From Genotype × Environment Interaction to Gene × Environment Interaction

Jose Crossa*

Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico

Abstract: Historically in plant breeding a large number of statistical models has been developed and used for studying genotype × environment interaction. These models have helped plant breeders to assess the stability of economically important traits and to predict the performance of newly developed genotypes evaluated under varying environmental conditions. In the last decade, the use of relatively low numbers of markers has facilitated the mapping of chromosome regions associated with phenotypic variability (e.g., QTL mapping) and, to a lesser extent, revealed the differetial response of these chromosome regions across environments (i.e., QTL × environment interaction). QTL technology has been useful for marker-assisted selection of simple traits; however, it has not been efficient for predicting complex traits affected by a large number of loci. Recently the appearance of cheap, abundant markers has made it possible to saturate the genome with high density markers and use marker information to predict genomic breeding values, thus increasing the precision of genetic value prediction over that achieved with the traditional use of pedigree information. Genomic data also allow assessing chromosome regions through marker effects and studying the pattern of covariablity of marker effects across differential environmental conditions. In this review, we outline the most important models for assessing genotype × environment interaction, QTL × environment interaction, and marker effect (gene) × environment interaction. Since analyzing genetic and genomic data is one of the most challenging statistical problems researchers currently face, different models from different areas of statistical research must be attempted in order to make significant progress in understanding genetic effects and their interaction with environment.

Received on: May 20, 2011 - Revised on: September 30, 2011 - Accepted on: October 09, 2011

Keywords: Genotype × environment interaction (GE), gene × environment interaction, environmental and genotypic covariables, Quantitative Trait Loci (QTL), molecular markers (MM), Genomics-enable prediction and selection.

INTRODUCTION

The presence of genotype × environment interaction (GE) in plant breeding multi-environment trials (MET) is expressed either as inconsistent responses of some genotypes relative to others due to genotypic rank change or as changes in the absolute differences between genotypes without rank change (i.e., heterogeneity of within-site variance). Several models are commonly used for describing the mean response of genotypes across environments and for studying and interpreting GE in agricultural experiments: linear models, bilinear models, and linear-bilinear models. Fixed-effect linear-bilinear models, such as the Sites Regression (SREG) [1, 2] and the Additive Main effect and Multiplicative Interaction (AMMI) models [3, 4] are used for studying genotypic response patterns across environments. In these models, the response patterns of genotypes and environments can be visualized graphically using biplots [5, 6] that allow the breeder to observe the high performing genotype(s) in a region(s) and/or sub-region(s). Recently, several review articles pointed out the merits and demerits of the fixed-effect linear-bilinear models [7-11]. One class of fixed-effect linear models, namely factorial regression (FR) models, and one class of bilinear models, namely partial least squares (PLS) regression, allow incorporating external environmental and genotypic covariables directly into the model and are useful for finding the climatic causes of GE or the genetic factors (molecular markers) influencing GE.

Linear mixed models have become widely accepted and used for analyzing MET in plant breeding [12-19]. The models naturally lead to a factor analytic (FA) [20, 21] form of the genetic variance-covariance for environments that is more parsimonious and flexible than other variance-covariance structures. Since these are linear-bilinear mixed models, they also have the usual advantages, i.e., that error variance modeling can be accommodated (in particular, heterogeneity of block and error variance between environments and within-environment spatial correlation) and that incomplete data are handled with ease. Furthermore, when genotypes are considered as random effects, coefficients of parentage can be incorporated into the FA model for GE, thereby obtaining more precise estimates of the breeding values of genotypes [16, 22, 18]. Furthermore, Burgueño *et al.* [19] showed how to use the FA model and its biplot for clustering sites and genotypes with statistically negligible crossover interaction (COI). The method proposed by Burgueño *et al.* [19] has two main advantages: (1) the descriptive biplot is the starting point for testing successive hypotheses about the suitability of combining sites and genotypes into subsets that decrease the amount of

*Address correspondence to this author at the Biometrics and Statistics Unit, International Maize and Wheat ImprovementCenter (CIMMYT), Apdo. Post 6-641, 06600 Mexico, D.F., Mexico; Tel: 58042004; Fax: 58047559; E-mail: j.crossa@cgiar.org

1875-5488/12 $58.00+.00

significant COI, and (2) the process of delineating mega-environments and (3) answering the question "who wins where?" can be done within the framework of linear mixed models.

As for comparing the predictive ability of linear-bilinear fixed-effect models, the study of Cornelius and Crossa [23] used a cross-validation scheme that splits the data of a MET into a training set and a validation set, and computes the root mean squared predictive error between the predictive value and the observed value of the model. Those authors evaluated the within-trial predictive assessment of several fixed-effect linear-bilinear models, and developed efficient shrinkage estimates of fixed linear-bilinear models that showed the same or better predictive ability than the traditional Best Linear Unbiased Prediction (BLUP) of the cell mean. However, Cornelius and Crossa [23] only studied the predictive ability of models within specific environments (within-trial prediction) but did not assess the prediction of performance of unobserved genotypes in other environments (between-trial prediction). Also, Piepho [21] performed cross-validation on linear-bilinear models but in a slightly different form. Fixed linear-bilinear models have been used for describing GE or the combination of genotype plus GE interaction (GGE) based on a biplot graph. Recently, Burgueño *et al*. [24] compared the predictive ability of various linear-bilinear models and mixed effects models using the factor analytic model; results show that for data sets with complex GE or GGE, modeling GE and GGE using the FA model improved the predictability of the model up to 5-7%. When GE and GGE were not complex, most models gave high predictability (FA versus no FA) and FA did not seem to lose much predictability (only 2%). Therefore, it was concluded that modeling GE and GGE is a good thing.

Selection in plant breeding is usually based on estimates of breeding values obtained with pedigree-based mixed models. In their multivariate formulation, these models can also accommodate genotype × environment (GE) interaction. These models have been used successfully for predicting breeding values in plants and animals. However, pedigree-based models cannot account for Mendelian segregation, a term that, under an infinitesimal additive model (e.g., Fisher [25]) and in the absence of inbreeding, explains one half of the genetic variability. Molecular markers allow tracing Mendelian segregation at several positions of the genome, which gives them enormous potential in terms of increasing the accuracy of estimates of genetic values and the genetic progress attainable when these predictions are used for selection purposes.

The models mentioned above assess GE from an overall perspective, that is, they do not attempt to decompose or partition the total GE into chromosome regions or even further into specific genes so that the physical genetic causes of GE can be identified in the chromosome. Although Quantitative Trait Loci (QTL) mapping has been routinely used in plant breeding, approaches that fully exploit data from MET to assess and study QTL × environment interaction (QEI) are very limited. The modeling of genetic (co)variances between environments, in combination with modeling of heterogeneous residuals, is an important condition for reaching reliable conclusions about the main effects of QTLs as well as the QEI that is caused by the different expressions of QTLs in different environments; linear mixed models are the natural framework for analyzing such complex data. Multi-environment QTL mapping approaches have been presented in the literature [26]. In these examples, single-trait QTL mapping is studied so that the problem is reduced to either a multi-trait or a multi-environment dimension, but not both. The QTL linear mixed model can be extended to cover both multi-trait and multi-environment cases by first identifying the genetic correlation between traits and/or environments and by imposing some structure on the (co)variance matrix, and then incorporating molecular marker information to extend the phenotypic model into the QTL model [27]. Marker-assisted selection (MAS) and the identification of molecular markers closely linked to QTLs [28] have been widely used in plant breeding to improve a few traits controlled by major genes. However, adoption of the technology has been limited because the bi-parental populations used for mapping QTLs are not easily used in breeding applications. Also, since MAS uses only partial information (few markers), it presents limitations for improving traits controlled by many loci with small effects because the few markers linked to significant QTLs explain only a small percentage of the total genetic variability (the problem of missing heritability).

On the other hand, genomic selection (GS) (or genome-wide selection) is an approach for improving quantitative traits [29] that uses all available molecular markers (MMs) across the genome to estimate genetic values. Reports on the use of GS in plants are few and refer mainly to computer simulation studies such as the research of Bernardo and Yu [30], who concluded that GS is superior to marker-assisted selection in maize. In recent articles, de los Campos *et al*. [31], Crossa *et al*. [32, 33], and Perez *et al*. [34] used Bayesian estimates from genomic parametric and semi-parametric regression and showed that models using MMs produced more accurate predictions of grain yield and other traits in maize and wheat than those based only on pedigree. Genomic selection has been validated in animal breeding for predicting breeding values [35, 36, 31].

In a usual genetic model, the phenotypic response of the i^{th} individual (y_i) is described as the sum of a genetic value, g_i, and a model residual, ε_i, such that the linear model for the genotypes (i=1,2,..,n) is represented as $y_i = \mu + g_i + \varepsilon_i$ (where μ is the general mean). One method for incorporating markers in models for GS is to define g_i as a parametric regression on marker covariates x_{ij} (which can take values of 1, 0, and -1 for a biallelic marker of a segregating population or values of 1 and -1 for inbred lines) of the form $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$, such that $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$ (j=1,2,…,p), where β_j is the regression of y_i on the j^{th} marker covariate. In matrix notation, the model is expressed as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Usually, the number of markers exceeds the number of individuals, and estimation of marker effects *via* ordinary least squares (OLS) is not feasible. In OLS, estimates are obtained to maximize model goodness-of-fit to the training set, and model complexity is not considered. When the number of MMs is large, this typically yields high mean-squared error of estimates of marker effects and poor predictive ability.

In the rest of the chapter I will show the basic fixed and mixed linear-bilinear models underlying GE, the models employed to develop QTL mapping and QTL×E interaction assessment, and the basic models for genomic-enable prediction that allow assessing gene × environment interaction..

STATISTICAL MODELS FOR ASSESSING GENOTYPE × ENVIRONMENT INTERACTION

The Basic Model

The basic two-way fixed effects linear model for GE analyses considers that the empirical mean response, $\bar{y}_{ij}$, of the i^th^ genotype (i=1,2,…,g) in the j^th^ environment (j=1,2,…,s) with r replications in each of the g×s cells is expressed as

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + (\tau\delta)_{ij} + \bar{\varepsilon}_{ij} \tag{1}$$

where μ is the grand mean over all genotypes and environments, τ_i is the main effect of the ith genotype, δ_j is the main effect of the jth environment, $(\tau\delta)_{ij}$ is the effect of the interaction (GE) of the ith genotype in the jth environment, and $\bar{\varepsilon}_{ij}$ is the average error, assumed to be NID $(0, \sigma^2_{\varepsilon}/r)$ (where σ^2_{ε} is the within-environment error variance, assumed to be constant, and r is the number of observations per cell). For a complete random model, it is assumed that τ_i, δ_j, and $(\tau\delta)_{ij}$ are normally and independently distributed, with variances σ^2_{τ}, σ^2_{δ}, and $\sigma^2_{\tau\delta}$, respectively. Adding the design effects to (1) with randomized complete blocks, or any type of incomplete block design, does not pose a problema. Furthermore, modeling the residulas by means of spatial analyses does not present further difficulty and is a practice that must be routinely used in any field experiment.

Yates and Cochran [37] introduced a model in which the GE term is linearly related to the environmental main effect $(\tau\delta)_{ij} = \vartheta_i \bar{y}_{.j}$, such that the stability parameter ϑ_i is the regression of genotype performance on the environment mean $\bar{y}_{.j}$. This was later more formally presented by Finlay and Wilkinson [38] and extended by Eberhart and Russell [39] to include the deviation from regression as another stability parameter (although this is in fact a lack or fit of the model).

Fixed Effect Linear-Bilinear Models

Williams [40] considered the model $\bar{y}_{ij} = \mu + \tau_i + \lambda\alpha_i\gamma_j + \bar{\varepsilon}_{ij}$, where λ is the largest singular value of $\mathbf{ZZ}'$ and $\mathbf{Z}'\mathbf{Z}$ (for $\mathbf{Z} = \bar{y}_{ij} - \bar{y}_{i.}$) and α_i and γ_j are the corresponding eigenvectors. Gollob [41] and Mandel [42] extended Williams' [40] work by considering the bilinear GE term as $(\tau\delta)_{ij} = \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk}$. Thus, the general formulation of the linear-bilinear model is

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{k=1}^{t} \lambda_k \alpha_{ij} \gamma_{jk} + \bar{\varepsilon}_{ij} \tag{2}$$

where the constant λ_k is the singular value of the k^th^ bilinear (multiplicative) component that is ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t$; the α_{ik} are elements of the k^th^ left singular vector of the true interaction and represent genotypic sensitivity to hypothetical environmental factors represented by the k^th^ right singular vector with elements γ_{jk}. The α_{ik} and γ_{jk} satisfy the constraints $\sum_{i=1}^{g} \alpha_{ik}\alpha_{ik'} = \sum_{j=1}^{s} \gamma_{jk}\gamma_{jk'} = 0$ for k≠k′ and $\sum_i \alpha^2_{ik} = \sum_j \gamma^2_{jk} = 1$. Gabriel [43] described the least squares fit of Eq. 2 and explained how the the GE term, $\mathbf{Z} = \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$, is subjected to singular value decomposition (SVD) after adjusting for the additive (linear) terms. Gauch [3] called Eq. 2 the Additive Main effects and Multiplicative Interaction (AMMI) model.

Other classes of linear-bilinear models described by Cornelius *et al*. [1] are the Genotypes Regression Model (GREG)

$$\bar{y}_{ij} = \mu_i + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij},$$

the Sites (environments) Regression Model (SREG)

$$\bar{y}_{ij} = \mu_j + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij},$$

the Completely Multiplicative Model

$$\bar{y}_{ij} = \mu + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij},$$

(COMM), and the Shifted Multiplicative Model (SHMM)

$$\bar{y}_{ij} = \beta + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij},$$

In matrix notation, these linear-bilinear models can be expressed as $\mathbf{Y}=\sum_{k=1}^{m}\beta_k \mathbf{X}_k + \mathbf{A\Lambda G'}+\mathbf{E}$ [44], where $\mathbf{Y}=[\bar{y}_{ij}]$, $\mathbf{X}_k=[x_{kij}]$, $\mathbf{E}=[\bar{\varepsilon}_{ij}]$, $\Lambda=\text{diag}(\lambda_k, k=1,2,\ldots,t)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t$, $\mathbf{A}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_t)$, $\mathbf{G}=(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_t)$, and $\mathbf{A'A}=\mathbf{G'G}=\mathbf{I}_t$. The x_{kij} are known constants and β_k, λ_k, α_{ik}, and γ_{jk} are parameters to be estimated. It is possible to study GE by finding low dimensional approximations by means of the singular value decomposition of the structure present in the two-way table.

The GREG linear-bilinear model defined above is a reparameterization of the stability analysis model of Finlay and Wilkinson [38] and the Eberhart and Russell [39] models that perform the linear regressions of genotype on environment means. In the GREG model the first multiplicative term, $\lambda_1\alpha_{i1}\gamma_{j1}$ is perceived as the genotype regressions, with coefficients α_{i1} on environmental indices γ_{j1} (the scale parameter λ_1 can be absorbed into α_{i1} or γ_{j1} or partially into each), and the deviation modeled as multiplicative components, provided that t>1.

There are several statistical as well as biological reasons to prefer SREG over AMMI for assessing COI and non-COI under the common situation of complex GE: (1) for the same number of bilinear terms, SREG is a more parsimonious model than AMMI; (2) SREG incorporates the main effect of genotypes directly into the statistical analysis of GE, that is, both effects genotypes and GE (GGE) are combined and estimated jointly; this is important for reaching breeders' objectives, which requires including the main performance of genotypes in the model; (3) sometimes the mixed SREG model can be fitted much more easily than the mixed AMMI model; and (4) the mixed SREG model, as proved by Burgueño *et al.* [19], is useful for delineating mega-environments using a formal statistical approach based on the factor analytic model.

Mixed Effect Linear-Bilinear Models

What if Genotypes or Environments, or both, are Random Effects?

The basic linear mixed model used for fitting data from g genotypes, s sites, and r replicates when searching for subsets of environments and/or genotypes with non-COI is

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Z}_r\mathbf{r} + \mathbf{Z}_g\mathbf{g} + \mathbf{e}$$

where $\mathbf{X}$ is the incidence matrix of 0s and 1s for the fixed effects of environments, and $\mathbf{Z}_r$ and $\mathbf{Z}_g$ are the incidence matrices of 0s and 1s for the random effects of replicates within environments and genotypes within environments, respectively. The random effect of genotypes within environments combines the main effects of genotypes and GE (GGE). Vector $\mathbf{b}$ denotes the fixed effects of environments and/or the effect of the design (i.e., replicates, incomplete blocks, etc.); vectors $\mathbf{r}$, $\mathbf{g}$, and $\mathbf{e}$ contain random effects of replicates within environments, genotypes within environments, and residuals within environments, respectively, and are assumed to be random and normally distributed with zero mean vectors and variance-covariance matrices $\mathbf{R}$, $\mathbf{G}$, and $\mathbf{E}$, respectively. The variance-covariance matrices $\mathbf{R}$ and $\mathbf{E}$ are assumed to have the simple variance component structure, i.e., $\mathbf{R}=\text{diag}(\sigma^2_{r_j}, j=1,2,\ldots,s) \otimes \mathbf{I}_r$ and $\mathbf{E}=\sigma^2_e \otimes \mathbf{I}_{rgs}$, where $\mathbf{I}_r$ and $\mathbf{I}_{rgs}$ are identity matrices of order r and r×g×s, respectively, $\sigma^2_{r_j}$, σ^2_e are the replicates within the jth environment and residual variances, respectively, and $\otimes$ is the Kronecker (or direct) product of the two matrices. The structure of $\mathbf{E}$ assumes that the residuals of the field plots at each environment (i.e., elements of vector $\mathbf{e}$) are not spatially correlated; however, when field information is available, the spatial model approach using models such as the two-dimensional auto-regressive procedure in the direction of rows and columns in the field can be incorporated into the analyses. The solution ($\hat{\mathbf{b}}$) for the vector of fixed environment means and the vectors of random effects ($\hat{\mathbf{r}}$ and $\hat{\mathbf{g}}$) are obtained from the mixed model equations.

The variance-covariance matrix $\mathbf{G}$ is indexed by two factors, environments and genotypes, and can therefore be written as the Kronecker product of two matrices indexing those factors, $\mathbf{G}=\sum_g \otimes \mathbf{I}_g$, where the jth diagonal element of the $s \times s$ matrix $\boldsymbol{\Sigma}_g$ is the genetic variance $\sigma^2_{g_j}$ within the jth environment, and the jj'th off-diagonal element is the genetic covariance $\rho_{jj'}\sigma_{g_j}\sigma_{g_{j'}}$ between environments j and j'; thus $\rho_{jj'}$ is the correlation of genetic effects between environments j and j'. As for the genotype factor, the identity matrix $\mathbf{I}_g$ (of order g) is used when it is assumed that the genotypes are not related, and the breeding value of each genotype will be predicted only by the value of the empirical responses of the genotype itself. The environmental component of $\mathbf{G}$, $\boldsymbol{\Sigma}_g$, can be modeled by the FA, whereas the genotypic component of $\mathbf{G}$ is modeled by the identity matrix $\mathbf{I}_g$, which assumes no relationship among genotypes. However, a relationship matrix **A** using the coefficient of parentage among genotypes can be used instead.

The Factor Analytic model

The environmental component $\boldsymbol{\Sigma}_g$ of the variance of random effects, **G**, can be modeled by the FA model, which expresses the random effect of the ith genotype in the jth environment as a linear function of latent variables x_{ik} with coefficients δ_{jk} for k = (1, 2, … t), plus a residual η_{ij} that is

$$g_{ij} = \Sigma_{k=1}^{t} x_{ik}\delta_{jk} + \eta_{ij}$$

where δ_{jk} is the factor loading of the jth environment in the kth latent factor, x_{ik} is the score of the ith genotype in the kth latent factor, and η_{ij} is the residual term. In matrix form the previous equation is expressed as

$$g_{ij} = (\boldsymbol{\delta}_1 \otimes \mathbf{I}_g)\mathbf{x}_1 + (\boldsymbol{\delta}_2 \otimes \mathbf{I}_g)\mathbf{x}_2 + ... + (\boldsymbol{\delta}_k \otimes \mathbf{I}_g)\mathbf{x}_k + \boldsymbol{\eta}$$

where vector $(\boldsymbol{\delta}_k \otimes \mathbf{I}_g)\mathbf{x}_k$ is of order gs × 1, and vector $\boldsymbol{\delta}$ is of order gs × 1; then it can be written as

$$\mathbf{g} = (\boldsymbol{\Lambda} \otimes \mathbf{I}_g)\mathbf{x} + \boldsymbol{\eta},$$

where $\boldsymbol{\Lambda} = \begin{bmatrix} \delta_{11} & \delta_{12} & . & . & \delta_{1k} \\ \delta_{21} & \delta_{22} & . & . & \delta k \\ . & . & . & . & . \\ . & . & . & . & . \\ \delta_{s1} & \delta_{s2} & . & . & \delta_{sk} \end{bmatrix}$ is a matrix of order s×k

with the kth column containing the environment loadings for the kth latent factor. Since it is assumed that the genotypes are unrelated, random effects $\mathbf{x}$ and $\boldsymbol{\delta}$ are independent and have a joint normal distribution with a mean vector of zero and variances V($\mathbf{x}$)=$\mathbf{I}_k \otimes \mathbf{I}_g = \mathbf{I}_{kg}$ and V($\boldsymbol{\eta}$)=$\boldsymbol{\Psi} \otimes \mathbf{I}_g$ (of order sg×sg), respectively, where $\boldsymbol{\Psi}$ is a diagonal matrix $(\sigma^2_{\eta_1}, \sigma^2_{\eta_1}, ..., \sigma^2_{\eta_s})$ of order s×s. Therefore, the variance of the random effects **g,** which separates the environmental and genotypic components, is given by

$$\mathbf{G} = (\Lambda \otimes \mathbf{I}_g)\mathbf{V}(\mathbf{x})(\Lambda' \otimes \mathbf{I}_g) + \mathrm{V}(\delta)$$

$$= (\boldsymbol{\Lambda} \otimes \mathbf{I}_g)(\mathbf{I}_k \otimes \mathbf{I}_g)(\boldsymbol{\Lambda}' \otimes \mathbf{I}_g) + (\boldsymbol{\Psi} \otimes \mathbf{I}_g)$$

$$= (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) \otimes \mathbf{I}_g = \mathrm{FA(k)} \otimes \mathbf{I}_g$$

where the elements of the matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$

$$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' = \begin{bmatrix} \delta_{11} & \delta_{12} & . & . & \delta_{1k} \\ \delta_{21} & \delta_{22} & . & . & \delta_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ \delta_{s1} & \delta_{s2} & . & . & \delta_{sk} \end{bmatrix} \begin{bmatrix} \delta_{11} & \delta_{21} & . & . & \delta_{s1} \\ \delta_{12} & \delta_{22} & . & . & \delta_{s2} \\ . & . & . & . & . \\ . & . & . & . & . \\ \delta_{1k} & \delta_{2k} & . & . & \delta_{sk} \end{bmatrix}$$

are estimates of the genetic variance within the jth environment (i.e., diagonal elements of $\boldsymbol{\Sigma}_g$, $\sigma^2_{g_j}$) and estimates of the genetic covariances between the jth and the j’th environments (i.e., off-diagonal elements of $\boldsymbol{\Sigma}_g$, $\rho_{jj'}\sigma_{g_j}\sigma_{g_{j'}}$).

Thus, FA can be interpreted as the linear regression of genotype and GE on latent environmental covariates (environmental loadings, δ_{jk}), with each genotype having a separate slope (genotypic scores, x_{ik}) 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.

A mixed-model analogue of AMMI or SREG has been developed using the factor analytic (FA) model for approximating the variance-covariance GE structure [12, 17, 14, 15]. Research conducted by Crossa *et al.* [16] and Burgueño *et al.* [19] described how to model variance-covariance GE and GGE using the FA model and how to incorporate the additive (relationship A) matrix and the additive × additive covariance matrix into the FA model based on pedigree information. Burgueño *et al.* [19] also described the equivalence between SREG2 and FA(2) for finding subsets of genotypes and environments without COI.

The Relationship Between the Factor Analytic and Sites Regression Models for Assessing Crossover genotype × Environment Interaction

In the FA model, the random effect of the ith genotype in the jth environment (g_{ij}) is expressed as a linear function of latent variables x_{ik} with coefficients δ_{jk} for k=(1,2,…t), plus a residual, η_{ij}, i.e., $g_{ij} = \mu_j + \Sigma_{k=1}^{t} x_{ik}\delta_{jk} + \eta_{ij}$ so that the ijth cell mean can be written as $y_{ij} = g_{ij} + \varepsilon_{ij}$. With only the first two latent factors being retained, g_{ij} is approximated by $g_{ij} \approx \mu_j + x_{i1}\delta_{j1} + x_{i2}\delta_{j2} + \eta_{ij}$. Therefore, there is a clear connection between the SREG2 and FA(2) models. A similar connection between the AMMI2 and FA(2) models was established by Smith *et al.* [14].

Table 1. Linear Mixed Models Used for Comparing the Prediction of the Missing Genotypes in the Three Maize Trials (M1-MET, M2-MET, and M3-MET). The Overall Mean is μ (Adapted from Burgueño *et al.* [24])

Model	Fixed effects		Random effects			
1	μ	Site	rep(site)		site×genotype	error
2	μ		rep(site)	Site	site×genotype	error
3FA	μ	Site	rep(site)		site(FA)×genotype	error
4FA	μ		rep(site)	Site	site(FA)×genotype	error

Under principal component rotation, the directions and projections of the vectors of FA(2) and SREG2 in the biplot are the same. Therefore, the property of the SREG by which the first principal component of SREG2 accounts for non-crossover interaction (non-COI) and the second principal component of SREG2 is due to COI variability should hold for FA(2) as well. It should be pointed out that the absolute values of genotypic and environmental scores under the FA(2) and SREG2 models may not necessarily be the same; the estimates of the random effects in the FA(2) model are BLUPs (Best Linear Unbiased Predictions), whereas the estimates in the fixed effects SREG2 model are least squares estimates, that is, Best Linear Unbiased Estimates (BLUEs). Furthermore, the standard errors of the estimable functions of fixed effects under SREG differ from those of predictable functions of a mixture of fixed and random effects under FA, and FA models are more flexible in handling unbalanced data (the SREG model does not handle missing data).

Prediction of Unobserved Individuals Based on Phenotypic Data While Modeling GE

Burgueño *et al*. [24] compared the predictive ability of linear mixed models when the GE is modeled by the FA model with that of simple linear mixed models when the GE is not modeled. A cross-validation scheme is used that randomly deletes some genotypes from environments for a 10-fold partition; the values for these genotypes are then predicted by the different models and correlated with their observed values in order to assess model accuracy. A total of six multi-environment trials (one potato trial, three maize trials, and two wheat trials) with GE of varying complexity were used in the evaluation. Although Burgueño *et al*. [24] analyzed six MET, here we present results from only three maize METs (M1-MET, M2-MET, and M3-MET). The models used for prediction are linear mixed models 1 and 2, which are two simple mixed models that had the random GE term not modeled, and two linear models, 3FA and 4FA, which had the GE modeled by the FA model (Table **1**).

For M1-MET, the best predictive model was model 3FA, which had a correlation of 0.878 that represents a 6% increase in accuracy with respect to simple linear mixed model 1, whereas model 4FA, with a correlation of 0.867, had a 4.7% increase in accuracy with respect to the simple linear mixed model 1 (Table **2**). Prediction of the overall genotypic effects of the simple linear mixed models in a complex GE setting was poor as compared with that of models that considered covariances between environments. The correlations between the observed and predicted values of the four models fitted to M2-MET and M3-MET were higher in the M2-MET than in the M3-MET because the GE in M2-MET was simpler than the GE in M3-MET. For M2-MET, the two FA models (3FA and FA) did show slightly better predictability (0.938) than their non-FA counterparts (models 1 and 2) (0.916) (Table **2**), and there was a 2.4% increase in the predictive ability of these models with respect to simple linear model 1. For M3-MET, the best predictive

Table 2. Correlations Between the Predicted and Observed Values of the Missing Genotypes for Three Maize Trials (M1-MET, M2-MET, and M3-MET), Across *Fold* for Four Models (1, 2, 3FA, 4FA). Numbers in Parentheses for Models 2, 3FA, and 4FA Denote the % Change in Correlations with Respect to Model 1 (Adapted from Burgueño *et al.* [24])

	Model			
	1	**2**	**3FA**	**4FA**
------------------M1-MET (overall mean=4.90 Mg ha^{-1})------------------				
Across *fold*	0.828	0.827 (-0.2)	0.878 (6.0)	0.867 (4.7)
-----------------M2-MET (overall mean=5.04 Mg ha^{-1})------------------				
Across *fold*	0.916	0.916 (0)	0.938 (2.4)	0.938 (2.4)
-------------------M3-MET (overall mean= 5.66 Mg ha^{-1})----------------				
Across *fold*	0.824	0.824 (0)	0.848 (2.9)	0.852 (3.4)

models were 3FA and 4FA, with correlations of 0.848 and 0.852, respectively. However, the percentage increases in correlations of these models over model 1 were 2.9% and 3.4% for models 3FA and 4FA, respectively, indicating that for more complex GE (M3-MET), the FA models increase the predictability more than the simple linear mixed model for less complex GE (M2-MET) (Table **2**).

Incorporating Pedigree Information into Linear-Bilinear Mixed Models While Modeling GE

As has been shown, the linear mixed version of SREG and AMMI naturally leads to a FA form for the genetic variance-covariance for environments that is more parsimonious and flexible than other variance-covariance structures. Since the above mentioned models are linear mixed models, they also have the usual advantages when compared with ordinary fixed effects linear-bilinear AMMI and SREG models. That is, error variance modeling can be accommodated, in particular, heterogeneity of block and error variance between environments and within-environment spatial correlation, and incomplete data are handled with ease. Furthermore, when genotypes are considered as random effects, the coefficients of parentage can be incorporated into the FA for modeling GE or GGE of the mixed versions of AMMI and SREG, respectively, hence obtaining more precise estimates of the breeding values of genotypes. The genetic covariance between any pair of related individuals (i and i'), due to their additive genetic effects, is equal to two times the coefficient of parentage (COP=$\mathbf{f}_{ii'}$), also known as the coefficient of coancestry, times the additive genetic variance, i.e., $2\mathbf{f}_{ii'}\sigma_a^2=\mathbf{A}\sigma_a^2$, where **A** is the additive relationship matrix. In self-pollinated species, $\mathbf{A}\sigma_a^2$ is the variance-covariance matrix of the breeding values (additive genetic effects). Closely related individuals contribute more to the prediction of breeding values of their relatives than do less closely related genotypes. Moreover, when one genotype is missing (either partially or totally), its breeding value can still be predicted from its relatives, albeit less efficiently than if the data were complete.

A linear mixed model used for fitting the data from g genotypes, s sites and r replicates, assuming the relationship of the genotypes is measured by the matrix COP=$\mathbf{f}_{ii'}$ (of order g), is

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_s \end{bmatrix} = \begin{bmatrix} \mathbf{1}\mu_1 \\ \mathbf{1}\mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{1}\mu_s \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_{R_1} & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_{R_2} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & \cdot & \cdots & \mathbf{Z}_{R_s} \end{bmatrix}\mathbf{r}$$

$$+\begin{bmatrix} \mathbf{Z}_{G_1} & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_{G_2} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & \cdot & \cdots & \mathbf{Z}_{G_s} \end{bmatrix}\mathbf{g}+\mathbf{e}.$$

where $\mathbf{y}_j$ is the vector of the response variable in the jth site (j=1,2,…,s), **1** is a vector of ones, μ_j is the population mean of the jth site, and $\mathbf{Z}_{R_j}$ and $\mathbf{Z}_{G_j}$ are the design matrices of the random effects of replicates and genotypes within the jth site, respectively. The variance-covariance matrices **R** and **E** are assumed to have the simple variance component structure $\mathbf{R}=\mathbf{\Sigma}_r\otimes\mathbf{I}_r$ and $\mathbf{E}=\mathbf{\Sigma}_e\otimes\mathbf{I}_{rg}$, where $\mathbf{I}_r$ and $\mathbf{I}_{rg}$ are the identity matrices of orders r and r×g, respectively; the $\mathbf{\Sigma}_r=\text{diag}\ (\sigma^2_{r_j}, j=1,2,...,s)$ and $\mathbf{\Sigma}_e=\text{diag}\ (\sigma^2_{e_j}, j=1,2,...,s)$ are the $s\times s$ replicate and error variance-covariance matrices among pairs of s sites, respectively; $\sigma^2_{r_j}$, $\sigma^2_{e_j}$ are the replicate and residual variances within the jth site, respectively; and ⊗ is the Kronecker (or direct) product of the two matrices. In covariance pattern models, it is assumed that residuals have a multivariate normal distribution with zero means and covariance matrix **E**. In this study, the structure of **E** assumes that the residuals of the field plots at each site (i.e., elements of vector e) are not spatially correlated, that is, $\mathbf{E}=\mathbf{\Sigma}_e\otimes\mathbf{I}_{rg}$. However, when the field location of the plots is recorded, the matrix **E** could be modeled as a structure that is less restrictive than $\mathbf{\Sigma}_e\otimes\mathbf{I}_{rg}$. Commonly, these spatial correlations among field plots are modeled using the two-dimensional auto-regressive procedure in the direction of the rows and columns in the field. The design effects given by the replicates and the incomplete blocks within replicates in environments can be easily incorporated into the above linear mixed model as random or fixed effects.

Vectors $\mathbf{r}$, $\mathbf{g}$, and $\mathbf{e}$ contain random effects of replicates within sites, genotypes within sites, and residuals within sites, respectively, and are assumed to be random and normally distributed with zero mean vectors and variance-covariance matrices R, G, E, respectively, such that

$$\begin{bmatrix} \mathbf{r} \\ \mathbf{g} \\ \mathbf{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}' \begin{bmatrix} \mathbf{R} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{E} \end{bmatrix} \right).$$

The variance-covariance matrix **G**, which combines the main effect of genotypes (breeding values) and GE, can be represented as $\mathbf{G} = \mathbf{\Sigma}_g \otimes \mathbf{A}$, where the jth diagonal element of the s×s matrix $\mathbf{\Sigma}_g$ is the additive genetic variance $\sigma^2_{a_j}$ within the jth site, and the jj'th element is the additive genetic covariance $\rho_{jj'}\sigma_{a_j}\sigma_{a_{j'}}$ between sites j and j'; thus $\rho_{jj'}$ is the correlation of additive genetic effects between sites j and j'. The variance-covariance matrix **G** can be modeled using the FA structure. The matrix $\mathbf{A}$=2$f_{ii'}$ is of order g×g and measures the relationship or covariance between relatives due to additive genetic effects. When genotypes are not related, **A** is replaced by $\mathbf{I}_g$ (identity matrix of order g) [14, 45] and the breeding value of each genotype will be predicted only by the value of the empirical response of the genotype itself.

An example of how to incorporate pedigree information while modeling GE using the FA is described by Crossa *et al.* [16] using data from a CIMMYT bread wheat (*Triticum aestivum* L.) international trial. Twenty-nine lines (1-29) were tested in 16 international sites, namely Mexico (MEX), USA (two sites, USA1 and USA2), Turkey (TKY), Israel (ISR), Bangladesh (BGD), India (IND), Pakistan (PKT), Syria (SYR), Spain, (SPN) (two sites, SPN1 and SPN2), Nepal (NPL), Kenya (KNY), Zimbabwe (ZBW), New Zealand (NZL), and Chile (CHL), in randomized complete block designs with three replications at each site. The response variable analyzed was grain yield (Mg ha^{-1}). There were five sets of sister lines (5, 6), (4, 19, 20), (14, 15), (21, 23, 24), and (28, 29).

Standard errors (SE) of BLUPs of breeding values of the lines for grain yield were smaller for models using information on relatives and when the main effect of genotypes and GE were modeled using FA structure of the **G** variance-covariance matrix (Fig. **1**). The standard fixed linear-bilinear model 1 where all the effects are considered as fixed effects had the largest SEs, followed by a simple mixed linear-bilinear model 2 (with the GE considered as random effect) without **A** (without modeling GE) and a simple mixed linear-bilinear model 3 with **A** (without modeling GE). In contrast, the mixed linear-bilinear model 4 that models the GE using the FA and includes information on relatives (**A**) had the smallest SE for most of the line-environment combinations. The lines in Fig. (**1**) (from top to bottom) graph the SEs of BLUPs of breeding values of the 29 genotypes computed from models 1, 2, 3, and 4 in each of the 16 sites included in the trial (NZL, USA1, TKY, ZBW, CHL, SPA2, MEX, SYR, ISR, SPA1, NPL, KNY, IND, PKT, BGD, and USA2). New Zealand (NZL) had the largest SE of the BLUPs of the breeding values and USA2 had the smallest. Only the SE of the BLUP of genotype 26 in NZL from model 4 was larger than those from models 2 and 3 (Fig. **1**). The benefits, in terms of precision, of including information on related lines are evident when comparing

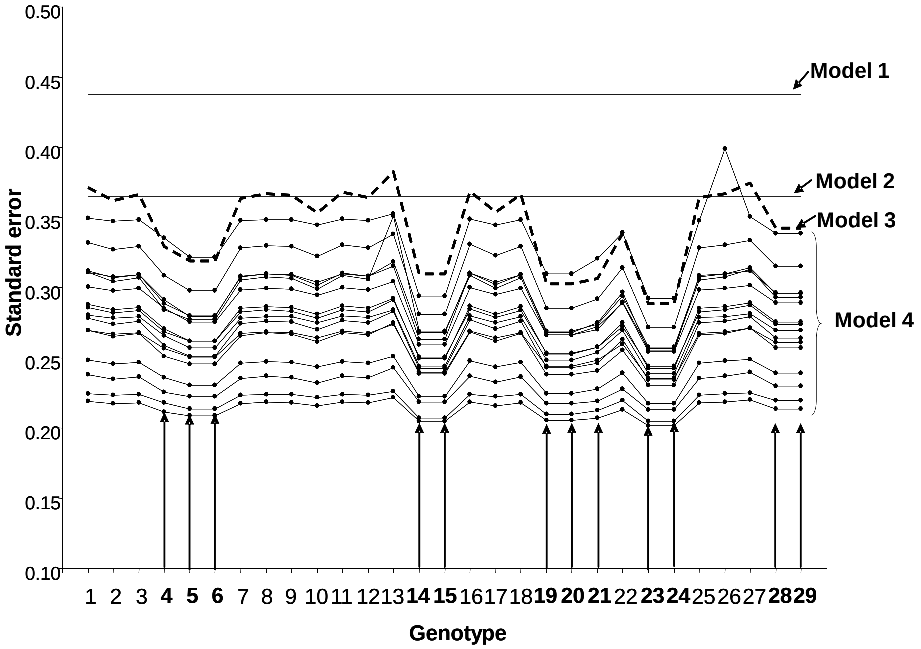

Fig. (1). Standard errors of BLUPs of 29 wheat genotypes (1-29) within 16 sites for grain yield for four linear-bilinear models (models 1, 2, 3, and 4). Sister lines are in bold (5,6), (4,19,20), (14,15), (21,23,24), (28,29). For model 15, each graph line represents one of the 16 sites. Model 1 is the standard fixed linear-bilinear model (SREG), model 2 is a simple mixed linear-bilinear model without **A** (not modeling GE), model 3 is a simple mixed linear-bilinear model using **A** (not modeling GE), and model 4 is a mixed linear-bilinear model that models the GE using the FA and includes information on relatives (**A**) (adapted from Crossa *et al.* [16]).

models 2 and 3; model 3 is similar to model 2, but includes information on relatives. It is clear that the SEs of sister lines were always smaller than the SEs of genotypes that had no relatives in the trial in all cases except for models 1 and 2, which do not incorporate this information. When modeling the main effects of genotype and GE in conjunction with the information on relatives, the improvement in the precision of the BLUP of the sisters, as well as of the other lines, is evident. It should be pointed out that no model is selected based on the SE of the means because SEs are model-dependent and one might be dealing with an inappropriate model despite its small SE. Fig. (**1**) only shows the natural impact of performing a more realistic statistical analysis instead of ignoring the relationship between genotypes and environments.

The standard biplots of a fixed effect linear-bilinear model (model 1 that excludes matrix **A**) and model 4 (FA including matrix **A**) are depicted in Figs. (**2** and **3**), respectively. The biplot of model 1 (Fig. **2**) is the usual biplot of a fixed SREG model that shows genotypes 19 and 20 as having a positive response in terms of genotype main effect and GE for most of the sites because they are in the same direction. Genotypes 16, 18, and 25, located on the opposite side of the biplot, have a negative response in all sites. Sites located farther away from the center, such as NZL, USA1, ZBW, ISRL, and SPN2, are the ones that discriminate the genotypes the most. Pairs of sister lines 19 and 20, and 14 and 15 are distinct from the others. Sister lines are scattered throughout the two dimensions of the biplot.

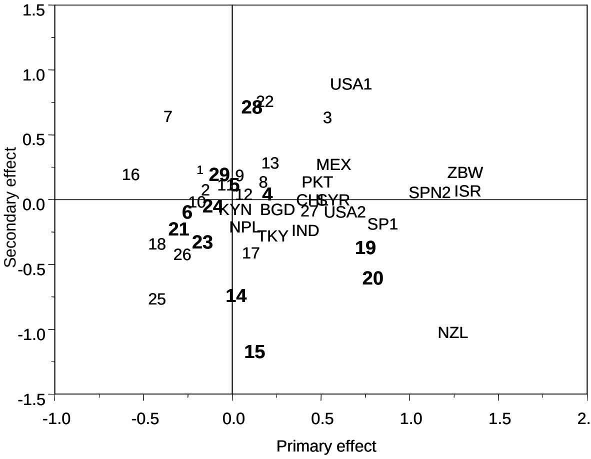

Fig. (2). Biplot of the standard fixed effect linear-bilinear model 1 for grain yield. Lines are 1-29. Sister lines are in bold (5,6), (4,19,20), (14,15), (21,23,24), (28,29). The 16 sites are MEX, USA1, USA2, TKY, ISR, BGD, IND, PKT, SYR, SPN1, SPN2, NPL, KNY, ZBW, NZL, and CHL (adapted from Crossa *et al.* [16]).

When the main effect of genotypes and GE are modeled together using the factor analytic variance-covariance model 4 (FA that includes information between relatives (Fig. **3**)), sister lines with a strong genetic association are held together, but others, such as genotype 4, tended to be farther away from their sister lines (19 and 20), because their COP was 0.792; a similar situation occurred with sister lines 28 and 29, with a COP of 0.774. Although the general patterns of response in terms of directions and projections of genotypes and sites in the biplot were not altered by using different models, the inclusion of information between relatives in conjunction with modeling the main effects of genotypes and GE by means of a factor analytic structure offers the best option for predicting breeding values of genotypes across different environments and studying the main effects of genotypes and GE. Models with FA for GE and pedigree information gave a more realistic view of the relationship between the genotypes themselves, between sites, and between their interactions.

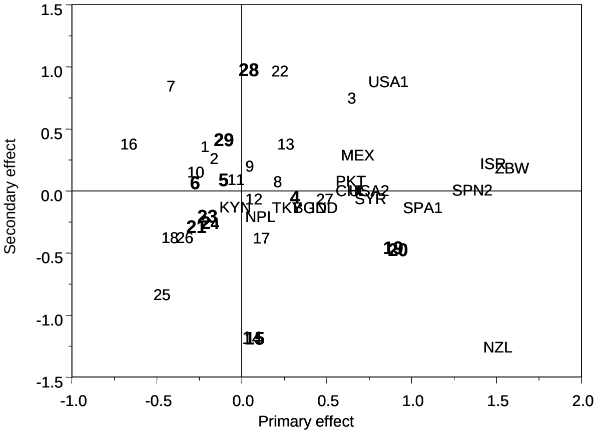

Fig. (3). Biplot of the mixed linear-bilinear model 4 with FA for GE and pedigree information (**A**) for grain yield. Lines are 1-29. Sister lines are in bold (5,6), (4,19,20), (14,15), (21,23,24), (28,29). The 16 sites are MEX, USA1, USA2, TKY, ISR, BGD, IND, PKT, SYR, SPN1, SPN2, NPL, KNY, ZBW, NZL, and CHL (adapted from Crossa *et al.* [16].

STATISTICAL MODELS FOR MAPPING QTLS AND STUDYING QTL × ENVIRONMENT INTERACTION (QEI)

This section is related to the issue of understanding the nature and causes of interaction. The factorial regression (FR) model (e.g., Vargas *et al.* [46] and van Eeuwijk *et al.* [47]) is useful for studying the effects of both genetic and environmental covariables and for developing functional relationships and predictability with explanatory covariables. In plant breeding, much research is directed at locating regions of the chromosomes that are involved in the physiological processes underlying phenotypical traits. These regions are called quantitative trait loci (QTLs). When the relative contributions of these regions to the explanation of the phenotype differ between genotypes, then QTL × environment interaction (QEI) occurs. The statistical problem can be interpreted as a multivariate multiple regression of phenotypic traits as observed over a set of environments on a set of genetic predictors. FR provides a suitable framework for QEI analysis. Crossa *et al.* [48] give examples of how FR can be used for assessing the chromosomal location of QTLs and QEI and the importance of their effects.

There are approaches in which the GE is modelled directly using regression on environmental (and/or genotypic) variables, rather than regression on the environmental mean, as originally proposed by Yates and Cochran [37]. A useful linear model for incorporating external environmental (or genotypic) variables is the FR model [49, 50]. The FR models are ordinary linear models that approximate the GE effects of Eq. 1 by the products of one or more of the following: (1) genotypic covariables (observed) × environmental potentialities (estimated); (2) genotypic sensitivities (estimated) × environmental covariables (observed); (3) scale factor (estimated) × genotypic covariables (observed) × environmental covariables (observed). The aim of FR is to replace, in the GE subspace, genotypic and environmental factors with a small number of genotypic and environmental covariables. Vargas *et al.* [51] further developed the statistical approaches described by Crossa *et al.* [48] and van Eeuwijk *et al.* [52] for modeling QTLs and QEI. The main objectives of their research were to demonstrate the use of: (1) FR for estimating effects and locations of QTL and QEI, and (2) FR for modeling and interpreting QEI in terms of products of genetic predictors and environmental variables.

In FR, genotypic covariables x_a (a =1…A), with values x_{ia} for the ith genotype can be introduced for the genotypic main effect, $G_i = x_{ia}\rho_a + \text{residual}$ where ρ_a is the regression coefficient for the regression of G_i on the genotypic covariables, x_{ia}. For more than one genotypic covariable, this becomes $G_i = \sum_{a=1}^{A} x_{ia}\rho_a + \text{residual}$. In the context of molecular markers and when attempting to map QTLs, the genotypic covariables, x_a, are replaced by the genetic predictors x_q, and the FR framework can also be used to do a genome scan for QTL effects. Analogous to the genotypic main effect, in FR, the environmental main effect, E_j, can also be regressed on environmental covariables z_b with values z_{jb} for the jth environment. The corresponding partitioning is $E_j = z_{jb}\beta_b + \text{residual}$ for one environmental covariable, or $E_j = \sum_{b=1}^{B} z_{jb}\beta_b + \text{residual}$ for multiple environmental covariables. The parameter β_b is the regression coefficient for the regression of the environmental main effect on the environmental covariables, z_{jb} for the jth environment.

Within a QTL analysis by FR, a multiple QEI model follows easily from models for GE: $(GE)_{ji} = \sum_{q=1}^{Q} x_{iq}\rho_{jq} + \text{residual}$ where ρ_{jq} represents a QEI effect, i.e., a differential QTL expression in relation to the main effect QTL expression, for the qth QTL in jth environment. QEI for a QTL q' can be further modeled by regressing it on an environmental covariable, z_b: $(GE)_{ij} = \nu_{q'b}x_{iq}\rho_{jb} + \text{residual}$. For multiple QTLs, this generalizes to $(GE)_{ij} = \sum_{q=1}^{Q}\sum_{b=1}^{B} \nu_{qb}x_{iq}\rho_{jb} + \text{residual}$. One or more QTL main effects can be tested by comparing the model $\bar{y}_{ij} = \mu + \sum_{q=1}^{Q} x_{iq}\rho_q + E_j + \text{residual}$ with the model $\bar{y}_{ij} = \mu + E_j$. When main effect QTL expression and QEI are considered together, this is equivalent to fitting different effects for the same QTL in different environments. A specific test for QE compares $\bar{y}_{ij} = \mu + \sum_{q=1}^{Q} x_{iq}\rho_q + E_j + \sum_{q=1}^{Q} x_{iq}\rho_{jq} + \text{residual}$ to $\bar{y}_{ij} = \mu + \sum_{q=1}^{Q} x_{iq}\rho_q + E_j + \text{residual}$. F-tests can be constructed from ratios of regression mean squares to independent error term.

An example including 211 F2-derived F3 maize families from a biparental cross evaluated across eight environments differing in the level of drought stress and soil nitrogen content is employed to illustrate the use of the fixed effect FR linear model for mapping QTLs and for studying the influence of external covariables on QEI [51]. Table **3** shows parts of the analysis of variance table at position 63 cM on chromosome 10 [51]. The first part shows the usual analysis of variance for a two-way table of grain yield measured in 211 genotypes with partitioning of the joint effect of G+GE into G and GE effects. The middle part of Table **3** shows the variability due to QTL+QE effects in parts of the genome other than chromosome 10 (i.e., due to QTLs on chromosomes 1-9), the variability due to G+GE after correction for QTLs on the other chromosomes, and the corresponding partitioning into G and GE components. The last part shows the partitioning of G+GE adjusted for the QTLs on chromosomes 1-9 into variation due to QTL+QE at position 63 cM on chromosome 10 and deviations from the QTL model. When the QEI effects were regressed on the set of environmental covariables, maximum temperature at flowering showed the closest relationship with the QEI effects and explained 23.8% of the QEI. The effect of this environmental covariable was highly significant by an F-test for the regression mean square over the deviations from the regression.

For grain yield, Fig. (**4**) depicts the profile of R^2_{QTL}, R^2_{QEI}, and $R^2_{QTL+QEI}$ and the corresponding critical values for α=0.01 based on 1000 randomizations. There is good reason to believe that there are environment-specific QTLs around 63 cM of chromosome 10 (QTL+QEI and QEI effects were both significant). In contrast, only main effect QTLs were observed in other chromosomes.

Table 3. Partitioning of Yield Variation at Position 63 cM on Chromosome 10. For Comparison, an Error Estimated from the Median Intra-Block Error was 0.75 (Adapted from Vargas *et al*. [51])

Source of variation	Degrees of freedom	Sum of squares	Mean squares
Environment (E)	7	12777.169	1825.310
G+GE	1680	3212.868	1.914
F2 family (G)	210	1382.102	6.581
GE	1470	1829.700	1.245
Total	1687	15988.970	
G+GE	1680	3212.868	1.914
QTL+QEI Chrom. 1-9	---*	1008.879	---
G+GE Chrom. 10 adj.	1680*	2203.988	---
F2 family (G) adj.	210	666.755	3.175
GE adj.	1470	1537.234	1.046
G+GE Chrom. 10 adj.	1680*	2203.988	---
QTL+QEI Chrom. 10 63 cM	8	93.868	11.733
QTL main effect	1	56.148	56.148
QEI	7	37.720	5.388
Max.Temp. Flow.	1	8.986	8.986
Residual QEI	6	28.72	4.787
Deviations	1672	2110.121	1.262

*For correction of the grain yield data due to genetic effects on chromosomes 1 through 9, degrees of freedom might be discounted.

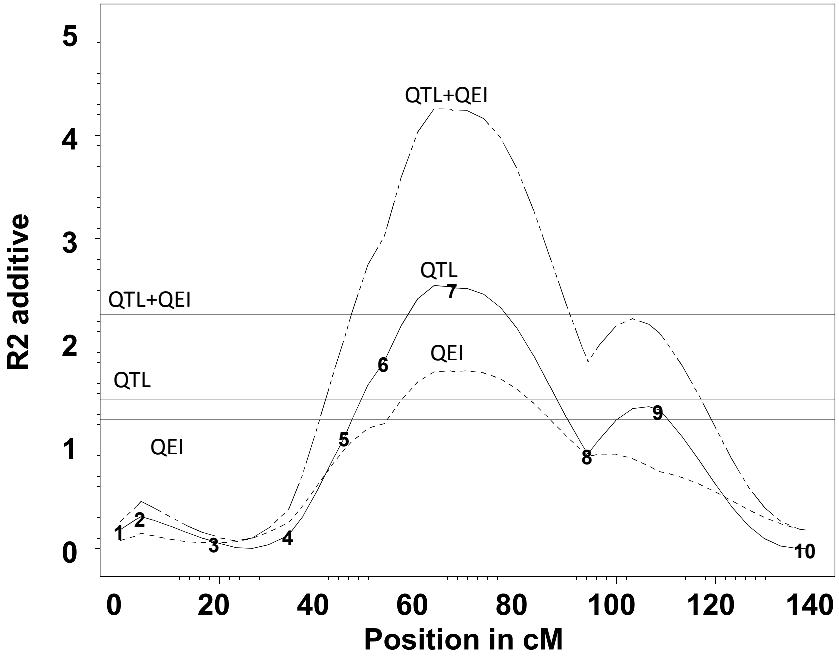

Fig. (4). Profile of R^2 for the additive effects of QTL (solid line), QEI (dotted line), and QTL+QEI (broken line) on grain yield for chromosome 10 (additive). The horizontal lines show the appropriate threshold for the effects QTL+QEI, QTL, and QEI (adapted from Vargas *et al*. [51]).

Linear Mixed Models for Multi-Trait Multi-Environment QTL Analysis

Plant breeders are interested in evaluating genotypes for multiple traits. A general formulation of a linear mixed model for the multi-trait multi-environment (MTME) model is presented by Malosetti *et al.* [53]. The initial model is $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$ where the response variable **y** is modelled by fixed and random effects factors **ß** and **u**, respectively; **X** and **Z** are design matrices assigning fixed and random effects to the observations, respectively. Random genetic effects are assumed to be normally distributed, **u**~N(**0**,**G**), with **G** the genetic (co)variance matrix. Finally, **e** is a vector of non-genetic residuals associated with each observation and normally distributed, **e**~N(**0**, **R**). The phenotypic (co)variance is given by V(**y**) = **ZGZ'** + **R**. From a breeder's point of view, **G** is of special interest, as it reflects the magnitude and pattern of relationships between genetic effects of traits. Exploiting the genetic correlations between traits is useful because it may be a primary trait with low heritability or a difficult-to-measure primary trait that is correlated with an easy-to-measure secondary trait with high heritability.

A QTL model arises by including the effect of a putative QTL as follows: $\mathbf{y} = \mathbf{X\beta} + \mathbf{X}^{\mathbf{QTL}}\boldsymbol{\alpha} + \mathbf{Zu}^* + \mathbf{e}$. The extra term in the model is composed of a design matrix $\mathbf{X}^{\mathrm{QTL}}$, which is derived from molecular marker information and a vector of fixed QTL effects (**α**). In an MTME model, vector **α** has dimensions JK×1 and contains the additive genetic QTL effects for all the traits in each of the environments. The random genetic effects, now collected in a vector **u***, result from the effects of QTLs outside the test region, that is, the genetic background. Genetic background effects are assumed to be normally distributed: **u** ~ N(0, **G***). Note that **G*** represents the part of the genetic (co)variance that is not explained by the QTL. The extension from a single-QTL model to a multi-QTL model is straightforward and given by

$$\mathbf{y} = \mathbf{X\beta} + \sum_{\mathbf{q}=\mathbf{1}}^{\mathbf{Q}} \mathbf{X}_{\mathbf{q}}^{\mathbf{QTL}}\boldsymbol{\alpha}^{\mathbf{q}} + \mathbf{Zu}^* + \mathbf{e}\,.$$

An example of the application of the MTME is presented in a wheat trial (S. Singh and M. Vargas, personal communication), where four diseases (Karnal bunt, tan spot, yellow rust, and leaf rust) were simultaneously measured in a biparental progeny. A multi-trait QTL scan was performed and the various significant QTLs detected are depicted in Figs. (**5** and **6**). Fig. **5** shows the joint profile and the two significant peaks for the same trait, Karnal bunt, one between 20 and 20 cM and the other one at about 95 cM chromosome 3A. On the other hand, Fig. **6** shows two significant peaks for two different traits, one for yellow rust at the beginning of chromosome 5B and another peak for Karnal bunt at about 20 cM. These findings suggest that some diseases are determined in similar regions of the chromosomes; they should be studied jointly rather than by single-trait QTL mapping.

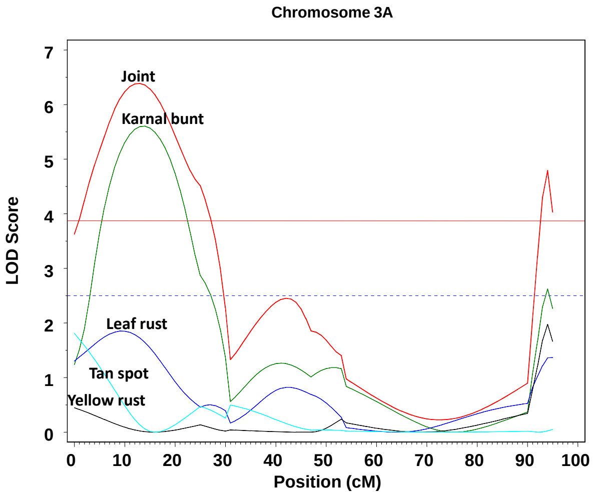

Fig. (5). Profile of LOD of chromosome 3A from the multi-trait analysis. The LOD profile of the joint analysis of all four traits (red) and the LOD profile of the marginal analyses for Karnal bunt (green), leaf rust (blue), tan spot (light blue), and yellow rust (black). The LOD threshold for the joint profile is the red horizontal line and the LOD threshold for the traits is the blue horizontal line.

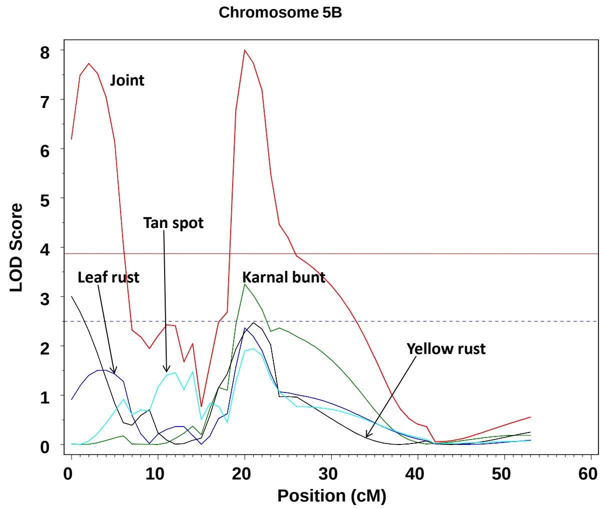

Fig. (6). Profile of LOD of chromosome 5B from the multi-trait analysis. The LOD profile of the joint analysis of all four traits (red) and the LOD of the marginal analyses for Karnal bunt (green), leaf rust (blue), tan spot (light blue) and yellow rust (black). The LOD threshold for the joint profile is the red horizontal line and the LOD threshold for the traits is the blue horizontal line.

STATISTICAL MODELS FOR GENOMIC SELECTION AND PREDICTION

Selection in plant breeding is based on estimates of breeding values obtained with pedigree-based mixed models [16-20]. These models have been used successfully for predicting breeding values in plants and animals. However, pedigree-based models cannot account for Mendelian segregation, a term that under an infinitesimal additive model [23] and in the absence of inbreeding, explains one half of the genetic variability. Molecular markers (MM) allow tracing Mendelian segregation at several positions of the genome; potentially, this may increase the accuracy of estimates of genetic values and of the genetic progress attainable when these predictions are used for selection purposes.

Genomic selection (GS) (or genome-wide selection) is an approach for improving quantitative traits [29, 35] that uses all available MM across the genome to estimate genetic values. Genomic selection has been validated in several species and populations in animal breeding [35, 36]. However, reports on the use of GS in plants are few and refer mainly to computer simulation studies such as the research of Bernardo and Yu [30], who concluded that GS was superior to marker assisted selection in maize. In recent articles, de los Campos *et al.* [31], Pérez *et al.* [34], and Crossa *et al.* [32, 33] validated GS in plant breeding using genomic regression and showed that models using MM were more accurate in predicting grain yield in wheat and maize than those based on pedigree only.

One method for incorporating markers into models for GS is to define genetic values (g_i , i=1,...,g) as a parametric regression on marker covariates x_{ij} (which can take values of 1, 0 or -1 for a biallelic marker of a segregating population, or values of 1 and 0 for inbred lines) of the form $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$. Phenotypes (y_i) can then be represented as $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$ (j=1,2,…,p), where β_j is the regression of y_i on the jth marker covariate, and ε_i is a model residual. This approach was first proposed by Meuwissen *et al.* [29] and is the most commonly used in GS. With high-density markers, the number of markers exceeds the number of individuals, and estimation of marker effects via ordinary least squares (OLS) is not feasible. Instead, penalized or Bayesian estimation methods are commonly used. Several penalized and Bayesian shrinkage estimation methods are available. Examples of the first group are Ridge Regression (RR) and the Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani [54] and Elastic Net (EN). The list of Bayesian models is extensive and includes Bayesian counterparts of RR (BRR) and LASSO [55] (BL) and the Elasctic Net (BEN).

We considered models for GS that differ in the information used (pedigree, molecular markers or both) and in the way molecular markers are incorporated into the model (parametric regression or semi-parametric regression). In all cases, the response variable is the average standardized performance of each line within each environment, that is $\bar{y}_i = \sum_{k=1}^{n_i} y_{ik} /(\mathrm{SD} \times n_i)$, where n_i is the number of replicates available for the ith line and SD is the (sample) standard deviation of the response variable $\{\bar{y}_i\}$. The conditional distribution of this phenotype given $\mu, \mathbf{g}, \sigma_\varepsilon^2$ is Gaussian,

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

where μ is an intercept, $\bar{\mathbf{y}} = \{\bar{y}_i\}$ and $\mathbf{g} = \{g_i\}$ are vectors of average phenotypes and genetic values, respectively, and σ_ε^2 is a residual variance.

A standard additive infinitesimal model (e.g., Fisher [25] and Henderson [57]) postulates that genetic values are multivariate normal, centered at zero, and with a co-variance matrix proportional to the numerator relationship matrix (**A**) computed from the pedigree, that is, $g_i = a_i$, where $\mathbf{a} = (a_1,..., a_n) \sim N\left(\mathbf{a}\middle|\mathbf{0}, \mathbf{A}\sigma_a^2\right)$, and σ_a^2 is the additive variance. The collection of unknowns in this model is $\left\{\mu, \mathbf{a}, \sigma_a^2, \sigma_\varepsilon^2\right\}$ and in a Bayesian setting, a prior density is assigned to these unknowns. Following standard assumptions, independent scaled inverse Chi-square distributions for the residual and the additive variance are chosen. The joint prior of this pedigree-based model becomes:

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

where $\chi^{-2}(.|.,.)$ is a scaled-inverse Chi-square density, and df and S are prior degree-of-freedom and scale parameters, respectively. The joint posterior distribution for this model is obtained combining the likelihood and prior, such that:

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

Inferences are based on samples obtained using a Gibbs sampler.

An alternative is to replace **A** with kinship matrix (**U**) estimated using marker genotypes. A kinship-based infinitesimal model (K) is obtained using $g_i = u_i$, where $\mathbf{u} = \left(u_1,..., u_n\right)' \sim N\left(\mathbf{u}\middle|\mathbf{0}, \mathbf{U}\sigma_u^2\right)$ and σ_u^2 is the associated variance parameter.

Penalized Parametric Regression on Marker Effects

In ridge regression, estimates are obtained by minimizing the residual sum of squares, $\min_{\boldsymbol{\beta}}\left\{\sum\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\}$, subject to the following constraint: $\sum_j \beta_j^2 \le t$ or, equivalently,

$$\hat{\boldsymbol\beta}_{RR} = \arg\min_{\boldsymbol\beta}\left\{\sum_{i=1}^{n}\left(y_i - \mathbf{x}'_{\mathbf{i}}\boldsymbol\beta\right)^2 + \lambda\left(t\right)\sum_j \beta_j^2\right\}.$$

The solution to this optimization problem can be shown to be $\hat{\boldsymbol\beta} = (\mathbf{X}'\mathbf{X} + \lambda(t)\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$, where $\lambda\left(t\right) > 0$ is a regularization parameter that induces shrinkage of estimates of effects towards zero and controls the trade-off between goodness of fit amd model complexity. Even though these estimates are biased, the sampling variance is reduced, yielding smaller mean-squared error and better predictive ability.

An alternative to ridge regression is to use LASSO [54]. Estimates in LASSO are obtained by minimizing the residual sum of squares, $\min_{\boldsymbol\beta}\left\{\sum\left(y_i - \mathbf{x}'_i\boldsymbol\beta\right)^2\right\}$, subject to the following constraint: $\sum_j\left|\beta_j\right| \le t$ or, equivalently,

$$\hat{\boldsymbol\beta}_{LASSO} = \arg\min_{\boldsymbol\beta}\left\{\sum_{i=1}^{n}\left(y_i - \mathbf{x}'_{\mathbf{i}}\boldsymbol\beta\right)^2 + \lambda\left(t\right)\sum_j\left|\beta_j\right|\right\}.$$

Unlike the quadratic penalty of ridge regression (L_2 norm, $\sum_j \beta_j^2$) the absolute-value penalty of LASSO (L_1 norm, $\sum_j\left|\beta_j\right|$) induces selection and shrinkage simultaneously as some predictor effects may take the value of zero. When predictors are highly co-linear empirical evidences show that RR has better prediction accuracy than LASSO; this situation is common in GS with dense molecular markers.

One variant of the traditional LASSO is the elastic net LASSO, which differs from LASSO in that it uses two penalties the L_2 and the L_1 norms such that

$$\hat{\boldsymbol\beta}_{EN} = \arg\min_{\boldsymbol\beta}\left\{\sum_{i=1}^{n}\left(y_i - \mathbf{x}'_{\mathbf{i}}\boldsymbol\beta\right)^2 + \lambda_1\left(t\right)\sum_j|\beta_j| + \lambda_2\left(t\right)\sum_j \beta_j^2\right\}$$

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

Bayesian Shrinkage Regression Methods

Estimates of regression coefficients derived from penalized optimization problems such RR, LASSO or EN are equivalent to posterior modes in certain class of Bayesian Models. In the Bayesian approach, inferences are based on the posterior distribution of the unknowns (hyper-parameters, H) given the data (**y**), p(H|**y**). Following Bayes' rule, this density is proportional to the product of the conditional distribution of the data given the unknowns, p(**y**|H), or the Bayesian likelihood times the prior density assigned to model unknowns p(H). In the models we are interest, the conditional distribution of the data given the parameters p(**y**|H) can be represented as the the product of independent normal densities centered at the regression function, $E(y_i \mid \mathbf{x}_i, \boldsymbol\beta) = \sum_{j=1}^{p} x_{ij}\beta_j$ and with common residual variance σ_ε^2, that is

$$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol\beta, \sigma_\varepsilon^2) = \prod_i N\left(y_i\left|\sum_j x_{ij}\beta_j, \sigma_\varepsilon^2\right.\right)$$

Marker effects are assigned identical and independent normal prior densities, $p(\boldsymbol\beta \mid \boldsymbol\omega) = \prod_j N\left(\beta_j\middle|\boldsymbol\omega\right)$ where $\boldsymbol\omega$ represents hyper-parameters indexing the prior density of marker effects. Following Bayes' rule, the posterior distribution is:

$$p\left(\boldsymbol\beta\middle|\mathbf{y}, \sigma_\varepsilon^2, \boldsymbol\omega\right) \propto p(\mathbf{y} \mid \mathbf{X}, \boldsymbol\beta, \sigma_\varepsilon^2)p(\boldsymbol\beta \mid \boldsymbol\omega)$$

$$= \prod_i N\left(y_i\left|\sum_j x_{ij}\beta_j, \sigma_\varepsilon^2\right.\right)\prod_j N\left(\beta_j\middle|\boldsymbol\omega\right)$$

The choice of prior density, $p(\boldsymbol\beta \mid \boldsymbol\omega)$, determines whether this posterior mode is equivalent to estimates obtained with the BRR, BL, BEN. In BRR the prior density of marker effects is Gaussian, centered at zero and with common variance, that is, $p(\beta_j|\sigma_{\beta_j}^2) = N\left(\beta_j\middle|0, \sigma_{\beta_j}^2\right)$ where $\sigma_{\beta_j}^2$ is a prior-variance of marker effects. In BL the prior assigned to marker effects is Double-Exponential (DE) centered at zero and with inverse-scale parameter, $\lambda/\sigma_\varepsilon^2$, that is, $p(\beta_j|\lambda, \sigma_\varepsilon^2) = DE\left(\beta_j\middle|0, \lambda/\sigma_\varepsilon^2\right)$.The prior density of the Bayesian LASSO represents a compromise between the normal and DE densities. Relative to the Gaussian density, the DE places higher mass at zero and have thicker tails, inducing a different type of shrinkage. In particular, relative to BRR, the prior used in the BL induces stronger shrinkage towards zero of estimates of effects of predictors having weak association with the response, and not as much shrinkage of estimates of effects of predictors having strong association with the response.

Penalized Semi-Parametric Regression on Marker Effects

An alternative to parametric regressions is to use semi-parametric methods such as reproducing kernel Hilbert spaces (RKHS) regression [56]. A Bayesian RKHS regression for molecular markers regards genetic values as random variables coming from a Gaussian process with a (co)variance structure that is proportional to a kernel matrix $\mathbf{K}$, that is, $\mathrm{Cov}\left(g_i, g_j\right) \propto K\left(\mathbf{x}_i, \mathbf{x}_j\right)$, where $\mathbf{x}_i$, $\mathbf{x}_j$ are vectors of marker genotypes for the ith and jth individuals, respectively, and $K(.,.)$ is a positive definite function evaluated in marker genotypes. One of the advantages of RKHS regression is that it can be used with almost any information set (e.g., covariates, strings, images, graphs). An advantage of RKHS is that the model is represented in terms of n unknowns, which gives RKHS a great computational advantage relative to parametric methods, when p>>n. A disadvantage of the RKHS is that no individual marker effect can be estimated.

The geostatistical models of Piepho [58] that uses the ridge regression BLUP has also the advantage of having only *n* instead of *p* unknowns when p >> n. The RKHS applied to marker data is related to the spatial models because the ridge regression BLUP corresponds to a spatial model with a quadratic covariance model, while RKHS just replaces this quadratic model with a Gaussian model.

Examples of the Identification of Marker Effects in Environments for Genomic Selection

Do Some (or all) Markers (Genes) have Different (or the same) Response Patterns in Environments?

Genomic selection is designed to improve complex traits; it focuses mainly on prediction and, as the number of markers increases, prediction of the genomic estimated breeding values (GEBV) is expected to become more accurate, whereas the marker effect is expected to decrease in absolute magnitude. However, parametric regression models also provide the opportunity to examine marker effects and study the possible differential response of markers in environments, that is, the gene × environment interaction effect. In general, the previously examined Bayesian shrinkage methods do not have an associated test for detecting chromosome regions; however, they can be routinely used for examining marker effects in certain chromosome regions.

Multivariate Analysis of Estimated Marker Effects in Environments

Parametric models such as those described in the previous section yield estimates of marker effects that are environment-specific. In the previous sections, we described how biplots [5, 43, 6] from singular-value decomposition can be used to assess GE. We used these techniques to study GE at the level of estimated marker effects.

Consider a matrix of estimated molecular marker effects, $\hat{\mathbf{B}}_{p\times q} = \left[\hat{\boldsymbol{\beta}}_1,...,\hat{\boldsymbol{\beta}}_q\right] = \left\{\hat{\beta}_{jk}\right\}$, whose columns, $\hat{\boldsymbol{\beta}}_k$, $k = 1,...,q$, are estimates of the effects of *p* markers in *q* different environments. The singular value decomposition of this matrix is $\hat{\mathbf{B}} = \mathbf{UDV}'$, where $\mathbf{U}_{p\times q} = \left[\boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_q\right] = \left\{\alpha_{jk}\right\}$ and $\mathbf{V}_{q\times q} = \left[\gamma_1,...,\gamma_q\right] = \left\{\gamma_{kl}\right\}$ are ortho-normal matrices that span the row (marker) and column (environment) spaces of $\hat{\mathbf{B}}$, respectively, and $\mathbf{D}_{q\times q}$ is a diagonal matrix whose non-null entries are the singular values of $\hat{\mathbf{B}}$, that is, $\mathbf{D} = \mathrm{Diag}\left\{\lambda_k\right\}$.

The biplot is constructed using the first and second components, that is, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, γ_1, and γ_2. Points in the biplot are the marker effects projected in the first two components, and are displayed using the coordinates provided by $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. The "environmental effects" are displayed as vectors whose coordinates are given by $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. The length of the vectors approximates the variance accounted for by the specific molecular marker and "environmental effect." Molecular markers represented in the same direction as the environments had positive effects on those environments, whereas molecular markers located in the opposite direction of the environmental vectors had negative effects on those environments. The cosine of the angle between two environments (or molecular marker effect) approximates the correlation of the two environments (or molecular marker), with an angle of zero indicating a correlation of +1, an angle of 90° (or -90°) a correlation of 0, and an angle of 180° a correlation of -1.

Biplots of marker effects for several maize datasets comparing maize lines genotyped with several hundred markers and phenotyped in different environments for male and female flowering and some diseases are shown in Crossa *et al.* [32, 33] using the Bayesian shrinkage M-BL model. The maize dataset is from the Drought Tolerant Maize for Africa project of CIMMYT's Global Maize Program and was obtained by genotyping 300 tropical inbred lines with 1148 SNPs. No pedigree was available for these data.Patterns of Co-Variability of Estimated Marker Effects Across Environments for Maize Flowering Genomic Data.

Traits included here were female flowering (FFL) (or days to silking), male flowering (MFL) (or days to anthesis), as well as the anthesis-silking interval (ASI) evaluated in 300 lines under severe drought stress (SS) and in well-watered (WW) environments [32]. The display of the first two component axes on estimated effects of the SNP markers in the six trait-environment combinations (MFL-SS, MFL-WW, FFL-SS, FFL-WW, ASI-SS, and ASI-WW) obtained from the M-BL model is depicted in Fig. (**7**). The correlation between trait-environment combinations using marker effects and phenotypic data, and the effects of the

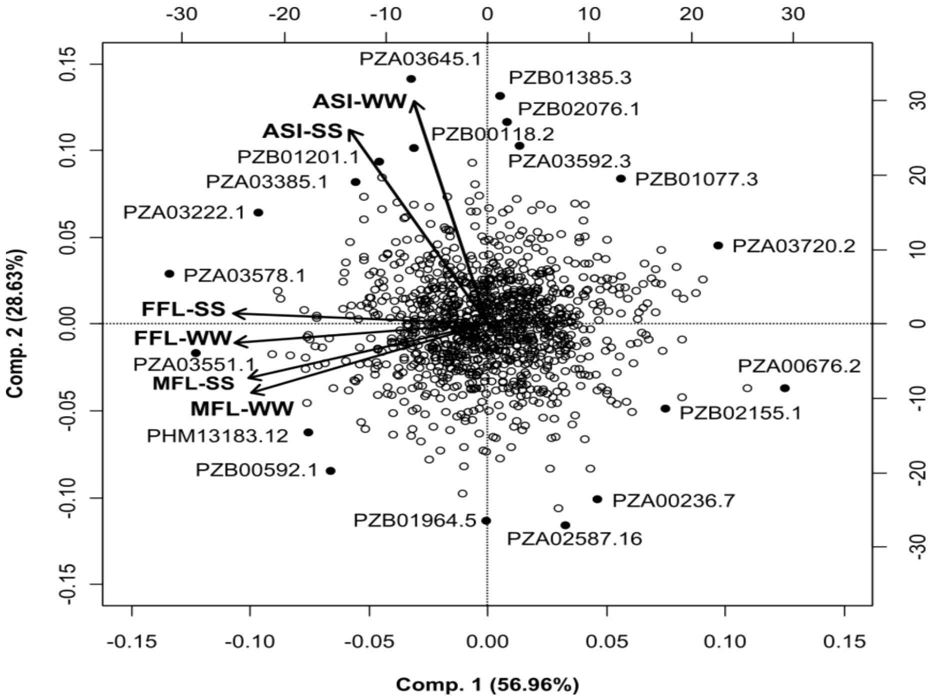

Fig. (7). Biplot of the first and second principal component axes (Comp. 1 and Comp. 2) of maize female flowering (FFL) and male flowering (MFL) effects of the 1148 SNPs estimated from the full data model M-BL of the maize dataset in each of two environments, severe water stress (SS) and well watered (WW). A total of six trait-environment combinations (FFL-SS, FFL-WW, MFL-SS, MFL-WW, SS-ASI, and WW-ASI) was formed. Only the effects of the 19 SNPs that are located far from the center of the biplot were identified with their corresponding SNP's name (filled-in circles) (from Crossa *et al*. [32]).

Table 4. Correlations Between Phenotypic Data (Upper Triangular) and Between Estimates of Marker Effects (Lower Triangular) from the Analysis of Six Trait-Environment Combinations (ASI-SS, ASI-WW, FFL-SS, FFL-WW, MFL-SS, and MFL-WW) of the Maize Flowering Time Trial Data (from Crossa *et al.* [32])

Trait-environment	ASI-SS	FFL-SS	MFL-SS	ASI-WW	FFL-WW	MFL-WW
ASI-SS	---	.446	.109	.728	.315	.109
FFL-SS	.472	---	.926	.221	.700	.633
MFL-SS	.095	.923	---	-.040	.678	.686
ASI-WW	.773	.266	-.037	---	.155	-.123
FFL-WW	.134	.497	.502	.066	---	.948
MFL-WW	-.051	.427	.505	-.173	.971	---

SNP markers most distant from the center of the biplot are in Tables **4** and **5**, respectively. Table **4** shows the correlations between phenotypic data (upper triangular) and between estimates of marker effects (lower triangular) from the analysis of six trait-environment combinations (ASI-SS, ASI-WW, FFL-SS, FFL-WW, MFL-SS, and MFL-WW) of the maize flowering trial data. Some trait-environment combinations are highly correlated (both phenotypically and genetically). The pattern of correlations between estimated SNP effects reflects the patterns of observed phenotypic correlations [32].

Clearly the two groups of trait-environment combinations are dominated more by the trait (ASI vs FFL and MFL) and less by the environment (SS and WW). Phenotypic outcomes and estimates of marker effects for ASI showed relatively small correlations with those of FFL and MFL; this is because ASI is defined as the difference between FFL and MFL, and these two traits are positively correlated.

Markers with relatively large (in absolute value) estimated effects are identified by name in Fig. (**7**), and their effects are shown in Table **5**. The marker effect on these traits should be interpreted differently than their effect on grain yield, since the favorable marker allele decreases both female and male flowering times, whereas for ASI, the optimal marker should give an ASI of 0. The alleles coded as 1 of SNPs whose estimated effects are located in the left and upper left corner of the biplot (i.e., PZA03551.1, PZA03578.1, PZA03222.1, PZA03385.1, PZB01201.1, and PZB00118.2) increase FFL, MFL, and ASI (they all have positive effects on all trait-environments combinations), whereas those SNPs located on the opposite side of the biplot (lower right corner) (i.e., PZA02587.16, PZA00236.7,

Table 5. Biplot of the Principal Component Analysis on the Marker Effects in Each of the Six Trait-Environment Combinations. Estimated Effect of the 19 SNP Molecular Markers Located Farthest from the Center of the Biplot in Fig. (7) from the Maize Flowering Time Trial Data (from Crossa *et al.* [32])

SNP	ASI-SS	FFL-SS	MFL-SS	ASI-WW	FFL-WW	MFL-WW
PZB02155.1	-.01567	-.02808	-.01713	-.01832	-.00957	-.00629
PZA03551.1	.00907	.03973	.02889	.00491	.01938	.01724
PZA03720.2	-.00062	-.02286	-.02511	.00301	-.01575	-.01822
PZA03578.1	.02019	.06300	.04031	.01414	.01310	.01141
PZA03592.3	.02969	-.00158	-.01165	.01705	-.00441	-.00700
PZA03645.1	.02668	.00951	-.00011	.04379	.00257	-.00135
PZA02587.16	-.01695	-.00925	-.00183	-.03975	-.00428	.00134
PZB01385.3	.03359	.00021	-.00960	.02742	-.00337	-.00757
PZA00236.7	-.03476	-.01352	-.00121	-.02162	-.00433	-.00327
PZA00676.2	-.02956	-.05885	-.03312	-.00786	-.01213	-.01068
PZB01077.3	.02016	-.01423	-.02687	.00962	-.00887	-.01099
PHM13183.12	-.00252	.03107	.03278	-.01239	.00734	.01058
PZA03385.1	.02674	.01727	.00896	.02199	.00725	.00306
PZB00592.1	-.01450	.01787	.02865	-.01194	.00966	.01302
PZB01201.1	.02927	.01670	.00614	.02306	.00431	.00128
PZB00118.2	.02900	.01084	.00216	.02443	.00285	-.00150
PZB01964.5	-.02849	.00531	.01561	-.02490	-.00003	.00222
PZA03222.1	.02976	.04733	.02447	.01574	.00902	.00545
PZB02076.1	.02229	-.00645	-.01467	.03059	-.00063	-.00484

PZB0255.1, and PZA00676.2) decrease the value of FFL, MFL, and ASI. Those SNPs whose presence is expected to increase or decrease traits across environments can be viewed as contributing to positive genetic correlations in FFL, MFL, and ASI between environments.

Despite the high heritability (between 0.74 and 0.87) found for flowering time and ASI in this maize trial, results show substantial interaction between molecular marker effects and environment. The biplot in Fig. (**7**) shows SNPs that had very contrasting effects across environments. For example, the minor alleles of SNPs whose estimated effects are located in the upper right corner of the biplot (PZA03592.3, PZB01077.3, and PZB02076.1) increase the anthesis-silking interval under drought and well-watered conditions (Table **5**), but decrease days to male and female flowering. In contrast, the minor alleles of SNPs whose estimated effects are located in the opposite quadrant of the biplot (lower left corner) (PZB00592.1, PHM13183.12, and PZB01964.5) showed a complete rank reversal with respect to the effects of SNPs PZA03592.3, PZB01077.3, and PZB01077.3 on those trait-environment combinations, i.e., a decrease in ASI under SS and WW, and an increase in male and female flowering times. These results are suggestive of important molecular marker effect × environment interaction, which in turn causes genotype × environment interaction.

Patterns of Co-Variability of Estimated Marker Effects Across Environments for Exserohilum Turcicum (NCLB) in Maize Genomic Data

This section shows marker effects from northern corn leaf blight (NCLB) disease caused by the fungus *Exserohilum turcicum* and evaluated in three international environments: El Batán (Mexico), Harare (Zimbabwe), and Mpongwe (Zambia) [31]. The patterns of marker effects in three environments for *Exserohilum turcicum* (NCLB) in the maize data are depicted in Fig. (**8**). The first two principal components explained 83.58% of the total variability in estimated SNP effects. The correlation among the three environments (based on phenotypic data or on estimates of marker effects) was 0.50 between Mpongwe and Harare and 0.26 between El Batán and Mpongwe and between El Batán and Harare. This pattern of correlations (depicted in Fig. **8**) may indicate the presence of different races of NCLB in these environments, i.e., a pathogen population in El Batán (Mexico) that may be different from populations in Harare and Mpongwe (Southern Africa).

The 16 SNPs with the largest effects (positive or negative) for NCLB are located farthest on the biplot in Fig. (**8**). The presence of the allele coded with 1 of the SNPs in group 1 should confer some degree of resistance to NCLB across environments. The opposite is true for SNPs in groups 2 and 3 (Fig. **8**); here the presence of the allele coded with 1

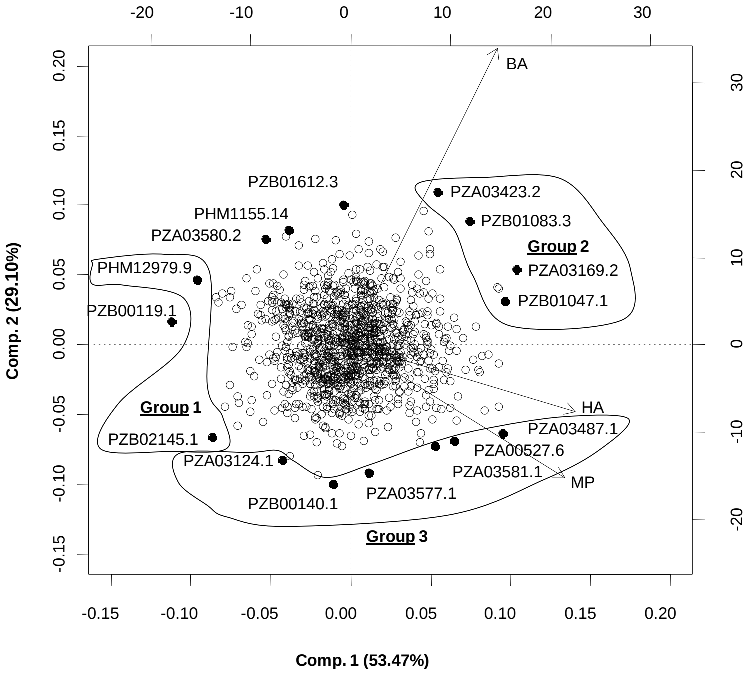

Fig. (8). Biplot of the first and second principal component axes (Comp. 1 and Comp. 2) of the *Exserohilum turcicum* (NCLB) disease effect of 1152 SNPs estimated from the full data M-BL model for the maize dataset in each of three environments: El Batán (México) (BA), Harare (Zimbabwe) (HA), and Mpongwe (Zambia) (MP). Only the effects of 24 SNPs located far from the center of the biplot were identified with the names of their corresponding SNPs. Three groups of environments and molecular markers are delineated as groups 1, 2, and 3 (from Crossa *et al*. [33]).

is expected to favor the disease in El Batán (group 2) and Harare and Mpongwe (group 3). Selection should therefore aim at decreasing this allele. It should be noted that caution must be exercised when interpreting the above results because they are based on one-year data. The prevalence of NCLB races can change from year to year, depending on changes in environmental conditions (e.g., temperature, rainfall, and relative humidity). The variance explained by these 16 SNPs with the largest positive or negative effects ranged between 0.3-1% of the total variance explained by all SNPs.

CONCLUDING REMARKS ON THE PLANT GENOMIC STUDIES

The results of the studies by de los Campos *et al.* [31] and Crossa *et al.* [32, 33] are encouraging; they indicate that models for GS can attain high predictive ability for genetic values of traits of economic interest under contrasting environmental conditions. The results indicate that genomic selection can be used effectively for selecting individuals whose phenotypes for various traits and in various environments have yet to be observed. As the number of available markers increases (as is expected in the near future), larger gains in predictive ability may be attained.

An advantage of models that include a parametric regression on marker covariates, such as M-BL, is that, in addition to estimating genetic values, they also provide information on (estimates of) “marker effects.” This information can be used to attain a better understanding of the genetic architecture of the traits under study and examine the patterns of response of marker effects across environments. In these studies, separate models were fitted to each trait-environment combination. An alternative to these single-environment models for genomic selection is to use multi-environment (or, equivalently, multi-trait) models where genetic values and marker effects on several traits/environments are jointly estimated. Multi-environment models allow borrowing information between correlated environments; thus it can be speculated that multi-environment genomic models can yield similar or even better predictions for individual environments. The literature on genomic selection has focused on single-trait models only; therefore the development of multi-environment and multi-trait models for genomic selection appears to be a natural next step.

Interestlingly, the results of these studies can also be used to generate a better design for field evaluations. For example, they show that prediction of unobserved lines in any of the correlated environments should be relatively accurate, and the scheme for testing these lines in any of those environments should be planned accordingly. It can be speculated that only one of the sets of correlated environments should be included in the trial, because any information lacking for the other environments can be borrowed from the one in use. However, unobserved lines in low correlated environments are expected to be poorly predicted. For example, concerning the analysis of NCLB in the maize data, results showed that estimated marker effects in the two African sites (Mpongwe and Harare) were positively correlated; however, they were negatively correlated with estimates of marker effects obtained for the same trait in Mexico (El Batán), which is suggestive of marker effect × environment interaction.

The analysis of genetic and genomic data is one of the most challenging and interesting mathematical and statistical problems researchers currently face. Therefore, different models from different areas of mathematical statistical research must be attempted in order to achieve a significant advance in understanding genetic effects. Models that consider complicated systems in biology and are able to study complex gene × gene interactions as well as gene × environment and gene × gene × environment are needed not only for improving prediction of the genetic values of individuals but also for understanding the chromosome regions related to important traits and the main climatic factors affecting these genomic regions. In a recent comprehensive study, Burgueño *et al.* [59] present multi-environment (multi-trait) models for GS and compare the predictive accuracy of these models with: (a) multi-environment analysis without pedigree and marker information, and (b) multi-environment pedigree or/and marker-based models. The authors described a statistical framework for incorporating pedigree and molecular marker information in models for multi-environment data and applied it to data that originated from wheat multi-environment trials. Two prediction problems relevant to plant breeders are considered: predicting the performance of untested genotypes ("newly" developed lines), and 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. Burgueño *et al.* [59] concluded that the evaluation of such trials can benefit greatly from using multi-environment GS models.

ACKNOWLEDGEMENTS

The author truly thank two anonymous reviewers that read, corrected, and suggested changes that have improved the quality of the chapter.

REFERENCES

[1] Cornelius, P.L.; Crossa, J.; Seyedsadr, M. Statistical tests and estimators for multiplicative models for cultivar trials. In: Kang, M.S.; Gauch, H.G.; Jr. Eds., Genotype-by-Environment Interaction. Boca Raton: CRC Press, **1996**, pp. 199-234.

[2] Crossa, J.; P.L. Cornelius. Site regression and shifted multiplicative model clustering of cultivar trial sites under heterogeneity of error variances. *Crop Sci.,* **1997**, *37*, 406-415.

[3] Gauch, H.G. Jr. Model selection and validation for yield trials with interaction. *Biometrics,* **1988**, *44*, 705-715.

[4] Gauch, H.G.; R.W. Zobel. Identifying mega-environments and targeting genotypes. *Crop Sci.,* **1997**, *37*, 311-326.

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

[6] Kempton, R.A. The use of biplots in interpreting variety by environment interactions. *J. Agric. Sci.,* **1984,** *103*, 123-135.

[7] Gauch, H.G. Statistical analysis of yield trials by AMMI and GGE. *Crop Sci.,* **2006**, *46*, 1488-1500.

[8] Yan, W.; Tinker, N.A. An integrated system of biplot analysis for displaying, interpreting, and exploring genotype-by-environment interactions. *Crop Sci.,* **2005**, *45*, 1004-1016.

[9] Yan, W.; Kang, M.S.; Ma, B.; Woods, S.; Cornelius, P.L. GGE biplot vs. AMMI analysis of genotype-by-environment data. *Crop Sci.,* **2007**, *47*, 643-655.

[10] Gauch, H.G.; Piepho, H.-P.; Annicchiarico, P. Statistical analysis of yield trials by AMMI and GGE: Further considerations. *Crop Sci.,* **2008**, *48*, 866-889.

[11] Yang, R.C.; Crossa, J.; Cornelius, P.L.; Burgueño, J. Biplot analysis of genotype × environment interaction: Proceed with caution. *Crop Sci.*, **2009**, *49*, 1564-1576.

[12] Piepho, H.P. Methods for comparing the yield stability of cropping systems- a review. *J. Agronomy Crop Sci*., **1998**, *180*, 193-213.

[13] Piepho, H.P. Ridge regression and extensions for genome-wide selection in maize. *Crop Sci.,* **2009**, *49*, 1-12.

[14] Smith, A.; Cullis, B.R.; Thompson, R. Exploring variety-environment data using random effects AMMI models with adjustment for spatial field trends: Part 1: Theory. In: Kang, M.S. (Ed.), Quantitative Genetics, Genomics and Plant Breeding. CABI Publishing, Oxford, **2002**.

[15] Smith, A.B.; Cullis, B.R.; Thompson, R. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *J. Agric. Sci.,* **2005**, *143*, 1-14.

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

[17] Piepho, H.P.; Mohring, J. Best linear unbiased prediction of cultivar effects for subdivided target regions. *Crop Sci.,* **2005**, *45*, 1151-1159.

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

[19] Burgueño, J.; Crossa, J.; Cornelius, P.L.; Yang, R.-C. Using factor analytic models for joining environments and genotypes without crossover genotype × environment interaction. *Crop Sci.,* **2008**, *48*, 1291-1305.

[20] Piepho, H.-P. Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics*, **1997**, *53*, 761-767.

[21] Piepho, H.-P. Empirical best linear unbiased prediction in cultivar trials using factor analytic variance-covariance structures. *Theor. Appl. Genet.,* **1998**, *97,* 195-201.

[22] Oakey, H.; Verbyla, A.; Pitchford, W.; Cullis, B.; Kuchel, H. Joint modeling of additive and non-additive genetic line effects in single field trials. *Theor. Appl. Genet.,* **2006**, *113*, 809-819.

[23] Cornelius, P.L.; Crossa, J. Prediction assessment of shrinkage estimators of multiplicative models for multi-environment cultivar trials. *Crop Sci.,* **1999**, *39*, 998-1009.

[24] Burgueño, J.; Crossa, J.; Miguel Cotes, J.; San Vicente, F.; Das, B. Prediction assessment of linear mixed models for multienvironment trials. *Crop Sci.,* **2011**, *51*, 944-954.

[25] Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburg.,* **1918**, *52*, 399-433.

[26] Malosetti, M.; Voltas, J.; Romagoza, I.; Ullrich, S.E.; van Eeuwjik, F.A. Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica,* **2004**, *137*, 139-145.

[27] Malosetti, M.; Boers, M.P.; Bink, M.C.A.M.; van Eeuwjik, F.A. 2006. Multi-trait QTL analysis based on mixed model with parsimonious covariance matrices. I: Proceedings of the 8th World Congress on genetics Applied to Livestock Production, August 13-18, Belo Horizonte, MG, Brazil. http://www.wcgalp8.org.br/wcgalp8. Article 25-04.

[28] Lande, R.; Thompson, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, **1990**, *124*, 743-756.

[29] Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic values using genome-wide dense marker maps. *Genetics,* **2001**, *157*, 1819-1829.

[30] Bernardo, R.; Yu, J. Prospects for genome-wide selection for quantitative traits in maize. *Crop Sci.,* **2007**, *47*, 1082-1090.

[31] de los Campos, G.; Naya, H.; Gianola, D.; Crossa, J.; Legarra, A.; Manfredi, E.; Weigel, K.; Cotes, J.M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, **2009**, *182*, 375-385.

[32] Crossa, J.; de los Campos, G.; Pérez, P.; Gianola, D.; Dreisigacker, S.; Burgueño, J.; Araus, J.L.; Makumbi, D.; Yan, J.; Singh, R.; Arief, V.; Banziger, M.; Braun, H.-J. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, **2010**, *186*, 1-12.

[33] Crossa, J.; Pérez, P.; de los Campos, G.; Mahuku, G.; Dreisigacker, S.; Magorokosho, C. Genomic selection and prediction in plant breeding. *J. Crop Improve.,* **2011**, *25*, 239-261.

[34] Pérez, P.; de los Campos, G.; Crossa, J.; Gianola, D. Genomic-enabled prediction based on molecular markers and pedigree using the BLR package in R. *Plant Genome,* **2010**, *3*(2), 106-116.

[35] Hayes, B.; Goddard, M. Genome-wide association and genomic selection in animal breeding. *Genome,* **2010**, *53*, 876-883.

[36] VanRaden, P.M. Genomic measures of relationship and inbreeding. *Interbull Bulletin,* **2007**, *37*, 33-36.

[37] Yates, F.; Cochran, W.G. The analysis of groups of experiments. *J. Agric. Sci.,* **1938**, *28*, 556-580.

[38] Finlay, K.W.; Wilkinson, G.N. The analysis of adaptation in a plant breeding programme. *Aust. J. Agric. Res.,* **1963**, *14*, 742-754.

[39] Eberhart, S.A.; Russell, W.A. Stability parameter for comparing varieties. *Crop Sci.,* **1966**, *6*, 36-40.

[40] Williams, E.J. The interpretation of interactions in factorial experiments. *Biometrika*, **1952**, *39*, 65-81.

[41] Gollob, H.F. A statistical model which combines features of factor analytic and analysis of variance. *Psychometrika,* **1968**, *33*, 73-115.

[42] Mandel, J. The partitioning of interaction in analysis of variance. *J. Res. Natl. Bur. Stand., Series B,* **1969**, *73*, 309-328.

[43] Gabriel, K.R. Least squares approximation of matrices by additive and multiplicative models. *J. R. Stat. Soc. Series B*, **1978**, *40*, 186-196.

[44] Cornelius, P.L.; Seyedsadr, M. Estimation of general linear-bilinear models for two-way tables. *J. Stat. Comput. Simul.,* **1997**, *58*, 287-322.

[45] Crossa, J.; Yang, R.-C.; Cornelius, P.L. Studying crossover genotype × environment interaction using linear-bilinear models and mixed models. *J. Agric. Biol. Environ. Stat.,* **2004**, *9*, 362-380.

[46] Vargas, M.; Crossa, J.; van Eeuwijk, F.A.; Ramirez, M.E.; Sayre, K. Using partial least squares, factorial regression and AMMI models for interpreting genotype×environment interaction. *Crop Sci.,* **1999**, *39*, 955-967.

[47] Van Eeuwijk, F.A.; Malosetti, M.; Yin, X.; Struik, P.C.; Stam, P. Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. *Aus. J. Agric. Res.,* **2005**, *56*, 883-894.

[48] Crossa, J.; Vargas, M.; van Eeuwijk, F.A.; Jiang, C.; Edmeades, G.O.; Hoisington, D. Interpreting genotype×environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theor. Appl. Genet.,* **1999**, *99*, 611-625.

[49] Denis, J.-B. Two-way analysis using covariates. *Statistics*, **1988**, *19*, 123-132.

[50] van Eeuwijk, F.A.; Denis, J.-B.; Kang, M.S. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In Kang, S.; Gauch, H.G. (Eds.). Genotype-by-environment interaction. CRC Press, Boca Raton, FL, **1996**.

[51] Vargas, M.; van Eeuwijk, F.A.; Crossa, J.; Ribaut, J.M. Mapping QTLs and QTL × environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. *Theor. Appl. Genet.,* **2006**, *112*, 1009-1023.

[52] van Eeuwijk, F.A.; Crossa, J.; Vargas, M.; Ribaut, J.M.. Variants of factorial regression for analysing QTL by environment interaction. Proceedings of the 11th Meeting of the EUCARPIA Section Biometrics in Plant Breeding. In Gallais, A.; Dillmann, C.; Goldringer, I. (eds.). Quantitative genetics and breeding methods: the way ahead. Paris, France, **2000**.

[53] Malosetti, M.; Ribaut, J.M.; Vargas, M.; Crossa, J.; van Eeuwijk, F.A. A multi-trait multi-environment QTL mixed model with application to drought and nitrogen stress trials in maize (*Zea mays* L.). *Euphytica*, **2008**, *161*, 241-257.

[54] Tibshirani, R. Regression shrinkage and selection *via* the LASSO. *J. R. Statist. Soc. B.,* **1996**, *58*, 267-288.

[55] Park, T.; Casella, G. The Bayesian LASSO. *J. Am. Stat. Assoc.,* **2008**, *103*, 681-686.

[56] Gianola, D.; van Kaam, J.B.C.H.M. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, **2008**, *178*, 2289-2303.

[57] Henderson, C.R. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **1975**, *31*, 423-447.

[58] Piepho, H.P. Ridge regression and extensions for genome-wide selection in maize. *Crop Sci.,* **2009**, *49*, 1-12.

[59] Burgueño, J.; de los Campos, G.; Weigel, K.; Crossa, J. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Sci.,* **2012**, *52,* 707-719.