

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. Multienvironment (multitrait) models allow borrowing of information across environments (traits), which could enhance prediction accuracy. This study presents multienvironment (multitrait) models for GS and compares the predictive accuracy of these models with: (i) multienvironment analysis without pedigree and marker information, and (ii) multienvironment pedigree or/and marker-based models. A statistical framework for incorporating pedigree and molecular marker information in models for multienvironment data is described and applied to data that originate from wheat (*Triticum aestivum* L.) multienvironment 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 multienvironment GS models.

J. Burgueño and J. Crossa, Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600, México D.F., México; G. de los Campos, Dep. of Biostatistics, Univ. of Alabama at Birmingham, Ryals Public Health Bldg. 443, Birmingham, AL 35294; K. Weigel, Dep. of Dairy Science, Univ. of Wisconsin, Madison, WI 53706. Received 3 June 2011. *Corresponding author (j.crossa@cgiar.org).

Abbreviations: Bayesian LASSO, Bayesian Least Absolute Shrinkage and Selection Operator; CV, cross-validation; D, diagonal; FA, factor analytic model; GE, genotype \times environment interaction; GS, genomic selection; **I**, identity matrix; M, molecular marker; P, pedigree; PM, pedigree + molecular marker; REML, restricted maximum likelihood; UN, unstructured.

IN PLANT BREEDING, multienvironment trials play a fundamental role for assessing the performance of genotypes across different environmental conditions, for studying genotype \times environment interaction (GE) and genotype stability, and for predicting the performance of untested genotypes. Genotype \times environment interaction can be incorporated into additive infinitesimal models (e.g., Fisher, 1918; Wright, 1921) by means of genetic and environmental covariances (e.g., Piepho, 1997, 1998, 2009; Smith et al., 2001; Crossa et al., 2004, 2006; Oakey et al., 2006; Burgueño et al., 2007, 2008, 2011). Such an approach allows deriving predictions of genetic values that borrow information across individuals through genetic relationships, and within individuals (across environments) through genetic and environmental covariances.

In pedigree-based additive infinitesimal models (e.g., Fisher, 1918; Wright, 1921; Henderson, 1975), genetic relationships are commonly incorporated by the use of the numerator relationship matrix $\mathbf{A} = \{a(i, i')\}$. This matrix, whose entries equal twice the

Published in Crop Sci. 52:707–719 (2012).

doi: 10.2135/cropsci2011.06.0299

Freely available online through the author-supported open-access option.

© Crop Science Society of America | 5585 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 multi-environment 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 multi-environment 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 multi-environment 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 multi-environment models for GS and to compare their predictive power with: (i) multi-environment analysis without pedigree and marker information, and (ii) multi-environment 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 multi-environment data is described and applied in a plant breeding context using a multi-environment 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 Multi-environment Mixed Model for Regression using Molecular Marker and Pedigree Data

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

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \cdot \\ \mathbf{y}_j \\ \cdot \\ \mathbf{y}_J \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdot & \cdot & \cdot & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \mathbf{0} & \cdot & \cdot & \cdot & \cdot & \mathbf{0} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \beta_j \\ \cdot \\ \beta_J \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdot & \cdot & \cdot & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \mathbf{0} & \cdot & \cdot & \cdot & \cdot & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \cdot \\ \mathbf{g}_j \\ \cdot \\ \mathbf{g}_J \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \epsilon_j \\ \cdot \\ \epsilon_J \end{bmatrix} \quad [1]$$

where \mathbf{y}_j is the vector of the response variable of phenotypic records (or means) collected in the j th environment; and \mathbf{X}_j and \mathbf{Z}_j are incidence matrices vectors of systematic effects, β_j , and random genetic effects, \mathbf{g}_j , in the j th environment (with elements \mathbf{g}_{ij} for the i th genotype). In reduced matrix notation,

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{g} + \epsilon \quad [2]$$

where $\mathbf{y} = (\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_J)'$, $\mathbf{g} = (\mathbf{g}'_1, \mathbf{g}'_2, \dots, \mathbf{g}'_J)'$, $\epsilon = (\epsilon'_1, \epsilon'_2, \dots, \epsilon'_J)'$.

In a standard multi-environment 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 $\text{Cov}(\mathbf{g}, \mathbf{g}') = \mathbf{G}_0 \otimes \mathbf{A}$, $\text{Cov}(\epsilon, \epsilon') = \mathbf{I}_g \otimes \mathbf{R}_0$, and $\text{Cov}(\mathbf{g}, \epsilon') = \mathbf{0}$, where \mathbf{I} is an 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 $\mathbf{A} = \{a(i, i')\}$ is a $g \times g$ numerator relationship matrix, whose entries are twice the kinship coefficients between pairs of lines. The $\text{Cov}(\mathbf{g}, \mathbf{g}') = \mathbf{G}_0 \otimes \mathbf{A}$ is usually represented as

$$\mathbf{G}_0 \otimes \mathbf{A} = \begin{bmatrix} \sigma_{a_1}^2 & \rho_{12}\sigma_{a_1}\sigma_{a_2} & \cdot & \cdot & \cdot & \rho_{1J}\sigma_{a_1}\sigma_{a_J} \\ \rho_{21}\sigma_{a_1}\sigma_{a_2} & \sigma_{a_2}^2 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho_{J1}\sigma_{a_1}\sigma_{a_J} & \cdot & \cdot & \cdot & \cdot & \sigma_{a_J}^2 \end{bmatrix} \otimes \mathbf{A}$$

where the j th diagonal element of the $J \times J$ matrix \mathbf{G}_0 is the additive genetic variance $\sigma_{a_j}^2$ within the j th environment, and the $j'j$ th element is the additive genetic covariance $\rho_{j'j}\sigma_{a_j'}\sigma_{a_j}$ between sites j' and j ; thus, $\rho_{j'j}$ is the correlation of the additive genetic effects between sites j' and j .

The $\mathbf{R}_0 = \{\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'})\}$ 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 \mathbf{I}_g in $\mathbf{I}_g \otimes \mathbf{R}_0$ with a diagonal matrix, $\mathbf{D} = \text{Diag}\{\sigma_j^2\}$.

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

$$[\mathbf{y}|\boldsymbol{\beta}, \mathbf{R}_0, \mathbf{G}_0] \sim \text{MVN}[\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}(\mathbf{G}_0 \otimes \mathbf{A})\mathbf{Z}' + \mathbf{I}_g \otimes \mathbf{R}_0] \quad [3]$$

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 parameters 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 second stage for handling the differences in standard errors among genotype estimates within environments is recommended (Welham et al., 2010) for multi-environment trials whose purpose is variety selection. In GS, the aim is to achieve accuracy in the genomic estimate breeding value based on marker or/and pedigree.

It should also be noted that the residuals of Model [2] comprise the confounded effects of nonadditive genetic variance and the error residual; this nonadditive genetic variance represents, in self-pollinated species, the additive \times additive epistasis that can be modeled using the relationship matrix \mathbf{A} , as shown by Burgueño et al. (2007).

Pedigree and Marker-based Prediction

In the models described by [1–3], $\mathbf{A} = \{a(i, i')\}$ represents a $g \times g$ matrix of additive relationships. Commonly, these are derived

from a pedigree (hereinafter denoted as \mathbf{A}_p). Alternatively, this matrix can be derived from molecular marker information (hereinafter denoted as \mathbf{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 $\mathbf{A}_M = \mathbf{M}\mathbf{M}'$, where $\mathbf{M} = \{m_{il}\}$ is a matrix containing centered and standardized genotypes ($i = 1, \dots, g$) at different markers ($l = 1, \dots, L$), that

is, $m_{il} = \frac{m_{il} - p_l}{\sqrt{p_l(1-p_l)}}$, where m_{il} counts the number of copies of the minor-frequency allele carried by the i th line at the l th locus, and p_l is the estimated minor-frequency allele of the l th marker. In Appendix A we show the equivalence of using $\mathbf{A}_M = \mathbf{M}\mathbf{M}'$ 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, $\mathbf{g}_p \sim N(\mathbf{0}, \mathbf{G}_{op} \otimes \mathbf{A}_p)$, represents a regression on a pedigree, and the other, a regression on markers, $\mathbf{g}_m \sim N(\mathbf{0}, \mathbf{G}_{om} \otimes \mathbf{A}_M)$, so Eq. [2] and [3] become

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_p + \mathbf{Z}\mathbf{g}_m + \boldsymbol{\varepsilon} \quad [4]$$

and

$$[\mathbf{y}|\boldsymbol{\beta}, \mathbf{G}_{op}, \mathbf{G}_{om}, \mathbf{R}_0] \sim \text{MVN}[\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}(\mathbf{G}_{op} \otimes \mathbf{A}_p + \mathbf{G}_{om} \otimes \mathbf{A}_M)\mathbf{Z}' + \mathbf{I}_g \otimes \mathbf{R}_0] \quad [5]$$

Modeling Covariance Structures

In the multi-environment model of Eq. [2–5], the within-line across-environment covariance matrices \mathbf{R}_0 and \mathbf{G}_0 are completely unstructured (UN). Each of the matrices involves $\frac{J(J+1)}{2}$ 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 \mathbf{G}_0 or \mathbf{R}_0 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 \mathbf{R}_0 or \mathbf{G}_0 .

The most restrictive and simple covariance structure is obtained by setting $\mathbf{R}_0 = \mathbf{I}_g\sigma_\varepsilon^2$, where σ_ε^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 $\mathbf{R}_0 = \text{Diag}\{\sigma_{\varepsilon_j}^2\}$, where $\sigma_{\varepsilon_j}^2$ (j

$= 1, \dots, J$) are environment-specific variance parameters. Fitting the model described above with $\mathbf{R}_0 = \text{Diag}\{\sigma_{\epsilon_j}^2\}$ and $\mathbf{G}_0 = \text{Diag}\{\sigma_{\epsilon_j}^2\}$ 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 $(\Lambda\Lambda' + \Psi) \otimes \mathbf{A} = \text{FA}(k) \otimes \mathbf{A}$ (for $\text{FA}(k)$ ($k \leq s$)), 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 Ψ 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 FA(1); for $k = 2$, FA(2) has two multiplicative components, and so on. Thus, 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_2^k . The number of parameters of a k -common factor model ($J \times k + J$) must at most equal 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, \dots, 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 k th 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 P_{FA-UN} (Model 9), M_{FA-UN} (Model 18), and PM_{FA-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 R_0 than in G_0 . Furthermore, the genomic-derived models (G_{0M}) produced higher association among environments than pedigree-derived models (G_{0P}).

Table 1. Description of the pedigree (P) and genomic (M) models used for the analysis.[†]

No.	Model Code	Covariance structure		
		R_0	G_{0P}	G_{0M}
Pedigree				
1	P_{I-I}	I	I	—
2	P_{D-I}	I	D	—
3	P_{FA-I}	I	FA(2)	—
4	P_{I-D}	D	I	—
5	P_{D-D}	D	D	—
6	P_{FA-D}	D	FA(2)	—
7	P_{I-UN}	UN	I	—
8	P_{D-UN}	UN	D	—
9	P_{FA-UN}	UN	FA(2)	—
Marker				
10	M_{I-I}	I	—	I
11	M_{D-I}	I	—	D
12	M_{FA-I}	I	—	FA(2)
13	M_{I-D}	D	—	I
14	M_{D-D}	D	—	D
15	M_{FA-D}	D	—	FA(2)
16	M_{I-UN}	UN	—	I
17	M_{D-UN}	UN	—	D
18	M_{FA-UN}	UN	—	FA(2)
Pedigree + marker				
19	PM_{I-I}	I	I	I
20	PM_{D-I}	I	D	D
21	PM_{FA-I}	I	FA(2)	FA(2)
22	PM_{I-D}	D	I	I
23	PM_{D-D}	D	D	D
24	PM_{FA-D}	D	FA(2)	FA(2)
25	PM_{I-UN}	UN	I	I
26	PM_{D-UN}	UN	D	D
27	PM_{FA-UN}	UN	FA(2)	FA(2)

[†]I = Identity (imposes independence and homoscedasticity, one parameter); D = Diagonal (imposes independence but allow heterogeneous variances, four parameters); FA(2) = two-common factor model (imposes heterogeneous variances and correlation between environments, 10 parameters); UN = unstructured (imposes heterogeneous variances and correlation between environments, 10 parameters).

Figure 1 displays the estimated loading (after varimax rotation) of each of the four environments on the first vs. the second common factor of G_0 derived from models using pedigree- (P_{FA-UN}) and marker-derived (M_{FA-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.

Table 2. Estimates of residual and genetic covariance matrices (correlation in the upper diagonal), variances and covariances in the diagonal and lower diagonal components of each matrix, respectively, derived from a full-data analysis of models using different covariance structures for four environments (1–4).

Model†	Env.‡	R_0				G_{OP}				G_{OM}			
		1	2	3	4	1	2	3	4	1	2	3	4
P_{FA-UN} (Model 9)	1	0.558	0.064	-0.248	0.021	0.573	-0.139	-0.131	-0.316	—	—	—	—
	2	0.035	0.556	0.390	0.250	-0.078	0.553	0.965	0.575	—	—	—	—
	3	-0.131	0.205	0.498	0.163	-0.080	0.583	0.660	0.593	—	—	—	—
	4	0.011	0.134	0.082	0.517	-0.185	0.331	0.373	0.598	—	—	—	—
M_{FA-UN} (Model 18)	1	0.526	0.153	-0.195	0.116	—	—	—	—	2.998	-0.258	-0.2080	-0.479
	2	0.084	0.573	0.484	0.259	—	—	—	—	-0.727	2.634	0.925	0.623
	3	-0.112	0.292	0.634	0.189	—	—	—	—	-0.547	2.280	2.305	0.648
	4	0.065	0.152	0.117	0.602	—	—	—	—	-1.290	1.572	1.529	2.414
PM_{FA-UN} (Model 27)	1	0.433	0.146	-0.220	0.183	2.665	-0.306	-0.344	-0.613	0.179	0.139	-0.011	-0.146
	2	0.068	0.499	0.390	0.225	-0.695	1.930	0.888	0.686	0.027	0.204	0.968	0.425
	3	-0.099	0.189	0.469	0.114	-0.670	1.471	1.422	0.771	-0.003	0.268	0.377	0.441
	4	0.082	0.108	0.053	0.462	-1.222	1.165	1.124	1.494	-0.036	0.112	0.159	0.343

† Model 9: P_{FA-UN} = factor analytic-pedigree for G_{OP} and unstructured for R_0 ; Model 18: M_{FA-UN} = factor analytic-genomic for G_{OM} and unstructured for R_0 ; Model 27: PM_{FA-UN} = factor analytic-pedigree for G_{OP} , factor analytic-genomic for G_{OM} and unstructured for R_0 .

‡ Env. = Environment.

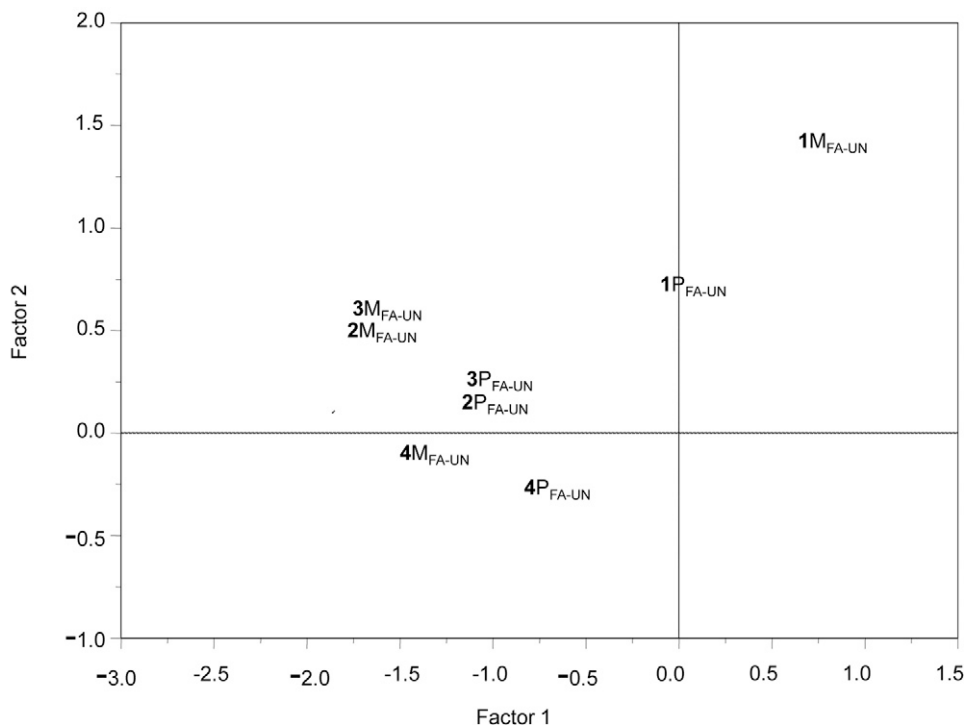


Figure 1. Estimated loadings (after varimax rotation) of each of the four environments on the first and second common factors of G_{OP} obtained from models using a pedigree-based (P_{FA-UN}) and marker-based (M_{FA-UN}) additive relationship matrix. The x–y coordinates are the estimated loadings after varimax rotation, and the numbers in the body of the plot are the environment labels (1–4). In both models, G_0 was modeled using a two-common factor model, and the residual covariance matrix, R_0 , was completely unstructured.

The residuals modeled using I, D, and UN had two effects that are confounded, the genetic variances that are represented by the additive \times 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 \times additive and their interactions with environments (Burgueño et al., 2007). Fitting the additive \times 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 \times additive epistasis and additive \times additive \times 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 0.317 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 PM_{D-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 (CV1) (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 1, 2, 3, and 4, and for Environments 2, 3, 4 together for 18 different models (Models 1–18) for two different cross-validation schemes (CV1 and CV2), each with 10-fold. For CV1 and CV2, the best predicted model among Models 1–9 and among Models 10–18 are underlined.

Model no.	CV1																	
	P _{I-I}	P _{D-I}	P _{FA-I}	P _{I-UN}	P _{D-UN}	P _{FA-UN}	M _{I-I}	M _{D-I}	M _{FA-I}	M _{I-D}	M _{D-D}	M _{FA-D}	M _{I-UN}	M _{D-UN}	M _{FA-UN}			
Environment 1	0.444	0.443	0.437	0.428	0.444	<u>0.447</u>	0.534	0.534	0.532	0.534	0.534	0.531	0.536	0.541	<u>0.550</u>			
Environment 2	0.416	0.416	0.410	0.365	0.350	<u>0.427</u>	<u>0.496</u>	0.495	0.467	0.496	0.495	0.463	0.466	0.450	0.490			
Environment 3	0.387	0.386	0.389	0.331	0.317	<u>0.397</u>	0.387	0.386	0.404	0.387	0.387	0.401	0.338	0.332	<u>0.406</u>			
Environment 4	0.441	0.441	0.432	0.402	0.409	0.435	0.438	0.438	0.433	0.438	0.437	0.428	0.420	0.417	<u>0.441</u>			
Environments E2–E4	0.414	0.413	0.409	0.413	0.405	<u>0.418</u>	0.439	0.438	0.432	0.439	0.439	0.428	0.408	0.397	<u>0.445</u>			
Environments E1–E4	0.420	0.420	0.415	0.408	0.408	<u>0.423</u>	0.464	0.464	0.457	0.465	0.465	0.452	0.442	0.434	<u>0.473</u>			
	CV2																	
Environment 1	0.449	0.449	<u>0.460</u>	0.449	0.447	0.454	0.550	0.550	0.552	0.550	0.550	0.553	0.556	<u>0.560</u>	0.558			
Environment 2	0.422	0.421	0.623	0.422	0.421	<u>0.626</u>	0.502	0.500	0.609	0.502	0.500	<u>0.611</u>	0.511	0.490	0.606			
Environment 3	0.374	0.374	0.633	0.374	0.374	<u>0.635</u>	0.392	0.391	0.581	0.392	0.390	<u>0.585</u>	0.384	0.368	0.556			
Environment 4	0.452	0.452	<u>0.533</u>	0.452	0.452	0.529	0.453	0.452	<u>0.513</u>	0.453	0.452	0.510	0.445	0.439	0.506			
Environments E2–E4	0.414	0.413	0.596	0.413	0.410	<u>0.596</u>	0.454	0.453	<u>0.571</u>	0.455	0.453	<u>0.571</u>	0.452	0.435	0.559			
Environments E1–E4	0.420	0.419	<u>0.565</u>	0.419	0.416	0.564	0.475	0.473	0.564	0.475	0.473	<u>0.565</u>	0.475	0.462	0.556			

[†] Model code: Model 1: P_{I-I} = identity-pedigree for G_{op} and identity for R_{ot} ; Model 2: P_{D-I} = diagonal-pedigree for G_{op} and identity for R_{ot} ; Model 3: P_{FA-I} = factor analytic-pedigree for G_{op} and identity for R_{ot} ; Model 4: P_{I-UN} = identity-pedigree for G_{op} and diagonal for R_{ot} ; Model 5: P_{D-UN} = diagonal-pedigree for G_{op} and diagonal for R_{ot} ; Model 6: P_{FA-UN} = factor analytic-pedigree for G_{op} and diagonal for R_{ot} ; Model 7: P_{I-UN} = identity-pedigree for G_{om} and unstructured for R_{ot} ; Model 8: P_{D-UN} = diagonal-pedigree for G_{om} and unstructured for R_{ot} ; Model 9: P_{FA-UN} = factor analytic-pedigree for G_{om} and unstructured for R_{ot} ; Model 10: M_{I-I} = identity-genomic for G_{om} and identity for R_{ot} ; Model 11: M_{D-I} = diagonal-genomic for G_{om} and identity for R_{ot} ; Model 12: M_{FA-I} = factor analytic-genomic for G_{om} and diagonal for R_{ot} ; Model 13: M_{I-UN} = identity-genomic for G_{om} and diagonal for R_{ot} ; Model 14: M_{D-UN} = diagonal-genomic for G_{om} and diagonal for R_{ot} ; Model 15: M_{FA-UN} = factor analytic-genomic for G_{om} and diagonal for R_{ot} ; Model 16: M_{I-UN} = identity-genomic for G_{om} and unstructured for R_{ot} ; Model 17: M_{D-UN} = diagonal-genomic for G_{om} and unstructured for R_{ot} ; Model 18: M_{FA-UN} = factor analytic-genomic for G_{om} and unstructured for R_{ot} .

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.

Model no.	19	20	21	22	23	24	25	26	27
Model code†	PM_{I-I}	PM_{D-I}	PM_{FA-I}	PM_{I-D}	PM_{D-D}	PM_{FA-D}	PM_{I-UN}	PM_{D-UN}	PM_{FA-UN}
CV1									
Environment 1	0.562	0.559	0.542	0.563	0.560	0.551	0.560	<u>0.564</u>	0.560
Environment 2	0.502	<u>0.504</u>	0.469	0.502	0.503	0.482	0.469	0.458	0.495
Environment 3	0.432	<u>0.433</u>	0.432	0.433	0.433	0.432	0.380	0.374	0.423
Environment 4	0.479	<u>0.481</u>	0.444	0.479	0.480	0.453	0.457	0.462	0.462
Environments E2–E4	0.471	<u>0.474</u>	0.445	0.471	0.472	0.454	0.437	0.431	0.459
Environments E1–E4	0.494	<u>0.496</u>	0.4722	0.494	0.495	0.480	0.469	0.464	0.486
CV2									
Environment 1	0.572	0.570	<u>0.580</u>	0.573	0.570	0.580	0.574	0.577	0.574
Environment 2	0.507	0.505	<u>0.646</u>	0.507	0.502	<u>0.646</u>	0.512	0.497	0.642
Environment 3	0.428	0.424	0.631	0.427	0.424	<u>0.632</u>	0.414	0.400	0.604
Environment 4	0.495	0.496	<u>0.548</u>	0.495	0.496	0.545	0.481	0.487	0.544
Environments E2–E4	0.480	0.479	<u>0.610</u>	0.480	0.477	0.609	0.475	0.465	0.599
Environments E1–E4	0.502	0.500	<u>0.603</u>	0.502	0.498	0.602	0.498	0.4897	0.592

†Model code: Model 19: PM_{I-I} = identity-pedigree (G_{OP}), identity-genomic (G_{OM}) with identity for R_0 ; Model 20: PM_{D-I} = diagonal-pedigree (G_{OP}), diagonal-genomic for (G_{OM}) with identity for R_0 ; Model 21: PM_{FA-I} = factor analytic-pedigree (G_{OP}), factor analytic-genomic for (G_{OM}) with identity for R_0 ; Model 22: PM_{I-D} = identity-pedigree (G_{OP}), identity-genomic (G_{OM}) with diagonal for R_0 ; Model 23: PM_{D-D} = diagonal-pedigree (G_{OP}), diagonal-genomic for (G_{OM}) with diagonal for R_0 ; Model 24: PM_{FA-D} = factor analytic-pedigree (G_{OP}), factor analytic-genomic for (G_{OM}) with diagonal for R_0 ; Model 25: PM_{I-UN} = identity-pedigree (G_{OP}), identity-genomic (G_{OM}) with unstructured for R_0 ; Model 26: PM_{D-UN} = diagonal-pedigree (G_{OP}), diagonal-genomic for (G_{OM}) with unstructured for R_0 ; Model 27: PM_{FA-UN} = factor analytic-pedigree (G_{OP}), factor analytic-genomic for (G_{OM}) with unstructured for R_0 .

(CV2). This is reflected in the results depicted in Fig. 2: averaged across environments, CV-correlations obtained from multienvironment models (P_{FA-UN} , M_{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, but not in E1 (Fig. 3). As shown in Table 2, genetic values in E2–E4 have high genetic correlations, while genetic values in E1 exhibit low genetic correlations with those from E2–E4. Models FA-D and FA-UN capitalize on these genetic correlations. For correlated environments E2–E4, the benefits in predictive ability come from two sources: from borrowing information from correlated environments by means of modeling GE and from using information regarding pedigree

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_{OP} , G_{OM} , 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 $y = X\beta + Zg + \epsilon$, where X is the incidence matrix for the fixed effects of environments β , and Z is the incidence matrix for the random effects of the genotypes g with $Cov(g, g') = \sigma_g^2 I_g$, and the random residual (ϵ) is $Cov(\epsilon, \epsilon') = \sigma_\epsilon^2 I_\epsilon$. In this model, the GE contains some of the nonadditive genetic effects, the GE effects, and parts of the error. Model II is $y = X\beta + Zg + \epsilon$, with the fixed effects of environments β and the random

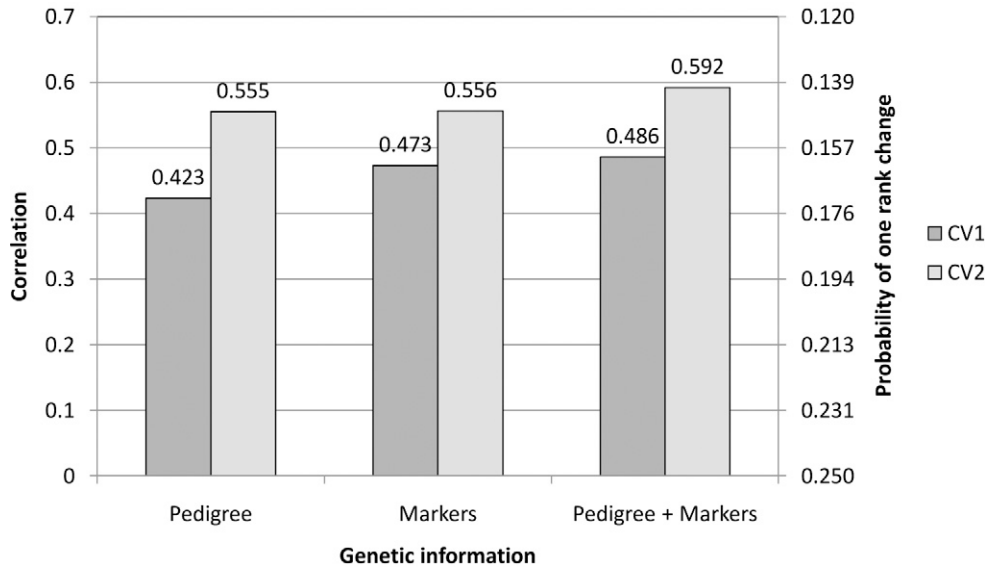


Figure 2. Average (across four environments) correlations between predicted and observed performance derived from models using only pedigree (Model 9 P_{FA-UN}), only markers (Model 18 M_{FA-UN}), or pedigree + markers (Model 27 PM_{FA-UN}) for two cross-validation schemes (CV1 and CV2). The right vertical axis shows the probabilities of one rank change for the given correlations.

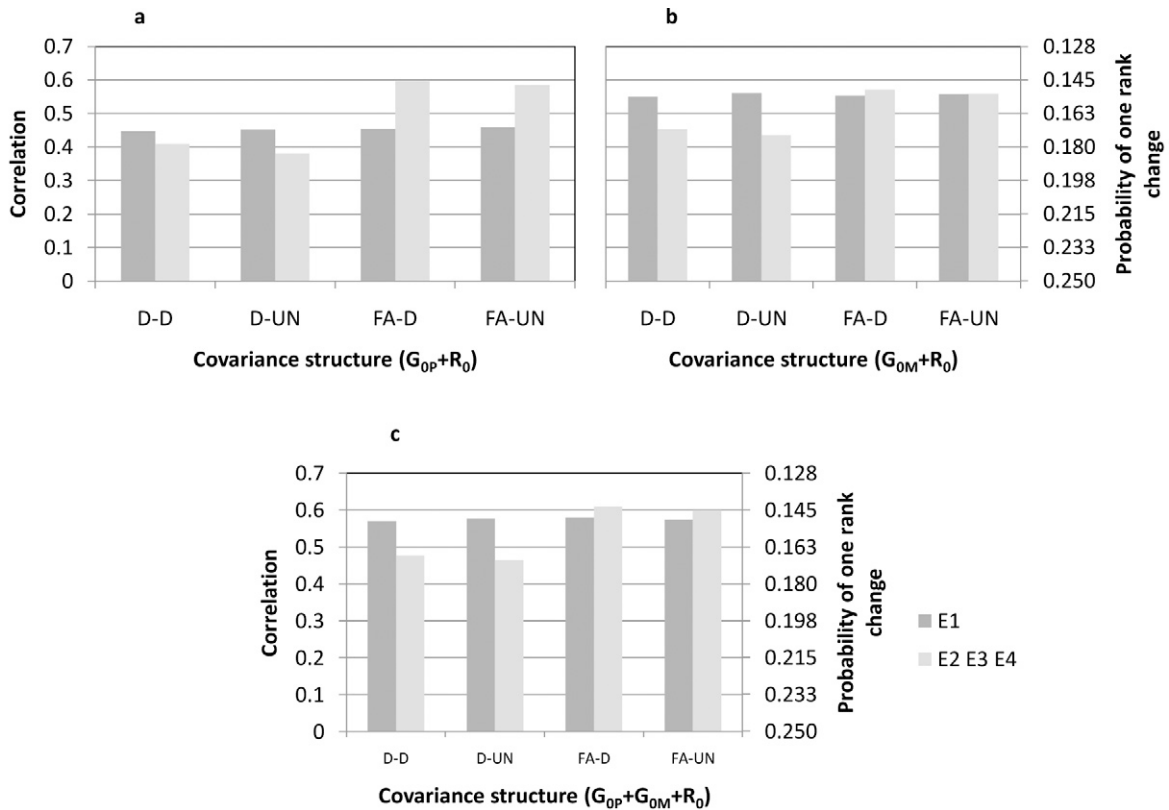


Figure 3. Correlation between predicted and observed performance in Environment 1 (E1) and average of Environments 2, 3, and 4 (E2, E3, and E4) obtained in CV2 using (a) pedigree, (b) markers, and (c) pedigree and marker-based models with different specifications for the residual and genetic covariance matrices (D = Diagonal, FA = two-common factors model, UN = unstructured). The right vertical axis shows the probabilities of one rank change for the given correlations.

effects \mathbf{g} that contain the genetic effects, the GE effects, and the error term, ϵ ; the random \mathbf{g} effects are modeled using the FA such that $\text{Cov}(\mathbf{g}, \mathbf{g}') = (\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}) \otimes \mathbf{I}_g = \text{FA}(k) \otimes \mathbf{I}_g$.

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 multi-environment models. However, multi-environment 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 multi-environment 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, $g_{ij} = \sum_{l=1}^L m_{il} b_{lj}$, where m_{il} counts the number of copies of the minor-frequency allele carried by the i th line at the l th locus, and b_{lj} represents the effect of the l 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 $\mathbf{M} = \{m_{il}\}$ is a matrix of marker genotypes, and $\mathbf{b}_j = (b_{1j}, \dots, b_{Lj})$ is a vector of marker effects. By stacking the genetic values of all lines in all environments into a single vector, we obtained $\mathbf{g} = [\mathbf{I}_J \otimes \mathbf{M}]\mathbf{b}$, where $\mathbf{g} = (\mathbf{g}'_1, \dots, \mathbf{g}'_J)$ and $\mathbf{b} = (\mathbf{b}'_1, \dots, \mathbf{b}'_J)$.

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

$$\text{Cov}(b_{lj}, b_{l'j'}) = \begin{cases} G_{j'j} & \text{if } l = l' \\ 0 & \text{otherwise} \end{cases}$$

In matrix notation this is expressed as $\text{Cov}(\mathbf{b}) = \mathbf{G}_0 \otimes \mathbf{I}_L$, where $\mathbf{G}_0 = \{b_{lj}, b_{l'j'}\}$ is a within-locus, across-environments covariance matrix of marker effects. Therefore,

$$\mathbf{b} \sim \text{MVN}(\mathbf{0}, \mathbf{G}_0 \otimes \mathbf{I}_L) \quad [1A]$$

The vector of genetic values, $\mathbf{g} = [\mathbf{I}_J \otimes \mathbf{M}]\mathbf{b}$, 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, $E[\mathbf{g}] = [\mathbf{I}_J \otimes \mathbf{M}]E[\mathbf{b}] = \mathbf{0}$ and

$$\begin{aligned} \text{Cov}(\mathbf{g}, \mathbf{g}') &= (\mathbf{I}_J \otimes \mathbf{M})\text{Cov}\{\mathbf{b}, \mathbf{b}'\}(\mathbf{M}' \otimes \mathbf{I}_J) \\ &= (\mathbf{I}_J \otimes \mathbf{M})(\mathbf{G}_0 \otimes \mathbf{I}_L)(\mathbf{M}' \otimes \mathbf{I}_J) \\ &= (\mathbf{G}_0 \otimes \mathbf{M})(\mathbf{M}' \otimes \mathbf{I}_J) \\ &= \mathbf{G}_0 \otimes (\mathbf{M}\mathbf{M}') \end{aligned}$$

Therefore, using $\mathbf{A} = \mathbf{M}\mathbf{M}'$ 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 $\text{Cov}(\mathbf{b}) = \mathbf{G}_0 \otimes \mathbf{I}_L$.

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 $i = 1, 2, \dots, g$) 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:

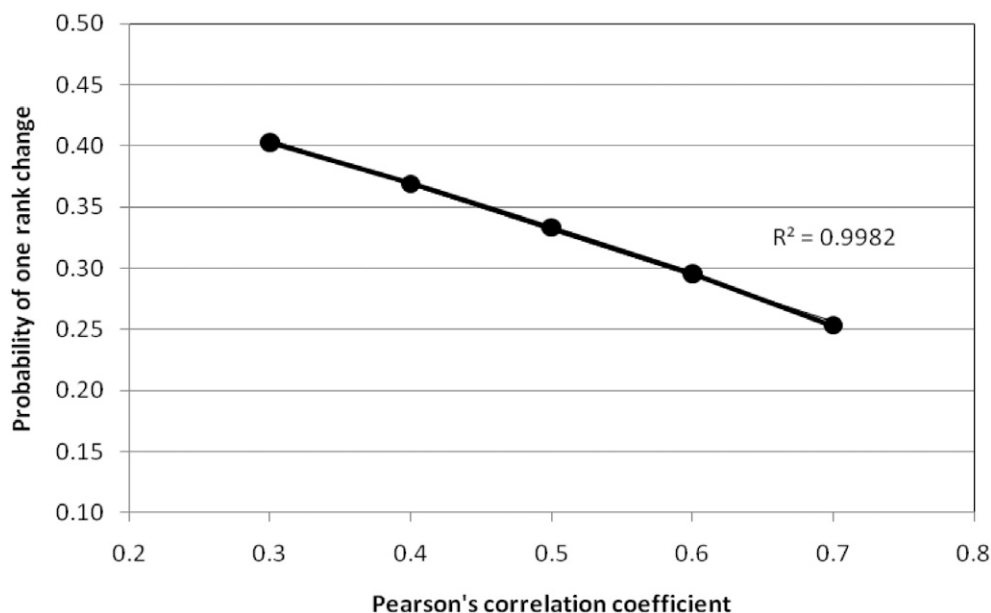


Figure B1. Relationship between Pearson's correlation coefficient and the probability that there will be a rank change between any two given genotypes if the observed or the predicted genotypic value is used.

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

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, $P[(X_1 - X_2)(Y_1 - Y_2) < 0] = \frac{1 - \tau}{2}$. Given

the relationship between Kendall's correlation coefficient

and Pearson's correlation coefficient (ρ), $\tau = \frac{\arcsin(\rho)}{\frac{\pi}{2}}$, the

probability that there will be a rank change between any two genotypes when using the observed or the predicted genotypic value is

$$P[(X_1 - X_2)(Y_1 - Y_2) < 0] = \frac{1 - \frac{\arcsin(\rho)}{\frac{\pi}{2}}}{2}$$

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

Table C1. Example of one-fold Cross-validation Scheme 1 (CV1) and Cross-validation Scheme 2 (CV2). The numbers in the table represent the actual standardized grain yield data of 599 genotypes in four environments (E1–E4). Missing genotypes are denoted as dots.

Genotype	CV1				CV2			
	E1	E2	E3	E4	E1	E2	E3	E4
1	1.7	-1.7	-1.9	0.1	1.7	-1.7	-1.9	0.1
2	0.3	-1.7
3	0.3	-0.6	-0.8	-1.1	0.3	-0.6	-0.8	-1.1
4	0.8	0.1	0.6	0.6	0.8	0.1	0.6	0.6
5	1.0	-0.3	1.6	-0.1	1.0	-0.3	1.6	-0.1
6	2.3	0.6	0.1	0.7	2.3	0.6	.	.
7	0.6	-0.3	0.1	0.0
8	0.6	-0.4	-0.7	1.0	0.6	-0.4	-0.7	1.0
9	-1.0	-1.8	-1.9	-1.5	.	-1.8	.	-1.5
10	-1.1	-1.6	-2.0	-0.6	-1.1	-1.6	-2.0	-0.6
11	1.7	-0.3	-0.2	0.3	1.7	-0.3	-0.2	0.3
12	0.9	-0.2	-0.4	0.7
13	1.2	-0.2	0.8	-0.2	.	-0.2	0.8	.
14	0.3	-0.7	-0.4	0.0	0.3	-0.7	-0.4	0.0
15	0.8	-0.6	-0.7	-0.2	0.8	-0.6	-0.7	-0.2
...
589	-0.8	0.7	0.7	2.6	-0.8	0.7	0.7	2.6
590	-0.9	0.1	1.1	2.8	-0.9	0.1	1.1	2.8
591	-1.1	.	2.0	.
592	-1.0	1.1	1.9	1.7	-1.0	1.1	1.9	1.7
593	-1.1	0.6	2.4	1.2	-1.1	0.6	2.4	1.2
594	-0.5	.	.	1.8
595	-1.2	1.4	1.6	1.7	-1.2	1.4	1.6	1.7
596	-1.1	0.1	2.1	0.6	-1.1	0.1	2.1	0.6
597	-1.2	0.5	2.0	1.8	.	0.5	.	1.8
598	-0.9	0.5	1.7	2.7
599	0.2	-0.5	-0.3	0.4	0.2	-0.5	-0.3	0.4

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

- ASReml. 2010. Version 3. VSN International Ltd., Hemel Hempstead, UK.
- Bernardo, R., and J. Yu. 2007. Prospects for genome-wide selection for quantitative traits in maize. *Crop Sci.* 47:1082–1090. doi:10.2135/cropsci2006.11.0690
- Braun, H.-J., S. Rajaram, and M. van Ginkel. 1996. CIMMYT's approach to breeding for wide adaptation. *Euphytica* 92:175–183. doi:10.1007/BF00022843
- Burgueño, J., J. Crossa, P.L. Cornelius, R. Trethowan, G. McLaren, and A. Krishnamachari. 2007. Modeling additive \times environment and additive \times additive \times environment using genetic covariances of relatives of wheat genotypes. *Crop Sci.* 47:311–320. doi:10.2135/cropsci2006.09.0564
- Burgueño, J., J. Crossa, P.L. Cornelius, and R.-C. Yang. 2008. Using factor analytic models for joining environments and genotypes without crossover genotype \times environment interaction. *Crop Sci.* 48:1291–1305. doi:10.2135/cropsci2007.11.0632
- Burgueño, J., J. Crossa, J. Miguel Cotes, F. San Vicente, and B. Das. 2011. Prediction assessment of linear mixed models for multi-environment trials. *Crop Sci.* 51:944–954. doi:10.2135/cropsci2010.07.0403
- Crossa, J., J. Burgueño, P.L. Cornelius, G. McLaren, R. Trethowan, and A. Krishnamachari. 2006. Modeling genotype \times environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Sci.* 46:1722–1733. doi:10.2135/cropsci2005.11-0427
- Crossa, J., G. de los Campos, P. Pérez, D. Gianola, S. Dreisigacker, J. Burgueño, J.L. Araus, D. Makumbi, J. Yan, R. Singh, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186:713–724. doi:10.1534/genetics.110.118521
- Crossa, J., P. Pérez, G. de los Campos, G. Mahuku, S. Dreisigacker, and C. Magorokosho. 2011. Genomic selection and prediction in plant breeding. *J. Crop Improv.* 25:239–261. doi:10.1080/15427528.2011.558767
- Crossa, J., R.-C. Yang, and P.L. Cornelius. 2004. Studying crossover genotype \times environment interaction using linear-bilinear models and mixed models. *J. Agric. Biol. Environ. Stat.* 9:362–380. doi:10.1198/108571104X4423
- de los Campos, G., and D. Gianola. 2007. Factor analysis models for structuring covariance matrices of additive genetic effects: A Bayesian implementation. *Genet. Sel. Evol.* 39:481–494. doi:10.1186/1297-9686-39-5-481
- de los Campos, G., D. Gianola, G.J.M. Rosa, K.A. Wiegand, and J. Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* 92:295–308. doi:10.1017/S0016672310000285
- de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182:375–385. doi:10.1534/genetics.109.101501
- de los Campos, G., and P. Pérez. 2010. BLR: Bayesian linear regression. Version 1.2. Available at <http://cran.r-project.org/web/packages/BLR/index.html> (verified 9 Nov. 2011).
- Fisher, R.A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* 52:399–433.
- Gianola, D., and D. Sorensen. 2004. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics* 167:1407–1424. doi:10.1534/genetics.103.025734
- Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genomic relationship information on genome-assisted breeding value. *Genetics* 177:2389–2397.
- Hayes, B., and M. Goddard. 2010. Genome-wide association and genomic selection in animal breeding. *Genome* 53:876–883. doi:10.1139/G10-076
- Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31:423–447. doi:10.2307/2529430
- Kirkpatrick, M., and K. Meyer. 2004. Direct estimation of genetic principal components: Simplified analysis of complex phenotypes. *Genetics* 168:2295–2306. doi:10.1534/genetics.104.029181
- Lynch, M., and K. Ritland. 1999. Estimation of pairwise relatedness with molecular markers. *Genetics* 152:1753–1766.
- Makowsky, R., N.M. Pajewski, Y.C. Klimentidis, A.I. Vazquez, C.W. Duarte, D.B. Allison, and G. de los Campos. 2011. Beyond missing heritability: Prediction of complex traits. *PLoS Genet.* 7:e1002051. doi:10.1371/journal.pgen.1002051
- 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. doi:10.1104/pp.105.063438
- 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.
- Meyer, K., and M. Kirkpatrick. 2005. Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. *Genet. Sel. Evol.* 37:1–30. doi:10.1186/1297-9686-37-1-1
- 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. doi:10.1007/s00122-006-0333-z
- 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 Gen.* 3:106–116. doi:10.3835/plantgenome2010.04.0005
- Piepho, H.P. 1997. Analyzing genotype–environment data by mixed models with multiplicative effects. *Biometrics* 53:761–766. doi:10.2307/2533976
- Piepho, H.P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor analytic variance covariance structure. *Theor. Appl. Genet.* 97:195–201. doi:10.1007/s001220050885
- Piepho, H.P. 2009. Ridge regression and extensions for genome-wide selection in maize. *Crop Sci.* 49:1165–1176. doi:10.2135/cropsci2008.10.0595
- Ritland, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet. Res.* 67:175–185. doi:10.1017/S0016672300033620
- Smith, A., B.R. Cullis, and R. Thompson. 2002. Exploring variety–environment data using random effects models with adjustment for spatial field trends: Part 1: Theory. *Quantitative genetics, genomics and plant breeding*. CAB Int., Oxford, UK.
- Smith, A.B., B.R. Cullis, and R. Thompson. 2001. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57:1138–1147. doi:10.1111/j.0006-341X.2001.01138.x
- So, Y.-S., and J. Edwards. 2011. Predictive ability assessment of linear mixed models in multienvironment trials in corn. *Crop Sci.* 51:542–552. doi:10.2135/cropsci2010.06.0338
- Van Raden, P.M. 2007. Genomic measures of relationship and inbreeding. *Interbull Bull.* 37:33–36.
- Welham, S.J., B.J. Gogel, A.B. Smith, R. Thompson, and B.R. Cullis. 2010. A comparison of analysis method for late-stage variety evaluation trials. *Aust. N. Z. J. Stat.* 52:125–149. doi:10.1111/j.1467-842X.2010.00570.x
- Wright, S. 1921. Systems of mating. 1. The biometric relations between parent and offspring. *Genetics* 6:111–123.