. This is an approximate approach for model selection that, in this case, offers a measure that is easy to interpret.

To calculate the probability of a change of rank, we considered a bivariate normal distribution $(X_i, Y_i)$, where $X_i$ (for $i = 1,2,...,g$) is the observed genotypic value and $Y_i$ is the predicted genotypic value. The Kendall correlation coefficient is defined as the probability of concordance minus the probability of discordance:
The probability of concordance can be interpreted as the probability that the rank between any two genotypes is the same using either the observed or the predicted genetic value. On the other hand, the probability of discordance can be interpreted as the probability that the rank between any two genotypes changes depending on whether observed or predicted genotypic values are used.

Therefore,

\[ P[(X_i - X_j) > 0] - P[(X_i - X_j) < 0] = \frac{1 - \frac{\pi}{2}}{2} \]

For the range of correlation found in this study (from 0.317 to 0.646), the relationship between the given correlation and the estimated probability of having one rank change is close to linear (Fig. B1).

**APPENDIX C**

**Cross-validation Schemes**

Table C1 shows an example of one-fold cross-validation (CV) in both the CV1 and CV2 schemes. In CV1, for each fold, 60 different genotypes are missing in all four environments. Missing genotypes are denoted as dots.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.7</td>
<td>-1.7</td>
<td>-1.9</td>
<td>0.1</td>
<td>1.7</td>
<td>-1.7</td>
<td>-1.9</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
<td>-0.6</td>
<td>-0.8</td>
<td>-1.1</td>
<td>0.3</td>
<td>-0.6</td>
<td>-0.8</td>
<td>-1.1</td>
</tr>
<tr>
<td>4</td>
<td>0.8</td>
<td>0.1</td>
<td>0.6</td>
<td>0.6</td>
<td>0.8</td>
<td>0.1</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>1.0</td>
<td>-0.3</td>
<td>1.6</td>
<td>-0.1</td>
<td>1.0</td>
<td>-0.3</td>
<td>1.6</td>
<td>-0.1</td>
</tr>
<tr>
<td>6</td>
<td>2.3</td>
<td>0.6</td>
<td>0.1</td>
<td>0.7</td>
<td>2.3</td>
<td>0.6</td>
<td>0.1</td>
<td>0.7</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>8</td>
<td>0.6</td>
<td>-0.4</td>
<td>-0.7</td>
<td>1.0</td>
<td>0.6</td>
<td>-0.4</td>
<td>-0.7</td>
<td>1.0</td>
</tr>
<tr>
<td>9</td>
<td>-1.0</td>
<td>-1.8</td>
<td>-1.9</td>
<td>-1.5</td>
<td>.</td>
<td>-1.8</td>
<td>.</td>
<td>-1.5</td>
</tr>
<tr>
<td>10</td>
<td>-1.1</td>
<td>-1.6</td>
<td>-2.0</td>
<td>-0.6</td>
<td>-1.1</td>
<td>-1.6</td>
<td>-2.0</td>
<td>-0.6</td>
</tr>
<tr>
<td>11</td>
<td>1.7</td>
<td>-0.3</td>
<td>-0.2</td>
<td>0.3</td>
<td>1.7</td>
<td>-0.3</td>
<td>-0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>12</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>13</td>
<td>1.2</td>
<td>-0.2</td>
<td>0.8</td>
<td>-0.2</td>
<td>.</td>
<td>-0.2</td>
<td>0.8</td>
<td>.</td>
</tr>
<tr>
<td>14</td>
<td>0.3</td>
<td>-0.7</td>
<td>-0.4</td>
<td>0.0</td>
<td>0.3</td>
<td>-0.7</td>
<td>-0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>15</td>
<td>0.8</td>
<td>-0.6</td>
<td>-0.7</td>
<td>-0.2</td>
<td>0.8</td>
<td>-0.6</td>
<td>-0.7</td>
<td>-0.2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>589</td>
<td>-0.8</td>
<td>0.7</td>
<td>0.7</td>
<td>2.6</td>
<td>-0.8</td>
<td>0.7</td>
<td>0.7</td>
<td>2.6</td>
</tr>
<tr>
<td>590</td>
<td>-0.9</td>
<td>0.1</td>
<td>1.1</td>
<td>2.8</td>
<td>-0.9</td>
<td>0.1</td>
<td>1.1</td>
<td>2.8</td>
</tr>
<tr>
<td>591</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-1.1</td>
<td>.</td>
<td>.</td>
<td>-1.1</td>
<td>2.0</td>
</tr>
<tr>
<td>592</td>
<td>-1.0</td>
<td>1.1</td>
<td>1.9</td>
<td>1.7</td>
<td>-1.0</td>
<td>1.1</td>
<td>1.9</td>
<td>1.7</td>
</tr>
<tr>
<td>593</td>
<td>-1.1</td>
<td>0.6</td>
<td>2.4</td>
<td>1.2</td>
<td>-1.1</td>
<td>0.6</td>
<td>2.4</td>
<td>1.2</td>
</tr>
<tr>
<td>594</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-0.5</td>
<td>.</td>
<td>.</td>
<td>-0.5</td>
<td>1.8</td>
</tr>
<tr>
<td>595</td>
<td>-1.2</td>
<td>1.4</td>
<td>1.6</td>
<td>1.7</td>
<td>-1.2</td>
<td>1.4</td>
<td>1.6</td>
<td>1.7</td>
</tr>
<tr>
<td>596</td>
<td>-1.1</td>
<td>0.1</td>
<td>2.1</td>
<td>0.6</td>
<td>-1.1</td>
<td>0.1</td>
<td>2.1</td>
<td>0.6</td>
</tr>
<tr>
<td>597</td>
<td>-1.2</td>
<td>0.5</td>
<td>2.0</td>
<td>1.8</td>
<td>.</td>
<td>0.5</td>
<td>1.8</td>
<td>.</td>
</tr>
<tr>
<td>598</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-0.9</td>
<td>.</td>
<td>-0.9</td>
<td>0.5</td>
<td>1.7</td>
</tr>
<tr>
<td>599</td>
<td>0.2</td>
<td>-0.5</td>
<td>-0.3</td>
<td>0.4</td>
<td>0.2</td>
<td>-0.5</td>
<td>-0.3</td>
<td>0.4</td>
</tr>
</tbody>
</table>
environments. In the CV2 scheme, one random genotype was missing in two random environments until approximately 240 missing cells were completed. To have 10 non-overlapping subsets, the two remaining environments, which have no missing data, will have missing data in another fold. For example, Genotype 2, which is missing in E1 and E2 in the fold represented in Table C1, will be missing in E3 and E4 in another fold.

**Acknowledgments**

The authors thank the numerous cooperators in national agricultural research institutes who carried out Elite Spring Wheat Yield Trials (ESWYT), and provided the phenotypic data presented here. We also thank the International Nursery and Seed Distribution Units at CIMMYT-Mexico for preparing and distributing the seed and computerizing the data. Kent Weigel acknowledges financial support from the National Association of Animal Breeders (Columbia, MO).

**References**


Pérez, P., G. de los Campos, J. Crossa, and D. Gianola. 2010. Genomic-enabled prediction based on molecular markers and