

# BAYES EMPÍRICO MULTIVARIADO PARA PREDECIR EL MÉRITO GENÉTICO EN PLANTAS

## MULTIVARIATE EMPIRICAL BAYES TO PREDICT THE PLANT BREEDING VALUES

J. Jesus Ceron-Rojas\*, Jaime Sahagún-Castellanos

Instituto de Horticultura, Departamento de Fitotecnia, Universidad Autónoma Chapingo. 56230. Chapingo, México. (jesusceronrojas@live.com.mx).

### RESUMEN

El mérito genético de las plantas es heredable y determina características fenotípicas como altura de planta y rendimiento de grano, y puede predecirse por medio de modelos bayesianos univariados o multivariados con base en la información fenotípica o genómica de las plantas. Estos modelos controlan la incertidumbre asociada a la predicción pero son computacionalmente demandantes, por lo cual se requieren modelos alternativos menos demandantes. Bayes empírico es un método de predicción en el cual la esperanza de la distribución posterior es el estimador del mérito genético. Éste es una variante del estimador bayesiano estándar y es eficiente; es robusto ante las especificaciones erróneas de la distribución *a priori* de los parámetros y las covarianzas de éstos pueden estimarse por verosimilitud restringida. Para predecir el mérito genético en el contexto Bayes empírico se propuso un modelo lineal multivariado, el cual incorpora las correlaciones genéticas entre caracteres, la información del pedigrí, la información genómica, y contiene al modelo lineal genómico multivariado y al modelo lineal estándar multivariado como casos particulares. El modelo genómico usa solo información genómica mientras que el modelo estándar usa sólo información del pedigrí en la predicción. Para comparar numéricamente la eficiencia de cada uno de los tres modelos se usaron las correlaciones entre los valores predichos y observados obtenidas con los datos de dos poblaciones de maíz (*Zea mays*) F<sub>2</sub> y una población de trigo (*Triticum aestivum* L.) doble haploide, cada una de éstas con tres características y un conjunto particular de marcadores moleculares y genotipos. En las tres poblaciones los resultados numéricos indicaron que el modelo propuesto proporciona predicciones más precisas que los otros dos. Concluimos que los resultados se deben a que el modelo propuesto usa en la predicción, además de las correlaciones genéticas entre caracteres, la información fenotípica y genómica.

### ABSTRACT

The plant breeding value is inheritable and determines phenotypic characteristics such as plant height, and grain yield, and it can be predicted by means of univariate or multivariate Bayesian models based on the phenotypic or genomic plants information. These models control the uncertainty associated to prediction better, but this comes at a high computational cost, so less demanding alternative models are required. Empirical Bayes is a prediction method in which the expectation of the posterior distribution is the estimator of the breeding value. This is a variant of the standard Bayesian estimator and is efficient; it is robust to the erroneous specifications of the *a priori* distribution of parameters, and the parameter covariances can be estimated through restricted maximum likelihood. A multivariate linear model was proposed to predict the breeding value within the empirical Bayes context. This model incorporates the genetic correlations between traits, pedigree information, genomic information, and contains the multivariate genomic linear model and the multivariate standard linear model as particular cases. The genomic model uses only genomic information, whereas the standard model uses only information from the pedigree in the prediction. To compare numerically the efficiency of each of the three models, the correlations between the predicted and observed values obtained with the data from two maize (*Zea mays*) F<sub>2</sub> populations and one double haploid wheat (*Triticum aestivum* L.) population, each of them with three characteristics and a particular set of molecular markers and genotypes, were used. In the three populations, the numerical results indicated that the model proposed provides more precise predictions than the other two. We concluded that the results were due to the fact that the model proposed used the genetic correlations between traits and the phenotypic, as well as genomic information, in the prediction.

\*Autor responsable ❖ Author for correspondence.

Recibido: febrero, 2015. Aprobado: febrero, 2016.

Publicado como ARTÍCULO en *Agrociencia* 50: 633-648. 2016.

**Key words:** Joint posterior distribution, molecular markers, multivariate linear model, *Triticum aestivum*, restricted likelihood, *Zea mays*.

**Palabras clave:** Distribución posterior conjunta, marcadores moleculares, modelo lineal multivariado, *Triticum aestivum*, verosimilitud restringida, *Zea mays*.

## INTRODUCCIÓN

La predicción del mérito genético en plantas y animales por lo general se hace con el modelo lineal mixto (Robinson, 1991) o con alguna aproximación bayesiana (Blasco, 2001; Sorensen y Gianola, 2002) basados en los registros fenotípicos y de pedigrí de los candidatos a selección. Sin embargo, Meuwissen *et al.* (2001) mostraron que la selección genómica (SG) incrementa la precisión (correlación entre los valores observados y predichos) de la predicción del mérito genético de los candidatos a selección y reduce los intervalos entre ciclos de selección hasta en dos tercios cuando el número de genotipos y marcadores moleculares (MM) usado en la predicción es suficientemente grande. En la SG los valores predichos del mérito genético, llamados méritos genómicos estimados (genomic estimated breeding values, GEBV por sus siglas en inglés), se obtienen multiplicando los efectos estimados de los MM en la población base por los valores codificados de los MM obtenidos después del primer ciclo de selección. Los GEBV son el instrumento de la SG y permiten seleccionar caracteres cuantitativos en ausencia de información fenotípica (Gianola, 2013; Beyene *et al.*, 2015).

Uno de los problemas más importantes en la SG es obtener GEBV suficientemente precisos para que la SG sea eficiente. Tal problema condujo a varias metodologías de predicción derivadas de los siguientes supuestos: 1) los efectos de los MM tienen distribución normal multivariada con media igual a cero y varianza constante, y 2) los efectos de los MM tienen distribución normal multivariada con media igual a cero pero la varianza de los MM tiene distribución *a priori* que puede ser uniforme, gamma, etc.. El punto 1) condujo al predictor lineal mejor insesgado genómico (genomic best linear unbiased predictor o GBLUP) (VanRaden, 2008) y al mejor predictor lineal insesgado bayesiano (Bayes-BLUP) (Verbyla *et al.*, 2009, 2010). El punto 2) condujo a metodologías bayesianas como Bayes A, B, C, D, etc. (de los Campos *et al.*, 2013; Gianola, 2013), que difieren sólo en el supuesto específico

## INTRODUCTION

The prediction of the plant and animal breeding values is generally done through the mixed linear model (Robinson, 1991) or with some Bayesian approximation (Blasco, 2001; Sorensen and Gianola, 2002) based on the phenotypic and pedigree records of the candidates for selection. However, Meuwissen *et al.* (2001) showed that genomic selection (GS) increases the accuracy (correlation between values observed and predicted) of the prediction of breeding values of the candidate for selection, and reduces the intervals between selection cycles in up to two thirds when the number of genotypes and molecular markers (MM) used in the prediction is sufficiently large. In GS, the predicted breeding values, or genomic estimated breeding values (GEBV), are obtained by multiplying the MM effects estimated in the training population by the coded MM values obtained after the first selection cycle. The GEBVs are the tool of GS and allow selecting quantitative traits in absence of phenotypic information (Gianola, 2013; Beyene *et al.*, 2015).

One of the most important problems in GS is to obtain sufficiently precise GEBV so that the GS efficiency can be the highest possible. This problem has led to several prediction methodologies derived from the following assumptions: 1) the MM effects have a multivariate normal distribution with mean equal to zero and constant variance, and 2) the MM effects have an *a priori* distribution that can be uniform, gamma, etc. Point 1) led to the genomic best linear unbiased predictor (GBLUP) (VanRaden, 2008) and to the best Bayes linear unbiased predictor (Bayes-BLUP) (Verbyla *et al.*, 2009, 2010). Point 2) led to Bayesian methodologies such as Bayes A, B, C, D, etc. (de los Campos *et al.*, 2013; Gianola, 2013), which differ only in the specific assumption that they make with regard to the *a priori* variance distribution of the marker effects.

In GS, the Bayesian methods were developed within the context of a phenotypic variable with the objective of improving the GBLUP accuracy; however, it has not been irrefutably demonstrated that the GBLUP accuracy is significantly lower than the Bayesian methods (Massman *et al.*, 2013). The Bayesian methods provide better control of the uncertainty associated to the prediction of the

que hacen respecto a la distribución *a priori* de la varianza de los marcadores.

En la SG los métodos bayesianos se desarrollaron en el contexto de una variable fenotípica con el objetivo de mejorar la precisión de GBLUP; sin embargo, no se ha mostrado de manera irrefutable que la precisión de GBLUP sea significativamente menor que la de los métodos bayesianos (Massman *et al.*, 2013). Los métodos bayesianos permiten un control mejor de la incertidumbre asociada a la predicción del mérito genético (Blasco, 2001) pero requieren métodos numéricos, como el muestreador de Gibbs (Casella y George, 1992), para estimar los efectos de los MM y cualquier otro parámetro asociado al mérito genético. Verbyla *et al.* (2009) señalan que Bayes B requiere hasta 2440 h de cómputo para que el muestreador de Gibbs converja, mientras que Bayes A y Bayes-BLUP requieren al menos 6 h de cómputo para la convergencia de tal algoritmo. Según Verbyla *et al.* (2009), a pesar de grandes diferencias en el tiempo de cómputo en los métodos indicados, cuando el número de genotipos y MM es grande, la precisión alcanzada por todos ellos es prácticamente igual (promedio, 0.6).

Los procedimientos de predicción con modelos univariados no toman en cuenta las correlaciones genéticas entre caracteres, a pesar que en la práctica la evaluación de las variedades requiere varios caracteres de manera simultánea. Por ejemplo, los mejoradores de rendimiento y calidad de grano registran datos fenotípicos que incluyen componentes de rendimiento (*e.g.*, peso de grano o biomasa), calidad del grano (*e.g.*, sabor, forma, color, contenido de nutrientes), y resistencia a estrés biótico y abiótico (Jia y Jannink, 2012). La predicción del mérito genético multivariado tiene la ventaja de incorporar las correlaciones genéticas entre caracteres. Esta información debe aumentar la eficiencia de la predicción del mérito genético; por ello, Calus y Veerkamp (2011) propusieron un procedimiento similar a Bayes A, y Hayashi e Iwata (2013) adaptaron Bayes D al caso multivariado. Sin embargo, se requieren alternativas computacionalmente menos demandantes sin afectar la precisión de la predicción: Bayes empírico es un método alternativo de predicción con propiedades estadísticas deseables. En éste, bajo el supuesto de que las varianzas de los parámetros son conocidas, la esperanza de la distribución posterior del mérito

breeding value (Blasco, 2001) but require numerical methods, such as Gibbs sampling (Casella and George, 1992), to estimate the MM effects and any other parameter associated to the breeding value. Verbyla *et al.* (2009) point out that Bayes B requires up to 2440 h of computing for the Gibbs sampling to converge, while Bayes A and Bayes-BLUP require at least 6 h of computing for the convergence of such algorithm. According to Verbyla *et al.* (2009), despite the great differences in the computing time with the methods indicated, when the number of genotypes and MM is large, the accuracy reached by all of them is virtually equal (0.6, average).

The prediction procedures with univariate models do not take into account the genetic correlations between traits, although in practice evaluating the cultivars requires several traits in a simultaneous manner. For example, breeders of yield and grain quality register phenotypic data that include yield components (*e.g.*, grain weight or biomass), grain quality (*e.g.*, flavor, shape, color, nutrient content), and resistance to biotic and abiotic stress (Jia and Jannink, 2012). The prediction of the multivariate breeding value has the advantage of incorporating the genetic correlations between traits. This information must increase the efficiency in the breeding value prediction; for this reason Calus and Veerkamp (2011) suggested a procedure similar to Bayes A, and Hayashi and Iwata (2013) adapted Bayes D to the multivariate case. However, alternatives computationally less demanding without affecting the precision of the prediction are required: empirical Bayes is an alternative prediction method with desirable statistical properties. In it, under the assumption that the variances of the parameters are known, the expectation of the posterior distribution of the breeding value is an empirical Bayes estimator of such value (Tempelman and Rosa, 2004). This is a variant of the standard Bayesian estimator and is quite efficient; in addition, it is robust to the erroneous specifications of the *a priori* distribution of the parameters (Lehmann and Casella, 1998).

In the GS programs the first selection cycle includes only phenotypic information, although the training population (where the first group of parents is selected) has phenotypic and MM data. When making selection only with phenotypic information, the information from the MM is not used. If the

genético se considera un estimador bayesiano empírico de tal mérito genético (Tempelman y Rosa, 2004). Éste es una variante del estimador bayesiano estándar y es muy eficiente; además, es robusto ante las especificaciones erróneas de la distribución a priori de los parámetros (Lehmann y Casella, 1998).

En los programas de SG el primer ciclo de selección incluye sólo información fenotípica, aunque la población base (donde se selecciona al primer grupo de progenitores) tiene datos fenotípicos y MM. Al hacer selección sólo con información fenotípica no se usa la información de los MM. Si se combina la información fenotípica y la información de los MM en la predicción, aumenta la precisión aun en el primer ciclo de selección debido al aumento de información en el modelo. Un problema similar ocurre cuando sólo algunos individuos de los candidatos a selección cuentan con MM y el resto no, como en el mejoramiento de plantas híbridas (Massman *et al.*, 2013) o en la selección de animales (Legarra *et al.*, 2009).

El objetivo del presente estudio fue proponer y evaluar, en el contexto Bayes empírico, un modelo lineal multivariado que usa información de pedigrí y genómica de manera conjunta para predecir el mérito genético de los candidatos a selección. En el modelo, la esperanza de la distribución posterior conjunta del mérito genético es el estimador bayesiano empírico. Los supuestos de este modelo son: 1) las varianzas y covarianzas genéticas son conocidas; 2) el efecto genómico y el efecto genético aditivo no explicado por los MM tienen distribución normal multivariada conjunta con media igual a cero y varianza común; 3) el mérito genético de los candidatos a selección es la suma del efecto genómico y del efecto genético aditivo no explicado por los MM. Además, se muestra que el modelo lineal multivariado genómico (que usa sólo información genómica en la predicción) y el modelo lineal multivariado estándar (que usa sólo información fenotípica y de pedigrí en la predicción) son casos particulares del modelo propuesto.

## MATERIALES Y MÉTODOS

### Poblaciones de maíz 1 y 2

En cada una de las dos Poblaciones de maíz (*Zea mays*) F<sub>2</sub>, se registraron tres variables: rendimiento de grano (RG,

phenotypic and the MM information are combined in the prediction, the precision will increase even during the first selection cycle due to the increase in the model information. A similar problem occurs when only some individuals of the candidates for selection have MM and the rest do not, as in the hybrid plants breeding (Massman *et al.*, 2013) or in animal selection (Legarra *et al.*, 2009).

The objective of this study was to propose and evaluate, within the empirical Bayes context, a multivariate linear model that uses pedigree and genomic information in a joint manner to predict the breeding value of the candidates for selection. In such a model the expectation of the joint posterior distribution of the breeding values is the empirical Bayes estimator. The basic assumptions of this model are: 1) the genetic variances and covariances are known; 2) the genomic effect and the additive genetic effect not explained by the MM have a joint multivariate normal distribution with mean equal to zero and common variance; 3) the breeding value of the candidates for selection is the sum of the genomic effect and the additive genetic effect not explained by the MM. Besides, it is shown that the genomic multivariate linear model (which uses only genomic information in the prediction) and the standard multivariate linear model (which uses only phenotypic and pedigree information in the prediction) are particular cases of the proposed model.

## MATERIALS AND METHODS

### Maize populations 1 and 2

In each one of the two F<sub>2</sub> maize (*Zea mays*) Populations, three variables were recorded: grain yield (RG, Mg ha<sup>-1</sup>), ear height (AM, cm), and plant height (AP, cm). The maize Population 1 had 199 MM and 247 genotypes, whereas in the maize Population 2 the number of MM was 259 and of genotypes 248. The estimated genetic correlations between RG and AM, RG and AP, and AM and AP in the maize Population 1 were, respectively, 0.53, 0.52 and 0.98, whereas in the maize Population 2, those correlations were: 0.58, 0.76 and 0.71.

### Population 3 (wheat population)

The wheat (*Triticum aestivum* L.) double haploid Population included 1279 MM and 599 genotypes. In it, the grain yield

Mg ha<sup>-1</sup>), altura de la mazorca en la planta (AM, cm), y altura de planta (AP, cm). La Población de maíz 1 tuvo 199 MM y 247 genotipos, mientras que en la Población de maíz 2 el número de MM fue 259 y el de genotipos 248. Las correlaciones genéticas estimadas entre RG y AM, RG y AP, y AM y AP en la Población de maíz 1 fueron, respectivamente, 0.53, 0.52 y 0.98, mientras que en la Población de maíz 2 las correlaciones fueron 0.58, 0.76 y 0.71.

### Población 3 (población de trigo)

La Población de trigo (*Triticum aestivum* L.) doble haploide incluyó 1279 MM y 599 genotipos. En ella se registró el rendimiento de grano (RG, Mg ha<sup>-1</sup>) en tres ambientes (RG1, RG2 y RG3). Para predecir el mérito genético de los candidatos a selección a RG1, RG2 y RG3 se le consideró una característica particular debido a que los genotipos se evaluaron en ambientes diferentes. Las correlaciones genéticas estimadas entre RG1 y RG2, RG1 y RG3, y RG2 y RG3 en la Población 3 fueron, respectivamente, -0.03, -0.21 y 0.73.

### El modelo lineal propuesto con una variable

Sea  $\gamma_q = \mathbf{X}\mathbf{u}_q$  un vector  $g \times 1$  ( $g$ =número de genotipos en la población) de méritos genómicos aditivos asociados a la característica  $q$  ( $q=1,2,\dots,t$ ;  $t$ =número de variables) de los candidatos a selección. Suponga que  $\gamma_q$  tiene distribución normal multivariada (NMV) con media  $\mathbf{0}$  y varianza  $\mathbf{G}\sigma_{\gamma_q}^2$ , i.e.,  $\gamma_q \sim \text{NMV}(\mathbf{0}, \mathbf{G}\sigma_{\gamma_q}^2)$ , donde  $\sigma_{\gamma_q}^2$  es la varianza genómica aditiva de  $\gamma_q$  y  $\mathbf{G} = \mathbf{X}\mathbf{X}'/k$  es la matriz  $g \times g$  de relaciones genómicas aditivas entre genotipos;  $\mathbf{X}$  es una matriz  $g \times m$  ( $m$ =número de MM en la población) de valores codificados de los MM ( $2-2p$  para el genotipo  $AA$ ,  $1-2p$  para el genotipo  $Aa$ , y  $-2p$  para el genotipo  $aa$ ) asociados a los efectos aditivos de los *loci* de los caracteres cuantitativos (quantitative trait loci, o QTL por sus siglas en inglés);  $p$  es la frecuencia del alelo  $A$  y  $1-p$  es la frecuencia del alelo  $a$  en el MM  $j$  ( $j=1,2,\dots,m$ );  $\mathbf{u}_q$  es un vector  $m \times 1$  de efectos aditivos de los QTL asociados a los  $m$  MM que afectan a la variable  $q$ ;  $k = \sum_{j=1}^m 2p_j(1-p_j)$  (Habier *et al.*, 2007) en una población  $F_2$  y  $k = \sum_{j=1}^m 4p_j(1-p_j)$  en una población doble haploide. Además, sea  $\mathbf{a}_q \sim \text{NMV}(\mathbf{0}, \mathbf{A}\sigma_{\mathbf{a}_q}^2)$  un vector  $g \times 1$  de méritos genéticos aditivos no explicados por los MM asociados a la variable  $q$ , donde  $\mathbf{A}$  es la matriz de relaciones numéricas y  $\sigma_{\mathbf{a}_q}^2$  es la varianza genética aditiva de  $\mathbf{a}_q$ . El modelo lineal combinado para la variable  $q$  ( $\mathbf{y}_q^*$ ) puede denotarse como  $\mathbf{y}_q^* = \mathbf{1}\mu_q + \mathbf{Z}\mathbf{a}_q + \mathbf{Z}\gamma_q + \mathbf{e}_q$ , o de manera equivalente, como:

(RG, Mg ha<sup>-1</sup>) was recorded in three environments (RG1, RG2, RG3). To predict the breeding value of the candidates for selection, RG1, RG2 and RG3, each of them was considered a particular characteristic because the genotypes were evaluated in different environments. The estimated genetic correlations between RG1 and RG2, RG1 and RG3, and RG2 and RG3, were -0.03, -0.21 and 0.73, respectively.

### The proposed univariate linear model

Let  $\gamma_q = \mathbf{X}\mathbf{u}_q$  be a  $g \times 1$  vector ( $g$ =number of genotypes in the population) of genomic breeding values associated to the characteristic  $q$  ( $q=1,2,\dots,t$ ;  $t$ =number of variables) of the candidates for selection. Assume that  $\gamma_q$  has multivariate normal distribution (NMV) with mean  $\mathbf{0}$  and variance  $\mathbf{G}\sigma_{\gamma_q}^2$ , i.e.,  $\gamma_q \sim \text{NMV}(\mathbf{0}, \mathbf{G}\sigma_{\gamma_q}^2)$ , where  $\sigma_{\gamma_q}^2$  is the additive genomic variance of  $\gamma_q$  and  $\mathbf{G} = \mathbf{X}\mathbf{X}'/k$  is the  $g \times g$  additive genomic relationships matrix between genotypes;  $\mathbf{X}$  is a  $g \times m$  matrix ( $m$ =number of MM in the population) of coded MM values ( $2-2p$ ,  $1-2p$  and  $-2p$  for the genotypes  $AA$ ,  $Aa$ , and  $aa$ , respectively) associated to the additive effects of the quantitative traits *loci* (QTL);  $p$  is the frequency of the allele  $A$  and  $1-p$  is the frequency of allele  $a$  in the MM  $j$  ( $j=1,2,\dots,m$ );  $\mathbf{u}_q$  is a  $m \times 1$  vector of additive effects of the QTL associated to the  $m$  MM that affect trait  $q$ ;  $k = \sum_{j=1}^m 2p_j(1-p_j)$  (Habier *et al.*, 2007) in a population  $F_2$  and  $k = \sum_{j=1}^m 4p_j(1-p_j)$  in a double haploid population. In addition, let  $\mathbf{a}_q \sim \text{NMV}(\mathbf{0}, \mathbf{A}\sigma_{\mathbf{a}_q}^2)$  be a  $g \times 1$  vector of additive genetic merits unexplained by the MM associated to the trait  $q$ , where  $\mathbf{A}$  is the numerical relationships matrix and  $\sigma_{\mathbf{a}_q}^2$  is the additive genetic variance of  $\mathbf{a}_q$ . The combined linear model for the trait  $q$  ( $\mathbf{y}_q^*$ ) can be denoted as  $\mathbf{y}_q^* = \mathbf{1}\mu_q + \mathbf{Z}\mathbf{a}_q + \mathbf{Z}\gamma_q + \mathbf{e}_q$ , or equivalently as:

$$\mathbf{y}_q = \mathbf{Z}\mathbf{a}_q + \mathbf{Z}\gamma_q + \mathbf{e}_q \quad (1)$$

where  $\mathbf{y}_q = \mathbf{y}_q^* - \mathbf{1}\mu_q \sim \text{NMV}(\mathbf{0}, \mathbf{V}_q)$  is a vector  $g \times 1$  of the  $q$  trait observations centered with regard to the  $q$  trait mean,  $\mu_q$ ;  $\mathbf{1}$  is a  $g \times 1$  vector of ones;  $\mathbf{V}_q = \mathbf{A}\sigma_{\mathbf{a}_q}^2 + \mathbf{G}\sigma_{\gamma_q}^2 + 2\text{Cov}(\mathbf{a}_q, \gamma_q) + \mathbf{I}_g\sigma_{\mathbf{e}_q}^2$ , and  $\text{Cov}(\mathbf{a}_q, \gamma_q) = \mathbf{G}\sigma_{\gamma_q}^2$  (i.e., the covariance between  $\mathbf{a}_q$  and  $\gamma_q$  is equal to the variance of  $\gamma_q$ );  $\gamma_q$ ,  $\mathbf{G}$  y  $\sigma_{\gamma_q}^2$  were defined before;  $\mathbf{Z}$  is a matrix of incidence (generally an identity matrix  $g \times g$ ) y  $\mathbf{e}_q \sim \text{NMV}(\mathbf{0}, \mathbf{I}_g\sigma_{\mathbf{e}_q}^2)$  is a  $g \times 1$  vector of residuals;  $\mathbf{I}_g$  is an identity matrix  $g \times g$  and  $\sigma_{\mathbf{e}_q}^2$  is the variance of residuals. The



$$\mathbf{y}_q = \mathbf{Z}\mathbf{a}_q + \mathbf{Z}\gamma_q + \mathbf{e}_q \quad (1)$$

donde  $\mathbf{y}_q = \mathbf{y}_q^* - \mathbf{1}\mu_q \sim \text{NMV}(\mathbf{0}, \mathbf{V}_q)$  es un vector de observaciones  $g \times 1$  de la variable  $q$  centradas respecto a la media  $\mu_q$ ;  $\mathbf{1}$  es un vector  $g \times 1$  de unos;  $\mathbf{V}_q = \mathbf{A}\sigma_{\mathbf{a}_q}^2 + \mathbf{G}\sigma_{\gamma_q}^2 + 2\text{Cov}(\mathbf{a}_q, \gamma_q) + \mathbf{I}_g\sigma_{\mathbf{e}_q}^2$ , y  $\text{Cov}(\mathbf{a}_q, \gamma_q) = G\sigma_{\gamma_q}^2$  (i.e., la covarianza entre  $\mathbf{a}_q$  y  $\gamma_q$  es igual a la varianza de  $\gamma_q$ );  $\gamma_q$ ,  $\mathbf{G}$  y  $\sigma_{\gamma_q}^2$  se definieron anteriormente;  $\mathbf{Z}$  es una matriz de incidencia (generalmente una matriz identidad  $g \times g$ ) y  $\mathbf{e}_q \sim \text{NMV}(\mathbf{0}, \mathbf{I}_g\sigma_{\mathbf{e}_q}^2)$  es un vector  $g \times 1$  de residuos;  $\mathbf{I}_g$  es una matriz identidad  $g \times g$  y  $\sigma_{\mathbf{e}_q}^2$  es la varianza de los residuos. Al modelo de la Ecuación 1 se le llamará modelo combinado.

**Distribución posterior conjunta de  $\mathbf{a}_q$  y  $\gamma_q$**

La distribución posterior conjunta de  $\mathbf{a}_q$  y  $\gamma_q$  puede escribirse como:

$$P(\mathbf{a}_q, \gamma_q / \mathbf{y}_q) \propto P(\mathbf{y}_q / \mathbf{a}_q, \gamma_q)P(\mathbf{a}_q / \gamma_q)P(\gamma_q) \quad (2)$$

donde el símbolo “ $\propto$ ” indica que  $P(\mathbf{a}_q, \gamma_q / \mathbf{y}_q)$  puede escribirse como el producto de la función de verosimilitud de  $\mathbf{y}_q$ ,  $P(\mathbf{y}_q / \mathbf{a}_q, \gamma_q) \propto \exp\left\{-\frac{1}{2}(\mathbf{y}_q - \mathbf{Z}\mathbf{a}_q - \mathbf{Z}\gamma_q)' \mathbf{R}^{-1}(\mathbf{y}_q - \mathbf{Z}\mathbf{a}_q - \mathbf{Z}\gamma_q)\right\}$ , la distribución condicional *a priori* de  $\mathbf{a}_q$  dado  $\gamma_q$ ,  $P(\mathbf{a}_q / \gamma_q) \propto \exp\left\{-\frac{1}{2}(\mathbf{a}_q - \gamma_q)' \mathbf{T}^{-1}(\mathbf{a}_q - \gamma_q)\right\}$  y la distribución *a priori* de  $\gamma_q$ ,  $P(\gamma_q) \propto \exp\left\{-\frac{1}{2}\gamma_q' \Phi^{-1}\gamma_q\right\}$ , donde  $\mathbf{R} = \mathbf{I}_g\sigma_{\mathbf{e}_q}^2$ ,  $\mathbf{T} = \mathbf{A}\sigma_{\mathbf{a}_q}^2 - \Phi$  y  $\Phi = \mathbf{G}\sigma_{\gamma_q}^2$ . Por las propiedades de la distribución NMV (Sorensen y Gianola, 2002),  $\gamma_q$  y  $\mathbf{T}$  son la esperanza y la varianza de  $\mathbf{a}_q / \gamma_q$ , respectivamente. Así, la Ecuación 2 es igual a:

$$P(\mathbf{a}_q, \gamma_q / \mathbf{y}_q) \propto \exp\left\{-\frac{1}{2}(\theta_q - \mathbf{D}\mathbf{d})' \mathbf{D}^{-1}(\theta_q - \mathbf{D}\mathbf{d})\right\} \quad (3)$$

El lado derecho de la Ecuación 3 es el núcleo de una distribución normal con media  $\mathbf{D}\mathbf{d}$  y varianza  $\mathbf{D}$ , donde

$$\theta'_q = [\mathbf{a}'_q \gamma'_q], \quad \mathbf{D}^{-1} = \begin{bmatrix} \mathbf{D}_{11}^{-1} & \mathbf{D}_{12}^{-1} \\ \mathbf{D}_{21}^{-1} & \mathbf{D}_{22}^{-1} \end{bmatrix}^{-1}, \quad \mathbf{D}_{11}^{-1} = \mathbf{R}^{-1} + \mathbf{T}^{-1},$$

$$\mathbf{D}_{12}^{-1} = \mathbf{D}_{21}^{-1} = \mathbf{R}^{-1} - \mathbf{T}^{-1}, \quad \mathbf{D}_{22}^{-1} = \mathbf{R}^{-1} + \mathbf{T}^{-1} + \Phi^{-1},$$

$$\mathbf{d} = \mathbf{1}_2 \otimes \mathbf{R}^{-1}\mathbf{y}_q, \quad \mathbf{1}_2 = [11] \text{ y “}\otimes\text{” denota al producto de Kronecker entre matrices (Langville y Stewart, 2004).}$$

model of Equation 1 will be called univariate combined linear model.

**Posterior joint distribution of  $\mathbf{a}_q$  and  $\gamma_q$**

The posterior joint distribution of  $\mathbf{a}_q$  and  $\gamma_q$  can be written as:

$$P(\mathbf{a}_q, \gamma_q / \mathbf{y}_q) \propto P(\mathbf{y}_q / \mathbf{a}_q, \gamma_q)P(\mathbf{a}_q / \gamma_q)P(\gamma_q) \quad (2)$$

where the symbol “ $\propto$ ” indicates that  $P(\mathbf{a}_q, \gamma_q / \mathbf{y}_q)$  can be written as the product of the likelihood function of  $\mathbf{y}_q$ ,  $P(\mathbf{y}_q / \mathbf{a}_q, \gamma_q) \propto \exp\left\{-\frac{1}{2}(\mathbf{y}_q - \mathbf{Z}\mathbf{a}_q - \mathbf{Z}\gamma_q)' \mathbf{R}^{-1}(\mathbf{y}_q - \mathbf{Z}\mathbf{a}_q - \mathbf{Z}\gamma_q)\right\}$ , the conditional *a priori* distribution of  $\mathbf{a}_q$  given  $\gamma_q$ ,  $P(\mathbf{a}_q / \gamma_q) \propto \exp\left\{-\frac{1}{2}(\mathbf{a}_q - \gamma_q)' \mathbf{T}^{-1}(\mathbf{a}_q - \gamma_q)\right\}$ , and the *a priori* distribution of  $\gamma_q$ ,  $P(\gamma_q) \propto \exp\left\{-\frac{1}{2}\gamma_q' \Phi^{-1}\gamma_q\right\}$ ,

$\mathbf{R} = \mathbf{I}_g\sigma_{\mathbf{e}_q}^2$ ,  $\mathbf{T} = \mathbf{A}\sigma_{\mathbf{a}_q}^2 - \Phi$  y  $\Phi = \mathbf{G}\sigma_{\gamma_q}^2$ . According to the properties of the NMV distribution (Sorensen and Gianola, 2002),  $\gamma_q$  and  $\mathbf{T}$  are the expectation and the variance of  $\mathbf{a}_q / \gamma_q$ , respectively. Thus, Equation 2 is equal to:

$$P(\mathbf{a}_q, \gamma_q / \mathbf{y}_q) \propto \exp\left\{-\frac{1}{2}(\theta_q - \mathbf{D}\mathbf{d})' \mathbf{D}^{-1}(\theta_q - \mathbf{D}\mathbf{d})\right\} \quad (3)$$

The right side of Equation 3 is the normal distribution kernel with mean  $\mathbf{D}\mathbf{d}$  and variance  $\mathbf{D}$ , where

$$\theta'_q = [\mathbf{a}'_q \gamma'_q], \quad \mathbf{D}^{-1} = \begin{bmatrix} \mathbf{D}_{11}^{-1} & \mathbf{D}_{12}^{-1} \\ \mathbf{D}_{21}^{-1} & \mathbf{D}_{22}^{-1} \end{bmatrix}^{-1}, \quad \mathbf{D}_{11}^{-1} = \mathbf{R}^{-1} + \mathbf{T}^{-1},$$

$$\mathbf{D}_{12}^{-1} = \mathbf{D}_{21}^{-1} = \mathbf{R}^{-1} - \mathbf{T}^{-1}, \quad \mathbf{D}_{22}^{-1} = \mathbf{R}^{-1} + \mathbf{T}^{-1} + \Phi^{-1},$$

$$\mathbf{d} = \mathbf{1}_2 \otimes \mathbf{R}^{-1}\mathbf{y}_q, \quad \mathbf{1}_2 = [11] \text{ and “}\otimes\text{” denotes the Kronecker product between matrices (Langville and Stewart, 2004).}$$

**Estimator of  $\theta_q$**

From Equation 3, the empirical Bayesian estimator of  $\theta_q = [\mathbf{a}'_q \gamma'_q]$  is:

$$\hat{\theta}_{BE_q} = \mathbf{D}\mathbf{d} \quad (4)$$

The components of variance:  $\sigma_{\mathbf{a}_q}^2$ ,  $\sigma_{\gamma_q}^2$  and  $\sigma_{\mathbf{e}_q}^2$ , can be estimated from the marginal distribution of  $\mathbf{y}_q$  using restricted maximum likelihood (Lynch and Walsh, 1998; Vattikuti et al., 2012).

**Estimador de  $\theta_q$**

Por la Ecuación 3, el estimador bayesiano empírico de  $\theta'_q = [\mathbf{a}'_q \ \gamma'_q]$  es:

$$\hat{\theta}_{BE_q} = \mathbf{D}\mathbf{d} \tag{4}$$

Los componentes de varianza  $\sigma_{\mathbf{a}_q}^2$ ,  $\sigma_{\gamma_q}^2$  y  $\sigma_{e_q}^2$  pueden estimarse por verosimilitud restringida a partir de la distribución marginal de  $\mathbf{y}_q$  (Lynch y Walsh, 1998; Vattikuti *et al.*, 2012).

**El modelo lineal multivariado**

Cuando se utilizan dos o más variables en la predicción del mérito genético, el modelo combinado de la Ecuación 1 puede escribirse como:

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{Z}\gamma + \mathbf{e} \tag{5}$$

donde, ahora,  $\mathbf{y}' = [\mathbf{y}'_1 \ \mathbf{y}'_2 \dots \mathbf{y}'_t] \sim \text{NMV}(0, \mathbf{V})$ ,  $\mathbf{a}' = [\mathbf{a}'_1 \ \mathbf{a}'_2 \dots \mathbf{a}'_t] \sim \text{NMV}(0, \mathbf{S})$ ,  $\gamma' = [\gamma'_1 \ \gamma'_2 \dots \gamma'_t] \sim \text{NMV}(0, \mathbf{\Omega})$  y  $\mathbf{e}' = [\mathbf{e}'_1 \ \mathbf{e}'_2 \dots \mathbf{e}'_t] \sim \text{NMV}(\mathbf{0}, \mathbf{\Psi})$  son vectores conformados por  $t$  subvectores  $g \times 1$  de observaciones ( $\mathbf{y}$ ), de efectos genéticos aditivos no explicados por los MM ( $\mathbf{a}$ ), de efectos genómicos aditivos ( $\gamma$ ), y de errores ( $\mathbf{e}$ ), respectivamente;  $\mathbf{V} = \mathbf{S} + 3\mathbf{\Omega} + \mathbf{\Psi}$ , donde  $\mathbf{S} = \mathbf{C} \otimes \mathbf{A}$ ,  $\mathbf{\Omega} = \mathbf{\Gamma} \otimes \mathbf{G}$  y  $\mathbf{\Psi} = \mathbf{E} \otimes \mathbf{I}_g$ ;  $\mathbf{C} = \{\sigma_{\mathbf{a}_{qi}}\}$  ( $q, i=1, 2, \dots, t$ ;  $t$ =número de variables) es la matriz de varianzas y covarianzas de los efectos genéticos aditivos no explicados por los MM ( $\mathbf{a}$ ),  $\mathbf{\Gamma} = \{\sigma_{\gamma_{qi}}\}$  es la matriz de varianzas y covarianzas de méritos genómicos aditivos ( $\gamma$ ),  $\mathbf{E} = \{\sigma_{e_{qi}}\}$  es la matriz de varianzas y covarianzas de los residuos;  $\mathbf{Z}$  es una matriz identidad (o de incidencia) de orden  $gt \times gt$ ;  $\mathbf{A}$ ,  $\mathbf{G}$  e  $\mathbf{I}_g$  se definieron en la Ecuación 1. Las matrices  $\mathbf{C} = \{\sigma_{\mathbf{a}_{qi}}\}$  y  $\mathbf{\Gamma} = \{\sigma_{\gamma_{qi}}\}$  pueden formarse a partir de las estimaciones de los componentes de varianzas  $\sigma_{\mathbf{a}_q}^2$ ,  $\sigma_{\gamma_q}^2$  y  $\sigma_{e_q}^2$ , y de las covarianzas respectivas (Vattikuti *et al.*, 2012).

**Estimación de  $\mathbf{a}$  y  $\gamma$**

Sea  $\theta' = [\mathbf{a}' \ \gamma']$ , donde  $\mathbf{a}' = [\mathbf{a}'_1 \ \mathbf{a}'_2 \dots \mathbf{a}'_t]$  y  $\gamma' = [\gamma'_1 \ \gamma'_2 \dots \gamma'_t]$  (Ecuación 5); la distribución posterior de  $\theta$  es similar a la distribución de  $\theta'_q = [\mathbf{a}'_q \ \gamma'_q]$  (Ecuación 3), así, el estimador bayesiano empírico de  $\theta$  es similar al de la Ecuación 4, *i.e.*,

$$\hat{\theta}_{BE} = \mathbf{D}\mathbf{d} \tag{6}$$

**The multivariate linear model**

When two or more traits are used in the prediction of the breeding value, the combined linear model from Equation 1 can be written as:

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{Z}\gamma + \mathbf{e} \tag{5}$$

where, now,  $\mathbf{y}' = [\mathbf{y}'_1 \ \mathbf{y}'_2 \dots \mathbf{y}'_t] \sim \text{NMV}(0, \mathbf{V})$ ,  $\mathbf{a}' = [\mathbf{a}'_1 \ \mathbf{a}'_2 \dots \mathbf{a}'_t] \sim \text{NMV}(0, \mathbf{S})$ ,  $\gamma' = [\gamma'_1 \ \gamma'_2 \dots \gamma'_t] \sim \text{NMV}(0, \mathbf{\Omega})$  y  $\mathbf{e}' = [\mathbf{e}'_1 \ \mathbf{e}'_2 \dots \mathbf{e}'_t] \sim \text{NMV}(\mathbf{0}, \mathbf{\Psi})$  are vectors made up of  $t$  sub-vectors  $g \times 1$  of observations ( $\mathbf{y}$ ), of additive genetic effects unexplained by the MM ( $\mathbf{a}$ ), of additive genomic effects ( $\gamma$ ), and of errors ( $\mathbf{e}$ ), respectively;  $\mathbf{V} = \mathbf{S} + 3\mathbf{\Omega} + \mathbf{\Psi}$ , where  $\mathbf{S} = \mathbf{C} \otimes \mathbf{A}$ ,  $\mathbf{\Omega} = \mathbf{\Gamma} \otimes \mathbf{G}$  and  $\mathbf{\Psi} = \mathbf{E} \otimes \mathbf{I}_g$ ;  $\mathbf{C} = \{\sigma_{\mathbf{a}_{qi}}\}$  ( $q, i=1, 2, \dots, t$ ;  $t$ =number of traits) is the matrix of variances and covariances of the additive genetic effects unexplained by the MM ( $\mathbf{a}$ ),  $\mathbf{\Gamma} = \{\sigma_{\gamma_{qi}}\}$  is the matrix of variances and covariances of additive genomic merits ( $\gamma$ ), and  $\mathbf{E} = \{\sigma_{e_{qi}}\}$  is the matrix of variances and covariances of the residues;  $\mathbf{Z}$  is an identity matrix (or of incidence) of order  $gt \times gt$ ;  $\mathbf{A}$ ,  $\mathbf{G}$  and  $\mathbf{I}_g$  are defined in Equation 1. The matrices  $\mathbf{C} = \{\sigma_{\mathbf{a}_{qi}}\}$  and  $\mathbf{\Gamma} = \{\sigma_{\gamma_{qi}}\}$  can be conformed from the estimations of the components of varianzas:  $\sigma_{\mathbf{a}_q}^2$ ,  $\sigma_{\gamma_q}^2$  and  $\sigma_{e_q}^2$ , and of the respective covariances (Vattikuti *et al.*, 2012).

**Estimation of  $\mathbf{a}$  and  $\gamma$**

Let  $\theta' = [\mathbf{a}' \ \gamma']$ , be a vector conformed by  $\mathbf{a}' = [\mathbf{a}'_1 \ \mathbf{a}'_2 \dots \mathbf{a}'_t]$  and  $\gamma' = [\gamma'_1 \ \gamma'_2 \dots \gamma'_t]$  (Equation 5); the posterior distribution of  $\theta$  is similar to the distribution of  $\theta'_q = [\mathbf{a}'_q \ \gamma'_q]$  (Equation 3), thus, the empirical Bayesian estimator of  $\theta$  is similar to Equation 4, *i.e.*,

$$\hat{\theta}_{BE} = \mathbf{D}\mathbf{d} \tag{6}$$

where, now, the components that make up the  $\mathbf{D}^{-1}$  matrix are:  $\mathbf{D}_{11}^{-1} = \mathbf{\Psi}^{-1} + (\mathbf{S} - \mathbf{\Omega})^{-1}$ ,  $\mathbf{D}_{12}^{-1} = \mathbf{D}_{21}^{-1} = \mathbf{\Psi}^{-1} - (\mathbf{S} - \mathbf{\Omega})^{-1}$  and  $\mathbf{D}_{22}^{-1} = \mathbf{\Psi}^{-1} + (\mathbf{S} - \mathbf{\Omega})^{-1} + \mathbf{\Omega}^{-1}$ ;  $\mathbf{d} = \mathbf{1}_2 \otimes \mathbf{\Psi}^{-1}\mathbf{y}$ ,  $\mathbf{\Psi}^{-1} = \mathbf{E}^{-1} \otimes \mathbf{I}_g$  and  $\mathbf{1}'_2 = [1 \ 1]$ .

**Prediction of the breeding value in the first cycle of selection**

In the first selection cycle the predictor of the breeding value of the candidates for selection ( $\hat{\theta}$ ) can be written as:

donde, ahora, los componentes que conforman la matriz  $\mathbf{D}^{-1}$  son:  $\mathbf{D}_{11}^{-1} = \Psi^{-1} + (\mathbf{S} - \Omega)^{-1}$ ,  $\mathbf{D}_{12}^{-1} = \mathbf{D}_{21}^{-1} = \Psi^{-1} - (\mathbf{S} - \Omega)^{-1}$  y  $\mathbf{D}_{22}^{-1} = \Psi^{-1} + (\mathbf{S} - \Omega)^{-1} + \Omega^{-1}$ ;  $\mathbf{d} = \mathbf{I}_2 \otimes \Psi^{-1} \mathbf{y}$ ,  $\Psi^{-1} = \mathbf{E}^{-1} \otimes \mathbf{I}_g$  y  $\mathbf{I}_2 = [1 \ 1]$ .

**Predicción del mérito genético en el primer ciclo de selección**

En el primer ciclo de selección el predictor del mérito genético de los candidatos a selección ( $\hat{\theta}$ ) puede escribirse como:

$$\hat{\theta} = \hat{\mathbf{a}} + \hat{\gamma} \tag{7}$$

donde  $\hat{\mathbf{a}}$  y  $\hat{\gamma}$  son subvectores de  $\hat{\theta}_{BE} = \mathbf{Dd}$  (Ecuación 6).

**Predicción del mérito genético después del primer ciclo de selección**

Para obtener los valores predichos de los candidatos a selección a partir del segundo ciclo de selección, es necesario estimar los valores del vector  $\mathbf{u}' = [\mathbf{u}'_1 \ \mathbf{u}'_2 \dots \mathbf{u}'_t]$  en la población base a partir de la igualdad  $\gamma = \mathbf{X}_t \mathbf{u}$ , donde  $\mathbf{X}_t = \mathbf{I}_t \otimes \mathbf{X}$ ,  $\mathbf{I}_t$  es una matriz identidad  $t \times t$  y  $\mathbf{X}$  es la matriz de valores codificados de los MM en la población base. Un estimador de  $\mathbf{u}$  en la población base es:

$$\hat{\mathbf{u}} = (\mathbf{X}'_t \mathbf{X}_t)^{-1} \mathbf{X}'_t \hat{\gamma} \tag{8}$$

donde  $\hat{\gamma}$  es el subvector de la Ecuación 6. Por la Ecuación 8, el predictor bayesiano empírico del mérito genético después del primer ciclo de selección es:

$$\hat{\theta}_l = \hat{\mathbf{a}} + \mathbf{W}_l \hat{\mathbf{u}} \tag{9}$$

donde  $\mathbf{W}_l = \mathbf{I}_t \otimes \mathbf{X}_l$  ( $l=2,3,\dots,N$ ;  $N$ =número de ciclos de selección),  $\mathbf{I}_t$  ya se definió y  $\mathbf{X}_l$  es la matriz de valores codificados de los MM obtenida en el ciclo de selección  $l$ . Así, desde el segundo ciclo de selección, en la Ecuación 9 sólo cambiarán los valores codificados de la matriz  $\mathbf{X}_l$ .

**Criterio para comparar la eficiencia de los modelos de predicción**

Como la precisión es igual a la correlación entre los valores predichos y observados, su valor máximo es 1. Suponga que  $\rho_c$  y  $\rho_g$  denotan, respectivamente, la precisión del modelo combinado y del modelo genómico, entonces:

$$\hat{\theta} = \hat{\mathbf{a}} + \hat{\gamma} \tag{7}$$

where  $\hat{\mathbf{a}}$  and  $\hat{\gamma}$  are sub-vectors of  $\hat{\theta}_{BE} = \mathbf{Dd}$  (Equation 6).

**Prediction of the breeding value after the first selection cycle**

In order to obtain the predicted values for the candidates for selection from the second selection cycle, it is necessary to estimate the values of the vector  $\mathbf{u}' = [\mathbf{u}'_1 \ \mathbf{u}'_2 \dots \mathbf{u}'_t]$  in the training population from the equation  $\gamma = \mathbf{X}_t \mathbf{u}$ , where  $\mathbf{X}_t = \mathbf{I}_t \otimes \mathbf{X}$ ,  $\mathbf{I}_t$  is the identity matrix  $t \times t$  and  $\mathbf{X}$  is the matrix of coded MM values in the training population. An estimator of  $\mathbf{u}$  in the training population is:

$$\hat{\mathbf{u}} = (\mathbf{X}'_t \mathbf{X}_t)^{-1} \mathbf{X}'_t \hat{\gamma} \tag{8}$$

where  $\hat{\gamma}$  is the sub-vector of Equation 6. From Equation 8, the empirical Bayesian predictor of the breeding value after the first selection cycle is:

$$\hat{\theta}_l = \hat{\mathbf{a}} + \mathbf{W}_l \hat{\mathbf{u}} \tag{9}$$

where  $\mathbf{W}_l = \mathbf{I}_t \otimes \mathbf{X}_l$  ( $l=2,3,\dots,N$ ;  $N$ =number of selection cycles),  $\mathbf{I}_t$  was defined previously, and  $\mathbf{X}_l$  is the matrix of coded MM values obtained in the selection cycle  $l$ . Thus, from the second selection cycle, the only thing that will change in Equation 9 will be the coded values in matrix  $\mathbf{X}_l$ .

**Criterion to compare the efficiency of the prediction models**

Since accuracy is equal to the correlation between predicted and observed values, their maximum value is 1. Assume that  $\rho_c$  and  $\rho_g$  denote the accuracy in the combined and the genomic linear model, respectively, then:

$$p = 100 \left( \frac{\rho_c}{\rho_g} - 1 \right) \tag{10}$$

is the efficiency (Bulmer, 1980) of the combined linear model with regard to the genomic linear model. Thus, if  $p=0$ , the **efficiency** of both models is equal ( $\rho_c = \rho_g$ );  $p>0$  si  $\rho_c > \rho_g$  (the **efficiency** of the combined model is greater than that of the genomic model) and if  $\rho_c < \rho_g$ ,  $p<0$  (the **efficiency** of the combined model is lower than that of the genomic model). Thus, Equation 10 allows determining the most efficient linear model to predict the genetic merit.



$$p = 100 \left( \frac{\rho_c}{\rho_g} - 1 \right) \quad (10)$$

es la eficiencia (Bulmer, 1980) del modelo combinado respecto al modelo genómico. Así, cuando  $p=0$  la **eficiencia** de ambos modelos es igual ( $\rho_c = \rho_g$ );  $p>0$  si  $\rho_c > \rho_g$  (la **eficiencia** del modelo combinado es mayor que la del modelo genómico) y  $p<0$  si  $\rho_c < \rho_g$  (la **eficiencia** del modelo combinado es menor que la del modelo genómico). Así, la Ecuación 10 permite determinar el modelo más adecuado, o más eficiente, para predecir el mérito genético.

## RESULTADOS Y DISCUSIÓN

### El modelo genómico está anidado en el modelo combinado

Uno de los resultados más importantes de la teoría de la SG es que la esperanza de la matriz de relaciones genómicas  $\mathbf{G}$  es igual a la matriz de relaciones numéricas  $\mathbf{A}$ , *i.e.*,  $E(\mathbf{G}) = \mathbf{A}$  (Habier *et al.*, 2007). Esto significa que  $\mathbf{G}$  es una realización particular de  $\mathbf{A}$  y que conforme el número de MM y genotipos se incrementa en la población base, el valor de  $\mathbf{G}$  se concentra cada vez más alrededor de  $\mathbf{A}$ , por lo que en el límite, puede asumirse que  $\mathbf{G}=\mathbf{A}$ . Lo mismo ocurre con la matriz de varianzas y covarianzas genómicas aditivas  $\mathbf{\Gamma}$  en relación con la matriz de varianzas y covarianzas genéticas aditiva  $\mathbf{C}$ . Es decir, conforme el número de MM y genotipos se incrementa, la matriz  $\mathbf{\Gamma}$  se aproxima cada vez más a  $\mathbf{C}$ , y en el límite  $\mathbf{\Gamma}=\mathbf{C}$ . Cuando  $\mathbf{G}=\mathbf{A}$  y  $\mathbf{\Gamma}=\mathbf{C}$ ,  $\mathbf{S}=\mathbf{\Omega}$  y las matrices que conforman la matriz  $\mathbf{D}^{-1}$ :  $\mathbf{D}_{11}^{-1} = \mathbf{\Psi}^{-1} + (\mathbf{S} - \mathbf{\Omega})^{-1}$ ,  $\mathbf{D}_{12}^{-1} = \mathbf{D}_{21}^{-1} = \mathbf{\Psi}^{-1} - (\mathbf{S} - \mathbf{\Omega})^{-1}$  y  $\mathbf{D}_{22}^{-1} = \mathbf{\Psi}^{-1} + (\mathbf{S} - \mathbf{\Omega})^{-1} + \mathbf{\Omega}^{-1}$ , se reducen a  $\mathbf{D}_{11}^{-1} = \mathbf{\Psi}^{-1}$ ,  $\mathbf{D}_{12}^{-1} = \mathbf{D}_{21}^{-1} = \mathbf{\Psi}^{-1}$  y  $\mathbf{D}_{22}^{-1} = \mathbf{\Psi}^{-1} + \mathbf{\Omega}^{-1}$ , y la matriz  $\mathbf{D}^{-1}$  se transforma en  $\begin{bmatrix} \mathbf{\Psi}^{-1} & \mathbf{\Psi}^{-1} \\ \mathbf{\Psi}^{-1} & \mathbf{\Psi}^{-1} + \mathbf{\Omega}^{-1} \end{bmatrix}^{-1}$ .

Esto indica que toda la información del mérito genético está concentrada en los efectos genómicos aditivos  $\gamma$  y que los valores del vector  $\mathbf{a}$  son nulos. En tal caso, el estimador bayesiano empírico  $\hat{\theta}_{BE} = \mathbf{Dd}$  (Ecuación 6) se convierte en el predictor del mérito genómico aditivo ( $\hat{\gamma}$ ) y puede denotarse como:

$$\hat{\gamma} = [\mathbf{\Psi}^{-1} + \mathbf{\Omega}^{-1}]^{-1} \mathbf{\Psi}^{-1} \mathbf{y} \quad (11)$$

## RESULTS AND DISCUSSION

### The genomic model is nested in the combined model

One of the most important results in the theory of GS is that the expectation of the genomic relationships matrix  $\mathbf{G}$  is equal to the numerical relationships matrix  $\mathbf{A}$ , *i.e.*,  $E(\mathbf{G}) = \mathbf{A}$  (Habier *et al.*, 2007). This means that  $\mathbf{G}$  is a particular realization of  $\mathbf{A}$  and that when the number of MM and genotypes increases in the training population, the value of  $\mathbf{G}$  will tend concentrated around  $\mathbf{A}$ , so that it can be assumed that at the limit,  $\mathbf{G}=\mathbf{A}$ . The same is true with the additive genomic variances and covariance matrix  $\mathbf{\Gamma}$  in relation to the additive genetic variances and covariances matrix  $\mathbf{C}$ . That is, when the number of MM and genotypes increases, the matrix  $\mathbf{\Gamma}$  approaches  $\mathbf{C}$ , and at the limit,  $\mathbf{\Gamma}=\mathbf{C}$ . When  $\mathbf{G}=\mathbf{A}$  and  $\mathbf{\Gamma}=\mathbf{C}$ ,  $\mathbf{S}=\mathbf{\Omega}$  and the matrices that make up the matrix  $\mathbf{D}^{-1}$ :  $\mathbf{D}_{11}^{-1} = \mathbf{\Psi}^{-1} + (\mathbf{S} - \mathbf{\Omega})^{-1}$ ,  $\mathbf{D}_{12}^{-1} = \mathbf{D}_{21}^{-1} = \mathbf{\Psi}^{-1} - (\mathbf{S} - \mathbf{\Omega})^{-1}$  and  $\mathbf{D}_{22}^{-1} = \mathbf{\Psi}^{-1} + (\mathbf{S} - \mathbf{\Omega})^{-1} + \mathbf{\Omega}^{-1}$ , are reduced to  $\mathbf{D}_{11}^{-1} = \mathbf{\Psi}^{-1}$ ,  $\mathbf{D}_{12}^{-1} = \mathbf{D}_{21}^{-1} = \mathbf{\Psi}^{-1}$  and  $\mathbf{D}_{22}^{-1} = \mathbf{\Psi}^{-1} + \mathbf{\Omega}^{-1}$ , and matrix  $\mathbf{D}^{-1}$  will be equal to  $\begin{bmatrix} \mathbf{\Psi}^{-1} & \mathbf{\Psi}^{-1} \\ \mathbf{\Psi}^{-1} & \mathbf{\Psi}^{-1} + \mathbf{\Omega}^{-1} \end{bmatrix}^{-1}$ .

This indicates that all breeding value information is concentrated in the additive genomic effects  $\gamma$  and that the values of vector  $\mathbf{a}$  are null. In such a case, the empirical Bayes estimator  $\hat{\theta}_{BE} = \mathbf{Dd}$  (Equation 6) becomes the predictor of the additive genomic merit ( $\hat{\gamma}$ ) and can be denoted as:

$$\hat{\gamma} = [\mathbf{\Psi}^{-1} + \mathbf{\Omega}^{-1}]^{-1} \mathbf{\Psi}^{-1} \mathbf{y} \quad (11)$$

This result indicates that the genomic linear model is a particular case of the combined linear model.

### The model with only phenotypic information is nested in the combined model

When the information of the MM is not used, matrix  $\mathbf{\Omega}$  is null and, in such a case,  $\hat{\theta}_{BE}$  becomes the predictor of the additive genetic effects ( $\hat{\mathbf{a}}$ ) and can be written as:

$$\hat{\mathbf{a}} = [\mathbf{\Psi}^{-1} + \mathbf{S}^{-1}]^{-1} \mathbf{\Psi}^{-1} \mathbf{y} \quad (12)$$

Este resultado indica que el modelo genómico es un caso particular del modelo combinado.

### El modelo con sólo información fenotípica está anidado en el modelo combinado

Cuando no se utiliza la información de los MM, la matriz  $\Omega$  es nula y, en tal caso  $\hat{\theta}_{BE}$  se convierte en el predictor de los efectos genético aditivos ( $\hat{\mathbf{a}}$ ) y puede escribirse como:

$$\hat{\mathbf{a}} = [\Psi^{-1} + \mathbf{S}^{-1}]^{-1} \Psi^{-1} \mathbf{y} \quad (12)$$

Esto demuestra que el modelo con sólo información fenotípica es un caso particular del modelo combinado. A  $\hat{\mathbf{a}}$  se le llamará predictor estándar.

### Precisión de los tres modelos de predicción

Los valores predichos del mérito genético de los candidatos a selección asociados a cada una de las tres variables de las dos poblaciones de maíz (Poblaciones 1 y 2) y de la población de trigo (Población 3) se denotaron como  $\hat{\theta}_1$ ,  $\hat{\theta}_2$  y  $\hat{\theta}_3$  para el modelo combinado (Ecuación 7);  $\hat{\gamma}_1$ ,  $\hat{\gamma}_2$  y  $\hat{\gamma}_3$  para el modelo genómico (Ecuación 11), y  $\hat{\mathbf{a}}_1$ ,  $\hat{\mathbf{a}}_2$  y  $\hat{\mathbf{a}}_3$  para el modelo estándar (Ecuación 12). Con los valores predichos y observados se calculó la precisión (correlación entre los valores predichos y observados) para cada una de las tres variables de los tres modelos; éstas se encuentran en el Cuadro 1.

### Evaluación numérica de los tres modelos de predicción

La eficiencia del modelo combinado respecto al modelo estándar y al modelo genómico; y la eficiencia del modelo estándar respecto al genómico, se evaluó por medio de la Ecuación 10 con los valores de las correlaciones presentadas en el Cuadro 1.

### Población de maíz 1

#### Eficiencia del modelo combinado en comparación con el modelo estándar

El valor de  $p$  (Ecuación 10) asociado a las correlaciones entre el rendimiento de grano (RG) y

This shows that the linear model with only phenotypic information is a particular case of the combined linear model. The  $\hat{\mathbf{a}}$  will be called standard predictor.

### Accuracy of the three prediction models

The predicted values of the breeding value of the candidates for selection associated to each one of the three traits of the two maize populations (Populations 1 and 2) and of the wheat population (Population 3) were denoted as  $\hat{\theta}_1$ ,  $\hat{\theta}_2$  and  $\hat{\theta}_3$  for the combined model (Equation 7);  $\hat{\gamma}_1$ ,  $\hat{\gamma}_2$  and  $\hat{\gamma}_3$  for the genomic model (Equation 11), and  $\hat{\mathbf{a}}_1$ ,  $\hat{\mathbf{a}}_2$  and  $\hat{\mathbf{a}}_3$  for the standard model (Equation 12). With the predicted and the observed values the accuracy (correlation between the predicted and observed values) was calculated for each one of the three traits of the three models; these are shown in Table 1.

### Numerical evaluation of the three prediction models

The efficiency of the combined model with regard to the standard model and the genomic model; and the efficiency of the standard model with regard to the genomic model, was evaluated through Equation 10 with the correlation values presented in Table 1.

### Maize population 1

#### Efficiency of the combined model compared with the standard model

The value of  $p$  (Equation 10) associated to the correlations between grain yield (GY) and its predicted values ( $\hat{\theta}_1$  and  $\hat{\mathbf{a}}$ ) was calculated as

$$p = 100 \left( \frac{0.883}{0.551} - 1 \right) = 60.2, \text{ where } 0.883 \text{ was the}$$

correlation between GY and  $\hat{\theta}_1$ , and 0.551 was the correlation between RG and  $\hat{\mathbf{a}}_1$ . Given that  $p=60.254$ , the efficiency of the combined model was 60.2 % higher than the efficiency of the standard model.

The value of  $p$  obtained from the correlation between the ear height (AM) and  $\hat{\theta}_2$  (0.767), and the correlation between AM and  $\hat{\mathbf{a}}_2$  (0.719), was

**Cuadro 1. Correlaciones obtenidas entre los valores predichos de los modelos estándar, genómico y combinado, y los valores de las observaciones de tres variables en dos poblaciones de maíz y una población de trigo.**  
**Table 1. Correlations obtained between the predicted values of the standard, genomic and combined models, and the values of the observations of three traits in two maize populations and one wheat population.**

Población	Variable	Modelo								
		Estándar			Genómico			Combinado		
		$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
Maíz 1	RG <sup>†</sup>	0.551			0.650			0.883		
	AM <sup>‡</sup>		0.719			0.130			0.767	
	AP <sup>§</sup>			0.229			0.123			0.830
Maíz 2	RG <sup>†</sup>	0.885			0.519			0.844		
	AM <sup>‡</sup>		0.772			0.610			0.866	
	AP <sup>§</sup>			0.559			0.614			0.675
Trigo 3	RG1 <sup>†</sup>	0.999			0.887			0.998		
	RG2 <sup>†</sup>		0.995			0.871			0.997	
	RG3 <sup>†</sup>			0.990			0.819			0.994

<sup>†</sup>Rendimiento de grano (Mg ha<sup>-1</sup>), <sup>‡</sup>Altura de mazorca (cm), <sup>§</sup>Altura de planta (cm) ❖ <sup>†</sup>Grain yield (Mg ha<sup>-1</sup>), <sup>‡</sup>Ear height (cm), <sup>§</sup>Plant height (cm).

sus valores predichos ( $\hat{\theta}_1$  y  $\hat{a}$ ) se calculó como  $p = 100 \left( \frac{0.883}{0.551} - 1 \right) = 60.2$ , donde 0.883 es la correlación entre RG y  $\hat{\theta}_1$ , y 0.551 es la correlación entre RG y  $\hat{a}_1$ . Como  $p=60.254$ , la eficiencia del modelo combinado es 60.2 % más alta que la eficiencia del modelo estándar.

El valor de  $p$  obtenido a partir de la correlación entre altura de mazorca (AM) y  $\hat{\theta}_2$  (0.767), y la correlación entre AM y  $\hat{a}_2$  (0.719), es  $p = 100 \left( \frac{0.767}{0.719} - 1 \right) = 6.7$ . Como  $p=6.7$ , el modelo combinado es 6.7 % más eficiente que el modelo estándar. Finalmente, el valor de  $p$  para altura de planta (AP) y sus valores predichos es igual a  $p = 100 \left( \frac{0.830}{0.229} - 1 \right) = 262.4$ , donde 0.830 es el valor de la correlación estimada entre AP y  $\hat{\theta}_3$ , y 0.229 es el valor de la correlación estimada entre RG y  $\hat{a}_3$ . En este último caso el modelo combinado es 262.4 % más eficiente que el modelo estándar.

El promedio de los tres valores de  $p$  obtenidos con las correlaciones entre los valores predichos y observados de las tres variables es igual a 109.8 %.

$p = 100 \left( \frac{0.767}{0.719} - 1 \right) = 6.7$ . Given that  $p=6.7$ , the combined model was 6.7 % more efficient than the standard model. Finally, the value of  $p$  for plant height (AP) and its predicted values was equal to

$$p = 100 \left( \frac{0.830}{0.229} - 1 \right) = 262.4, \text{ where } 0.830 \text{ was}$$

the value of the correlation estimated between AP and  $\hat{\theta}_3$ , and 0.229 was the value of the estimated correlation between RG and  $\hat{a}_3$ . In this last case, the combined model was 262.4 % more efficient than the standard model.

The average of the three values of  $p$  obtained with the correlations between the predicted and observed values of the three traits was equal to 109.8 %. This means that the combined model was more adequate for predicting the breeding value because it has an efficiency that is 1.1 greater than the standard model.

### Efficiency of the combined model compared with the genomic model

The average efficiency of the combined model was 366.9 % higher than the genomic

Esto significa que el modelo combinado es más adecuado para predecir el mérito genético debido a que tiene una eficiencia 1.1 veces mayor que el modelo estándar.

### **Eficiencia del modelo combinado comparado con el modelo genómico**

La eficiencia promedio del modelo combinado fue 366.9 % más alta que en el modelo genómico. Esto se debe a que los valores de las correlaciones entre AM y  $\hat{\gamma}_2$  (0.130) (Columna 7, Cuadro 1), y entre AP y  $\hat{\gamma}_3$  (0.123) (Columna 8, Cuadro 1) son muy bajos. Así, en este caso el modelo combinado tuvo una precisión 3.7 veces mayor que el modelo genómico, por lo cual puede asumirse que es más adecuado para predecir el mérito genético en este conjunto de datos.

### **Eficiencia del modelo estándar comparado con el modelo genómico**

De nuevo, debido a los valores bajos de las correlaciones estimadas entre AM y  $\hat{\gamma}_2$  (0.130) y entre AP y  $\hat{\gamma}_3$  (0.123), la eficiencia promedio del modelo estándar respecto al modelo genómico fue 174.7 % más alta. Es decir, el modelo genómico fue casi dos veces menos eficiente que el estándar. Esto se debe a que en la Población de maíz 1 el número de marcadores fue sólo de 199.

En resumen, el modelo combinado fue casi cuatro veces más eficiente que el modelo genómico y 1.1 más eficiente que el modelo estándar. Es evidente que el modelo combinado es más adecuado para predecir el mérito genético que los otros dos modelos en este conjunto de datos.

## **Población de maíz 2**

### **Eficiencia del modelo combinado comparado con el modelo estándar y con el modelo genómico**

Un procedimiento similar al de la Población de maíz 1 permite demostrar que la eficiencia promedio del modelo combinado fue 9.4 % más alta que en el modelo estándar, y 38.2 % más alta que en el modelo genómico, respectivamente. A pesar de que el número de marcadores aumentó relativamente poco en la Población de maíz 2 (sólo 60 MM más que en

model efficiency. This is because the estimated correlation values between AM and  $\hat{\gamma}_2$  (0.130) (Column 7, Table 1), and between AP and  $\hat{\gamma}_3$  (0.123) (Column 8, Table 1) were the lowest. In this case the combined model efficiency was 3.7 higher than the genomic model efficiency, so it is more adequate for predicting the breeding value in this set of data.

### **Efficiency of the standard model compared with the genomic model**

Again, due to the low estimated correlation values between AM and  $\hat{\gamma}_2$  (0.130) and between AP and  $\hat{\gamma}_3$  (0.123), the average efficiency of the standard model with regard to the genomic model was 174.7 % higher. That is, the genomic model was almost twice less efficient than the standard model. This is because in the maize Population 1 the number of markers was only 199.

In short, the combined model was almost 4 times more efficient than the genomic model, and 1.1 more efficient than the standard model. It is evident, then, that the combined model is more adequate to predict the breeding value than the other two models in this set of data.

## **Maize population 2**

### **Efficiency of the combined model compared with the standard model and with the genomic model**

A similar procedure to the one performed with maize Population 1 allows demonstrating that the average efficiency of the combined model was 9.4 % higher than the average efficiency of the standard model, and 38.2 % higher than in the genomic model, respectively. Although the number of markers increased relatively little in the maize Population 2 (only 20 MM more than in the maize Population 1), the efficiency of the combined model with regard to the genomic one was only 38.2 % higher, which indicates that the increase in the number of markers increased the efficiency of the genomic model. However, the combined model was more efficient than the other two models, so it is also advisable to use it to predict the breeding value in this set of data.

la Población de maíz 1), la eficiencia del modelo combinado respecto al genómico es sólo 38.2 % más alta, lo cual indica que el incremento del número de marcadores aumentó la eficiencia del modelo genómico. Sin embargo, el modelo combinado fue más eficiente que los otros dos modelos, por lo que también es recomendable utilizarlo para predecir el mérito genético en este conjunto de datos.

### **Eficiencia del modelo estándar comparado con el modelo genómico**

La eficiencia promedio del modelo estándar fue sólo 29.4 % superior a la del modelo genómico. Tal resultado indica que el aumento en el número de MM incrementó la eficiencia del modelo genómico.

### **Población 3**

#### **Eficiencia del modelo combinado comparado con el modelo estándar y con el modelo genómico**

La eficiencia promedio del modelo combinado respecto al modelo estándar y al genómico fue sólo 0.2 y 16.15 %, respectivamente. Debido al número de MM (1279) y de genotipos (599) en la Población 3, la eficiencia del modelo combinado respecto al modelo estándar y genómico fue mucho menor que en las Poblaciones de maíz 1 y 2. En este caso, tanto el modelo combinado como el modelo estándar podrían ser adecuados para predecir el mérito genético.

Los resultados de la Población 3 se deben a que la precisión del modelo estándar (Ecuación 12) es muy alta (Cuadro 1) porque el rendimiento de grano provino de una especie autógama. Por lo tanto, aunque el número de MM es grande, éstos contribuyeron muy poco a la precisión del modelo combinado.

#### **Eficiencia del modelo estándar comparado con el modelo genómico**

Aunque el número de MM fue relativamente alto, la eficiencia promedio del modelo estándar fue superior a la del modelo genómico en 15.9 %. Como ya se indicó, esto se debe a que la precisión del modelo estándar para esta población es muy

#### **Efficiency of the standard model compared with the genomic model**

The average efficiency of the standard model was only 29.4 % higher than that of the genomic model. This result indicates that the increment in the number of MM increased the efficiency of the genomic model.

### **Population 3**

#### **Efficiency of the combined model compared with the standard model and with the genomic model**

The average efficiency of the combined model with regard to the standard model and the genomic one was only 0.2 and 16.15 %, respectively. Due to the number of MM (1279) and of genotypes (599) in Population 3, the efficiency of the combined model with regard to the standard model and genomic model was lower than in maize Populations 1 and 2. In this case, both, the combined model and the standard model, could be adequate for predicting the breeding value.

The results from Population 3 are explained because the accuracy of the standard model (Equation 12) is very high (Table 1) due to the grain yield coming from an autogamous species. Therefore, although the number of MM is large, these contribute very little to the accuracy of the combined model.

#### **Efficiency of the standard model compared with the genomic model**

Despite the number of MM being relatively high, the average efficiency of the standard model was higher than that of the genomic model by 15.9 %. As it has been indicated, this is because the accuracy of the standard model for this population is very high (Table 1). However, the correlations obtained in the genomic model between the predicted values and the observed values were higher in Population 3 than in maize Populations 1 and 2 (Table 1), which suggests that when increasing the number of MM, the precision of the genomic model also increased.

According to the results of the three prior populations, the combined model was in general



alta (Cuadro 1). Sin embargo, las correlaciones obtenidas en el modelo genómico entre los valores predichos y los observados fueron más altas en la Población 3 que en las Poblaciones de maíz 1 y 2 (Cuadro 1). Esto sugiere que al aumentar el número de MM, la precisión del modelo genómico se incrementó.

De acuerdo con los resultados de las tres poblaciones anteriores, el modelo combinado fue más eficiente que los otros dos modelos, aunque, conforme el número de marcadores y genotipos aumentó, la eficiencia del modelo combinado respecto al modelo genómico se redujo. La eficiencia observada del modelo combinado en los resultados de las tres poblaciones debe atribuirse a que usa en la predicción dos fuentes de información: fenotípica y genómica. Entonces, si se usa en el primer ciclo de selección, la precisión de la selección en tal ciclo aumentará.

#### **Ventaja del modelo genómico respecto al modelo estándar**

La manera usual de predecir el mérito genético en plantas y animales en la SG es sustituir la matriz de relaciones numéricas (**A**) por la matriz de relaciones genómicas (**G**) en las ecuaciones de predicción. Por ello, la ecuación de predicción del modelo genómico (Ecuación 11) y del modelo estándar (Ecuación 12), son formalmente equivalentes. Cuando el número de MM y genotipos es grande, ambos modelos tienden a proporcionar predicciones que se asemejan cada vez más (Cuadro 1, Población 3). Sin embargo, la ventaja del modelo genómico respecto al estándar radica en la posibilidad de reducir los intervalos entre ciclos de selección en más de dos tercios. Así, el modelo genómico es más eficiente que el modelo estándar cuando la eficiencia se mide por año y no por ciclo de selección. Según Beyene *et al.* (2015), la selección genómica requiere 1.5 años para completar un ciclo de selección, mientras que la selección fenotípica requiere 4 años por cada ciclo de selección.

#### **Importancia del modelo combinado**

Existen varios métodos bayesianos (Gianola, 2013) y no bayesianos (VanRaden, 2008) para predecir el mérito genético en el contexto univariado bajo

more efficient than the other two models, although, as the number of markers and genotypes increased, the efficiency of the combined model with regard to the genomic model was reduced. The efficiency observed from the combined model in the results of the three populations must be necessarily attributed to the fact that the model used two sources of information in the prediction: phenotypic and genomic. Then if the combined model is used in the first selection cycle, the precision of the selection in that cycle will increase.

#### **Advantage of the genomic model with regard to the standard model**

In the GS the usual way of predicting the plant and animal breeding values in breeding programs is to substitute the numerical relationships matrix (**A**) by the genomic relationships matrix (**G**) in the prediction equations. Therefore, the prediction equation of the genomic model (Equation 11) and the standard model (Equation 12), are formally equivalent. When the number of MM and genotypes is large, both models tend to provide predictions that are increasingly more similar (Table 1, Population 3). However, the advantage of the genomic model with regard to the standard model lies in the possibility of reducing the intervals between selection cycles in more than two thirds. Thus, the genomic model is more efficient than the standard model when the efficiency is measured per year and not per selection cycle. According to Beyene *et al.* (2015), the genomic selection requires 1.5 years to complete a selection cycle, while the phenotypic selection requires 4 years for each selection cycle.

#### **Importance of the combined model**

There are several Bayesian (Gianola, 2013) and non-Bayesian (VanRaden, 2008) methods to predict the breeding value in the univariate context under the assumption that the number of genotypes and MM is sufficiently large in the base population. In practice, however, not all the candidates for selection (plants or animals) have molecular markers. Therefore, a model such as the one proposed could be easily adapted to this case, thus increasing the precision in the prediction.

el supuesto de que el número de genotipos y MM es suficientemente grande en la población base. En la práctica, sin embargo, no todos los candidatos a selección (plantas o animales) cuentan con marcadores moleculares. Por ello, un modelo como el propuesto podría adaptarse fácilmente a este caso, aumentando así la precisión de la predicción.

### Bayes empírico comparado con GBLUP

Debido a que los efectos de los MM tienen distribución normal multivariada, Bayes empírico y GBLUP debería proporcionar resultados muy similares (Robinson, 1991) cuando se usa el mismo modelo de predicción. Esto se debe a que los supuestos de GBLUP y Bayes empírico son básicamente los mismos y porque, cuando las varianzas de los parámetros son conocidas, GBLUP se considera un caso particular de los métodos bayesianos (Blasco, 2001).

Finalmente, ¿cómo predecir el mérito genético? ¿por medio del Bayes empírico propuesto, por GBLUP o con alguna de las aproximaciones bayesianas existentes? Los modelos bayesianos estándares proporcionan un mejor control de la incertidumbre asociada a la predicción del mérito genético (de los Campos *et al.*, 2013, Gianola, 2013), lo cual se consigue con mucho trabajo de cómputo (Verbyla *et al.*, 2009). GBLUP, por su parte, requiere el conocimiento de las varianzas de los parámetros para que sus predicciones sean insesgadas; cuando tales varianzas son desconocidas, las propiedades estadísticas de las predicciones de GBLUP son también desconocidas (Gianola, 2013). De acuerdo con Blasco (2001), la elección de un modelo de predicción sobre otro debería estar basada en que el modelo elegido ofrezca una solución que los otros no ofrecen, de la facilidad para resolver el problema, y de la confianza en sus resultados. Este último punto es el de mayor importancia, ya que si el investigador se siente cómodo con un determinado método, significa que conoce sus limitaciones y ventajas y sabe qué esperar del modelo al utilizarlo en un análisis estadístico específico.

### CONCLUSIONES

El modelo propuesto, con la información conjunta del pedigrí y genómica en el contexto Bayes

### Empirical Bayes compared with GBLUP

Because the MM effects have a multivariate normal distribution, empirical Bayes and GBLUP should give very similar results (Robinson, 1991) when the same prediction model is used. This is because the assumptions of GBLUP and empirical Bayes are basically the same and because, when the variances of the parameters are known, GBLUP is considered a particular case of the Bayesian methods (Blasco, 2001).

Finally, how could the breeding value be predicted? Through the empirical Bayes proposed, through GBLUP or with some of the existing Bayesian approximations? The standard Bayesian models provide a better control of the uncertainty associated to the prediction of the breeding value (de los Campos *et al.*, 2013, Gianola, 2013), which is attained with much computation work (Verbyla *et al.*, 2009). GBLUP, in turn, requires the knowledge of variances of the parameters; when these variances are unknown, the statistical properties of GBLUP are also unknown (Gianola, 2013). According to Blasco (2001), the election of one prediction model over another one should be based on the fact that the chosen model offers a solution that the others do not, on how ease is to solve the problem, and on the trust in their results. This last point has the greatest importance, since if the researcher feels comfortable with a specific method, it means that he/she knows its limitations and advantages and knows what to expect from the model when using it in a specific statistical analysis.

### CONCLUSIONS

The model proposed, with joint information on pedigree and genome within the empirical Bayes context, provided more precise predictions than the other two models because the predictions incorporate not only the phenotypic and genomic information, but also the genetic correlations between traits.

—End of the English version—



empírico, proporcionó predicciones más precisas que los otros dos modelos porque en la predicción se

incorporan la información fenotípica y genómica, y también las correlaciones genéticas entre caracteres.

### AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), por la Beca Postdoctoral otorgada a J. Jesus Ceron-Rojas a través del Programa de Estancias Posdoctorales Vinculadas al Fortalecimiento de la Calidad del Postgrado Nacional, Convocatoria 2013 (3), la cual hizo posible el presente estudio.

Al doctor José Crossa, jefe de la Unidad de Biometría y Estadística del Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT), quien gentilmente nos proporcionó los datos de las dos poblaciones de maíz y de la población de trigo para la validación numérica de los resultados teóricos del presente estudio.

### LITERATURA CITADA

- Blasco, A. 2001. The Bayesian controversy in animal breeding. *J. Anim. Sci.* 79: 2023-2046.
- Beyene, Y., K. Semagn, S. Mugo, A. Tarekegne, R. Babu, B. Meise, P. Sehabiague, D. Makumbi, C. Magorokosho, S. Oikeh, J. Gakunga, M. Vargas, M. Olsen, B. M. Prasanna, M. Banziger, and J. Crossa. 2015. Genetic gains in grain yield through genomic selection 1 in eight bi-parental maize populations under drought stress. *Crop Sci.* 55: 154-163.
- Bulmer, M. G. 1980. *The Mathematical Theory of Quantitative Genetics. Lectures in Biomathematics.* University of Oxford: Clarendon Press. 254 p.
- Calus, M. P. L., and R. F. Veerkamp. 2011. Accuracy of multi-trait genomic selection using different methods. *Genet. Selection Evol.* 43: 26. <http://www.gsejournal.org/content/43/1/26>. (Consulta: Febrero 2015).
- Casella, G., and E. I. George. 1992. Explaining the Gibbs sampler. *The Am. Stat.* 46: 167-174.
- de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193: 327-345.
- Gianola, D. 2013. Priors in whole-genome regression: the bayesian alphabet returns. *Genetics* 194: 573-596.
- Hayashi, T., and H. Iwata. 2013. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinf.* 14: 34.
- Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177: 2389-2397.
- Jia, Y., and J. L. Jannink. 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 192: 1513-1522.
- Langville, A. N., and W. J. Stewart. 2004. The Kronecker product and stochastic automata networks. *J. Comp. Appl. Math.* 167: 429-44.
- Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92: 4656-4663.
- Lehmann, E. L., and G. Casella. 1998. *Theory of Point Estimation.* 2nd Ed. Springer-Verlag New York. 589 p.
- Lynch, M., and B. Walsh. 1998. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Inc. Publisher Sunderland, Massachusetts, USA. 980 p.
- Massman, J. M., A. Gordillo, R. E. Lorenzana, and R. Bernardo. 2013. Genomewide predictions from maize single-cross data. *Theor. Appl. Genet.* 126: 13-22.
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819-1829.
- Robinson, G. K. 1991. That BLUP is a good thing: The estimation of random effects. *Stat. Sci.* 6: 15-51.
- Sorensen, D., and D. Gianola. 2002. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics.* Springer, New York. 740 p.
- Tempelman, R. J., and G. J. M. Rosa. 2004. Empirical Bayes approach to mixed model inference in quantitative genetics. *In: Saxto, A. M. (ed). Genetics Analysis of Complex Traits Using SAS.* Cary N.C., SAS Institute Inc. pp: 149-176.
- VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91: 4414-4423.
- Vattikuti, S., J. Guo, and C. C. Chow. 2012. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet* 8 (3): e1002637. DOI: 10.1371/journal.pgen.1002637.
- Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genet. Res. Camb.* 91: 307-311.
- Verbyla, K. L., P. J. Bowman, B. J. Hayes, and M. E. Goddard. 2010. Sensitivity of genomic selection to using different prior distributions. *MCM Proceeding* 4 (Supp 1) S5.