University of São Paulo
"Luiz de Queiroz" College of Agriculture

Predicting the performance of untested maize single cross hybrids based on information from genomic relationship matrix and genotype by environment interaction

Matheus Dalsente Krause

Dissertation presented to obtain the degree of Master in Science. Area: Genetics and Plant Breeding

Piracicaba
2018

Matheus Dalsente Krause
Agronomist

Predicting the performance of untested maize single cross hybrids based on information from genomic relationship matrix and genotype by environment interaction

Advisor:
Prof. Dr. **ANTONIO AUGUSTO FRANCO GARCIA**

Dissertation presented to obtain the degree of Master in Science. Area: Genetics and Plant Breeding

Piracicaba
2018

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Krause , Matheus Dalsente
Predicting the performance of untested maize single cross hybrids based on information from genomic relationship matrix and genotype by environment interaction / Matheus Dalsente Krause . – – Piracicaba, 2018 .
63 p.

Dissertação (Mestrado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Seleção genômica 2. GBLUP 3. Ensaios para múltiplos ambientes 4. Variância-covariância I. Título.

DEDICATORY

To my wife, Aline.

ACKNOWLEDGMENTS

- To the “Luiz de Queiroz” College of Agriculture - University of São Paulo (ESALQ/USP) for all support to develop my Master’s Degree and this project.
- To the Coordination for the Improvement of Higher Education Personnel (CAPES) for my scholarship.
- To my advisor Prof. Dr. Antonio Augusto Franco Garcia for believing in my potential as a graduate student.
- To Embrapa Maize and Sorghum, an important partner of this project.
- All students (and friends) in the Statistical Genetics Laboratory at ESALQ/USP. We have made a great team!
- To all professors and staff of Genetics Department and Exact Sciences Department at ESALQ/USP.
- To my parents for their support.

Thanks.

SUMMARY

RESUMO

Predição de híbridos simples de milho não avaliados com informações da matriz de parentesco realizada e interação genótipos por ambientes

A fenotipagem em ensaios de múltiplos ambientes (MET) tem papel importante para acessar a resposta diferencial de híbridos de milho em diferentes regiões alvo de melhoramento, o que se deve a interação genótipos por ambientes (GxE). Neste contexto, um modelo efetivo de seleção genômica (GS) para predição do desempenho de híbridos não avaliados em MET é essencial para maximizar os ganhos genéticos e alocar eficientemente o orçamento dos programas de melhoramento. Desta forma, os objetivos deste estudo foram (*i*) avaliar as acurácias preditivas de modelos GBLUP (do inglês, *Genomic Best Linear Unbiased Prediction*) na predição da produtividade de grãos de híbridos simples de milho tropical não avaliados, usando modelos genético-estatísticos que levam em consideração a interação GxE através de uma estrutura de variância-covariância (VCOV) do tipo fator analítico (FA) e (*ii*) investigar a utilidade da matriz de parentesco realizada em combinação com diferentes estruturas de VCOV para efeitos genéticos e de resíduos em diferentes níveis de ambientes em desbalanceamento. As predições foram realizadas em duas situações: (CV1) híbridos não avaliados em nenhum ambiente e (CV2) híbridos avaliados em alguns ambientes e em outros não. Foram fenotipados 156 híbridos simples de milho em 12 ambientes para a característica produtividade de grãos. O genótipo dos híbridos foi inferido com base nas informações de marcadores SNP (do inglês, *single nucleotide polymorphism*) das linhagens parentais, obtidos via GBS (do inglês, *genotyping-by-sequencing*). Modelos que contemplaram informações de ambientes relacionados apresentaram acurácia preditiva superior em relação aos modelos que ignoraram tal informação. Modelos com matriz de parentesco realizada e interação GxE mantiveram acurácias preditivas acima de 0,400 com até 66% dos ambientes em desbalanceamento (oito ambientes selecionados ao acaso), sendo superior em relação a modelos FA que não contemplaram informação da matriz de parentesco realizada. A modelagem dos efeitos genéticos apresentou melhores resultados que a modelagem dos efeitos residuais. Estes resultados destacam a importância de incluir a matriz de parentesco realizada e de estruturas de VCOV que permitam que os modelos levem em consideração informações de híbridos aparentados, assim como de ambientes correlacionados na predição de híbridos simples de milho tropical não avaliados. Os procedimentos e modelos utilizados neste estudo podem ser facilmente estendidos a outras culturas em que MET desempenha um papel importante no processo de melhoramento.

Palavras-chave: Seleção Genômica; GBLUP; Ensaios para Múltiplos Ambientes; Variância-Covariância

ABSTRACT

Predicting the performance of untested maize single cross hybrids based on information from genomic relationship matrix and genotype by environment interaction

Phenotyping in multi-environment trials (MET) plays an important role to access the differential response of maize hybrids across target breeding regions due to genotype by environment (GxE) interaction. In this context, an effective model of genomic selection (GS) to predict the performance of untested hybrids in MET is essential to maximize genetic gains and to efficiently allocated the breeding programs' budget. Therefore, the goals of this study were (*i*) to evaluate the predictive accuracies of GBLUP (Genomic Best Linear Unbiased Prediction) models to predict grain yield performance of unobserved tropical maize single-cross hybrids, using models that consider GxE interaction by fitting a factor analytic (FA) variance-covariance (VCOV) structure, and (*ii*) to investigate the usefulness of genomic relationship information in combination with different VCOV for genetics and residuals effects, under different levels of unbalanced environments. Predictions were performed for two situations: (CV1) untested hybrids, and (CV2) hybrids evaluated in some environments but missing in others. Phenotypic data of grain yield was measured in 156 maize single-cross hybrids at 12 environments. Hybrids genotypes were inferred based on their parents (inbred lines) via SNP (single nucleotide polymorphism) markers obtained from GBS (genotyping-by-sequencing). Models that borrowed information from correlated environments presented higher predictive accuracy over those that ignored it. Models with genomic relationship information and GxE interaction were able to keep predictive accuracies up 0.400 with less than 66% of missing environments (eight environments randomly selected), being superior than FA models that not accounted genomic information. Modeling genetic effects was more important than residuals effects. These results highlight the importance of including genomic relationship information and VCOV structures that allow models to borrow information from relatives, as well as from correlated environments for predictions of unobserved tropical maize single-cross hybrids. The procedures and models applied in this study can be easily extended to other crops in which MET plays an important role in the breeding process.

Keywords: Genomic Selection; GBLUP; Multi-Environment Trials; Variance-Covariance

1 INTRODUCTION

Maize (*Zea mays* L.) plays an important role in global food security, being a key crop for 460 million inhabitants in sub-Saharan Africa, Asia, and Latin America (Prasanna, 2016, p.62). Worldwide, in terms of animal supply chain, 60-70% of harvested maize is used as livestock feed (Gwirtz and Garcia-Casal, 2014). It is also has been important for biofuel production (Shiferaw *et al.*, 2011). By 2050 the expected population on Earth is 9.8 billion people (UN DESA, 2017), and whereas maize yields remain low in many developing countries, Brazil can be an important player to feed the world. Brazil already is a big maize producer ranking in the third position globally (USDA, 2017), and to properly face the task, the challenge to release fast superior cultivars to the market has intensified.

Genomic Selection (GS) has been recently incorporated into plant breeding programs (Jonas and de Koning, 2016) and is a promising tool to predict unobserved maize single-cross hybrids (Burgueño *et al.*, 2012; Crossa *et al.*, 2017). Proposed by Meuwissen *et al.* (2001), it consists in predicting the genetic merit of a genotype based on molecular markers information covering the whole genome. The availability of large-scale genomic information for most crops, due to cost-effective high-throughput sequencing technologies, is an important contributor for the success of GS (Crossa *et al.*, 2017). Molecular markers information can be used to estimate the identity-by-state relationship between pair of individuals (VanRaden, 2008; Powell *et al.*, 2010) and this information can be accounted by linear mixed models based on the genomic realized relationship matrix $\mathbf{A}$.

One of the most resource-demanding phase in a breeding program consists of hybrids phenotyping in multi-environment trials (MET) (Fritsche-Neto *et al.*, 2010). MET are crucial to access hybrids performance, allowing breeders to quantify the differential response of hybrids across target breeding regions or environments, phenomena known as genotype-by-environment (GxE) interaction. High levels of predictive accuracies have been found for GS models that incorporate both genomic information and GxE interaction (Jarquín *et al.*, 2014; Acosta-Pech *et al.*, 2017). Data from MET are usually unbalanced due to the natural process of selection; hybrids with poor performance are discarded and new entries are added every year (Piepho *et al.*, 2008; Dawson *et al.*, 2013). Therefore, GS models that properly deal with unbalanced data in MET leads to better predictive accuracies across environments.

The implementation of GS into breeding programs reshaped the breeder's equation; rather than genetic gains per cycle of selection (generation interval), gains per unit time/annual rate have been taken into account and more efficiently allocated budget in the breeding program (Heffner *et al.*, 2010; Hickey *et al.*, 2012). Thus, this dissertation have the goal of evaluating the flexibility of linear mixed models to account complex variance-covariance structures, considering genomic and MET information, under different levels of missing data across environments.

References

Acosta-Pech, R., J. Crossa, G. de los Campos, S. Teyssèdre, B. Claustres, S. Pérez-Elizalde, and P. Pérez-Rodríguez, 2017 Genomic models with genotype ×

environment interaction for predicting hybrid performance: an application in maize hybrids. Theoretical and Applied Genetics **130**: 1431–1440.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa, 2012 Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Science **52**: 707–719.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. de los Campos, J. Burgueño, J. M. González-Camacho, S. Pérez-Elizalde, Y. Beyene, S. Dreisigacker, R. Singh, X. Zhang, M. Gowda, M. Roorkiwal, J. Rutkoski, and R. K. Varshney, 2017 Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science **22**: 961–975.

Dawson, J. C., J. B. Endelman, N. Heslot, J. Crossa, J. Poland, S. Dreisigacker, Y. Manès, M. E. Sorrells, and J. L. Jannink, 2013 The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crops Research **154**: 12–22.

Fritsche-Neto, R., M. C. Gonçalves, R. Vencovsky, and C. L. D. Souza Junior, 2010 Prediction of genotypic values of maize hybrids in unbalanced experiments. Crop Breeding and Applied Biotechnology **10**: 32–39.

Gwirtz, J. A. and M. N. Garcia-Casal, 2014 Processing maize flour and corn meal food products. Annals of the New York Academy of Sciences **1312**: 66–75.

Heffner, E. L., A. J. Lorenz, J. L. Jannink, and M. E. Sorrells, 2010 Plant breeding with Genomic selection: Gain per unit time and cost. Crop Science **50**: 1681–1690.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos, 2012 Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Science **52**: 654–663.

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt, J. Lorgeou, F. Piraux, L. Guerreiro, P. Pérez, M. Calus, J. Burgueño, and G. de los Campos, 2014 A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics **127**: 595–607.

Jonas, E. and D. J. de Koning, 2016 Goals and hurdles for a successful implementation of genomic selection in breeding programme for selected annual and perennial crops. Biotechnology and Genetic Engineering Reviews **32**: 18–42.

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.

Piepho, H. P., J. Möhring, A. E. Melchinger, and A. Büchse, 2008 BLUP for phenotypic selection in plant breeding and variety testing. Euphytica **161**: 209–228.

Powell, J. E., P. M. Visscher, and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics **11**: 800–805.

Prasanna, B., 2016 Developing and Deploying Abiotic Stress-Tolerant Maize Varieties in the Tropics: Challenges and Opportunities. In *Molecular Breeding for Sustainable Crop Improvement*, edited by V. R. Rajpal, S. R. Rao, and S. Raina, volume 11, chapter 3, pp. 61–77.

Shiferaw, B., B. M. Prasanna, J. Hellin, and M. Bänziger, 2011 Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. Food Security **3**: 307–327.

UN DESA, 2017 World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100.

USDA, 2017 World Agricultural Supply and Demand Estimates.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414–4423.

2 LITERATURE REVIEW

2.1 Maize Breeding and Production in Brazil

The remarkable researchers Shull (Shull, 1908, 1909, 1911) and East (East, 1908), in the early twentieth century, discovered the superiority of the single-cross hybrid in their work with endogamy and hybridization, which is due to heterosis. In the 1960, the maize single-cross hybrid was introduced into the agriculture and gradually replaced open pollination varieties, mainly in the United States of America (Hallauer *et al.*, 2010, p.4). In Brazil, the first single-cross hybrid was developed by Agroceres in 1945, with inbred lines from Federal University of Viçosa in partnership with Agronomic Institute of Campinas (Fornasieri Filho, 2007).

In the 1990s, a new concept of maize production, called "safrinha" (second crop season), was introduced in Brazil. The main factor for its introduction was an innovation in the soil cultivation, called "Sistema de Plantio Direto" (no-till farming), which allows farmers to start early the first crop season (Fornasieri Filho, 2007). In this system, the first crop season is planted from September to November and the second from January to March. In the first crop season, plants have favorable growing conditions as the increase of temperature and rainfall plus a reduced intensity of plant disease and insect pests. On the other hand, these conditions are the opposite in the second crop season. From the end of January, the intensity of rainfall and averages temperatures decreases, and moreover, field crops have to face the spore load plus pest infestations not efficiently controlled from the first crop season. Hence, beyond the change in the system of production, breeding for these specific conditions are also a requirement.

Hybrids with broad adaptability and good stability are some of the factors that have improved the maize supply chain in Brazil. Over the past four decades, improvements of 225% in the average yield per hectare, of 362% in the country production and of 42% in the cultivated area, have established maize as the second most cultivated crop in Brazil (Conab, 2017b).

For the first and second maize crop seasons of 2016/2017, farmers in Brazil could choose the best cultivar among 315 registered at the Ministry of Agriculture, Livestock and Food Supply in Brazil (Filho and Borgui, 2016). Among all possibilities, 214 genotypes had some transgenic event (whether for resistance to herbicides, insect pests or both). Approximately 68% of cultivars were single-cross hybrids, 17% were three-way crosses, 6% were double cross and 9% were open pollinated cultivars and other materials. Regarding maturity groups, there was a predominance of early group (68%), followed by extra-early (23%) and intermediate maturity group (3%). The estimated production for these crop seasons was 91,468.4 million tons, grown on 17,077.1 million hectares, being the second crop season responsible for approximately 67% of the harvest (Conab, 2017a).

2.2 Linear Mixed Models in Plant Breeding

A linear mixed model is a statistical model containing both fixed and random effects, with the exception of the mean μ and the vector of residuals ϵ, respectively (Searle *et al.*, 1992; Galwey, 2006; Mrode, 2014). The general form of a linear mixed model is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

Where:

$\boldsymbol{y}$: is a n x 1 vector of phenotypes
$\boldsymbol{X}$: is the incidence matrix of fixed effects with dimension n x p
$\boldsymbol{b}$: is a p x 1 vector of fixed effects
$\boldsymbol{Z}$: is the incidence matrix for random effects with dimension n x q
$\boldsymbol{u}$: is a q x 1 vector of random effects (*e.g.*, breeding values of individuals)
$\boldsymbol{\epsilon}$: is a n x 1 vector of residuals

Both fixed ($\boldsymbol{b}$) and random ($\boldsymbol{u}$) effects can be computed by the mixed models equations (MME) presented by Henderson (1950) as follow:

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}y \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}y \end{bmatrix}$$

The estimator $\hat{\boldsymbol{u}}$ follows $\mathbf{u} \sim \text{NMV}(\mathbf{0}, \mathbf{G})$, being $\mathbf{G}$ the variance-covariance (VCOV) structure of individuals. Standard models assume unrelated individuals, then $\mathbf{G} = \mathbf{I}\sigma_a^2$, in which $\mathbf{I}$ is an identity matrix and σ_a^2 the variance component for additive effects. The vector of residuals $\boldsymbol{\epsilon}$ follows $\boldsymbol{\epsilon} \sim \text{NMV}(\mathbf{0}, \mathbf{R})$, being $\mathbf{R} = \mathbf{I}\sigma_\epsilon^2$ the VCOV structure for residuals. The generalized least squares solution for $\hat{\boldsymbol{b}}$ is the best linear unbiased estimator (BLUE), and for $\hat{\boldsymbol{u}}$ is the best linear unbiased predictor (BLUP). In practical sense, estimations of $\mathbf{G}$ and $\mathbf{R}$ comes prior to estimation of $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{u}}$, giving rise to empirical BLUE and BLUP (Piepho *et al.*, 2008).

Mixed models equations proposed by Henderson (1950, 1963) were initially applied for livestock and nowadays are widespread for animal breeding (Quaas and Pollak, 1980; Bienefeld *et al.*, 2007; VanRaden, 2008). In plant breeding, according to Piepho *et al.* (2008), its application have arrived later due several reasons, as the great availability of phenotypic records for the same line across environments, an unlikely situation in animal breeding. Specifically in maize breeding, mixed models have been used to select individuals and also to estimate variance components (Bernardo, 1996; Eller *et al.*, 2008; Piepho *et al.*, 2008; Arnhold *et al.*, 2009; Mi *et al.*, 2011; Poland *et al.*, 2011; DoVale and Fritsche-Neto, 2013), to account genotype by environment (GxE) interaction (Balestre *et al.*, 2009; Piepho, 2009; Burgueño *et al.*, 2011; Mendes *et al.*, 2012; Malosetti *et al.*, 2013) and to predict unobserved genotypes based on a genomic selection (GS) model (Bernardo and Yu, 2007; Piepho, 2009; Schrag *et al.*, 2009; Riedelsheimer *et al.*, 2012; Massman *et al.*, 2013; Albrecht *et al.*, 2014; Bernardo, 2014).

2.2.1 Modeling Variance-Covariance Structures and Relationship Matrix

The matrices $\mathbf{G} = \mathbf{I}\sigma_a^2$ and $\mathbf{R} = \mathbf{I}\sigma_\epsilon^2$ described above assumes independence and homogeneity of variance among individuals and residuals. In terms of genetic effects, this assumption may be unrealistic for individuals from the same breeding program. Different VCOV structures, some of them presented below, can be considered to best-fit an specific biological situation (Piepho *et al.*, 2008; Van Eeuwijk *et al.*, 2016).

(i) Identity (I)

It admits that all individuals are independent and with homogeneous variance. It is the simplest VCOV structure, usually used as default in statistical models. Analysis of variance (ANOVA) depends on this assumption.

$$\begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

(ii) Diagonal (D)

It admits independence and heterogeneous variances. In practice, can be interpreted as each hybrid has its own variance component.

$$\begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_J^2 \end{bmatrix}$$

(iii) Compound symmetry (CS)

It assumes homogeneous variances and the same pairwise correlation for all hybrids. Therefore, the estimation of two parameters are required.

$$\begin{bmatrix} \sigma_1^2 + \sigma^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma^2 & \dots & \sigma_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 + \sigma^2 \end{bmatrix}$$

(iv) First-order autoregressive (AR1)

It assumes homogeneous variances and correlations that decline exponentially with distance. It is commonly used for symmetrical measurements in space or time.

$$\begin{bmatrix} \sigma^2 & \rho & \rho^2 & \dots & \rho^{d(1,J)} \\ \rho & \sigma^2 & \rho & \dots & \rho^{d(2,J)} \\ \rho^2 & \rho & \sigma^2 & \dots & \rho^{d(3,J)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{d(J,1)} & \rho^{d(J,2)} & \rho^{d(J,3)} & \dots & \sigma^2 \end{bmatrix}$$

(v) Unstructured (UN)

It allows heterogeneous variances and covariances between individuals. The number of parameters to be estimated is $t(t+1)/2$, and as the number of environments t increase, fitting such model can become computational prohibitively and impractical (Kelly *et al.*, 2007).

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1J} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1J} & \sigma_{2J} & \dots & \sigma_J^2 \end{bmatrix}$$

(vi) Factor analytic of order k - FA(k)

Proposed by Piepho (1997, 1998) and Smith *et al.* (2001), it is a parsimonious structure that provides a good approximation to the unstructured matrix. It has been used in multi-environment trials (MET) analysis (Smith *et al.*, 2015). Ψ: specific variance for each environment; λ_j: the coefficient (or loading) for environment j.

$$\begin{bmatrix} \lambda_1^2 + \Psi_1 & \lambda_1\lambda_2 & \dots & \lambda_1\lambda_J \\ \lambda_2\lambda_1 & \lambda_2^2 + \Psi_2 & \dots & \lambda_2\lambda_J \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_J\lambda_1 & \lambda_J\lambda_2 & \dots & \lambda_J^2 + \Psi_J \end{bmatrix}$$

The inclusion of factor analytic structure in the framework of MET allows predictions to take into account information from correlated environments (Smith *et al.*, 2015). In an implicit model, GxE is included into the model in a single term that predicts each hybrid within an environments (i = 1, 2, ..., t hybrids within j = 1, 2, ..., J environments), together with an appropriate VCOV structure (Gezan *et al.*, 2016). In MET, the VCOV matrices of genetics and residuals effects are defined as Cov(g,g') = $\boldsymbol{G_0} \otimes \boldsymbol{I_g}$ and Cov(ϵ,ϵ') = $\boldsymbol{I_n} \otimes \boldsymbol{R_0}$, respectively, where $\boldsymbol{G_0}$ is a T x T matrix of genetic effects within environments, $\boldsymbol{I_g}$ is an identity matrix of order g that assumes independent and homogeneous variances between individuals, $\boldsymbol{I_n}$ is an identity matrix of order n that assumes independent and homogeneous variances among environments, $\boldsymbol{R_0}$ = Cov($\epsilon_{ij}, \epsilon_{ij'}$) is a J x J covariance matrix of residuals within environments and $\otimes$ is the Kronecker product. Both $\boldsymbol{G_0}$ and $\boldsymbol{R_0}$ can be modeled with different VCOV structures.

The matrices $\mathbf{G} = \mathbf{I}\sigma_a^2$ and Cov(g,g') = $\boldsymbol{G_0} \otimes \boldsymbol{I_g}$, for single and multiples environments, respectively, can be modeled by the numerator relationship matrix $\mathbf{A} = \{a(i, i^{'})\}$ (Henderson, 1976). The resemblance between relatives for any pair of individuals, for additive effects, is twice the coancestry coefficient ($\boldsymbol{f}_{i,i'}$) (Wright, 1921) multiplied by the variance component of additive effects σ_a^2. Therefore, $\mathbf{A} = 2[\boldsymbol{f}_{i,i'}]$ is the additive relationship matrix and $\boldsymbol{A}\sigma_a^2$ is the VCOV structure (Crossa *et al.*, 2006). The matrix $\mathbf{A}$ can be obtained by pedigree information (expected) (Bernardo, 1996; Crossa *et al.*, 2010; Albrecht *et al.*, 2014), by molecular markers (realized) (Habier *et al.*, 2009; Hayes *et al.*, 2009; Balestre *et al.*, 2010; Crossa *et al.*, 2010; Massman *et al.*, 2013; Albrecht *et al.*, 2014; Technow *et al.*, 2014) or both sources (Burgueño *et al.*, 2012; Albrecht *et al.*, 2014). The realized relationship matrix can be computed by the methodology proposed by VanRaden (2008).

In the past few years, the cost of molecular information have decreased due to cost-effective high-throughput sequencing technologies (Gorjanc *et al.*, 2017). The large availability of genomic information for several crops reshaped the breeding programs, bringing for breeders

modern technologies known as genomic selection (Meuwissen *et al.*, 2001), in which mixed models can also be applied.

2.3 Genomic Selection and Multi-Environment Trials

Proposed by Meuwissen *et al.* (2001), genomic selection (GS) is a form of marker-assisted selection (MAS) in which markers covering the whole genome are used to predict the genomic estimated breeding value (GEBV) of an individual. In practical sense, for GS be implemented in a breeding program it is necessary to split the program into three populations, named as: (*i*) Training Population (TRN); (*ii*) Validation or Testing Population (TST) and (*iii*) Breeding Population (Heffner *et al.*, 2009; Jannink *et al.*, 2010; Nakaya and Isobe, 2012; Desta and Ortiz, 2014; Ferrão *et al.*, 2017).

The first data set is the Training Population (TRN), where individuals must be genotyped and phenotyped for the traits of interest. There is no standard procedure for selecting which individuals should be included in this data set, but the guideline is superior genotypes or those of breeder's interest (Nakaya and Isobe, 2012; Desta and Ortiz, 2014; Ferrão *et al.*, 2017). The TRN is used to define the predictive model and to estimate the allelic effects, the later will be account to estimate the GEBV of individuals only with genomic information (Heffner *et al.*, 2009; Ferrão *et al.*, 2017).

The Validation or Testing Population (TST), slightly smaller than TRN, should also be genotyped and phenotyped for the traits of interest. Its aim is to evaluate the predictive accuracy of the model previously defined through the correlation between the GEBVs and the true phenotype value of the individuals (*e.g.* adjusted means) (Ferrão *et al.*, 2017). The GS model will be applied in the Breeding Population (just genotyped) and the GEBV of unobserved individuals will be predicted based on genomic information (Heffner *et al.*, 2009; Nakaya and Isobe, 2012; Desta and Ortiz, 2014). Therefore, the efficiency of predictions relies on the genetic relatedness between individuals from TRN and the Breeding Population (Nakaya and Isobe, 2012; Desta and Ortiz, 2014). GS models based on linear mixed models have been used in maize breeding programs with different purposes (Bernardo and Yu, 2007; Piepho, 2009; Schrag *et al.*, 2009; Riedelsheimer *et al.*, 2012; Massman *et al.*, 2013; Bernardo, 2014; Krchov and Bernardo, 2015) due to its simplicity and low computational demands (Heslot *et al.*, 2015).

Despite the gradual reduction of the cost of genotyping, the cost of phenotyping does not exhibit the same behavior (Bernardo, 2008; Krchov and Bernardo, 2015). The cost of one maize yield-trial plot is US$ 13.00, assuming two lines per plot and considering more than 2,001 plots evaluated (for less plots, the price should increases) (Tech Services Inc., 2018). The cost of GBS genotyping in a sequencing coverage (x) of 2x is US$ 25,00 per maize line (Gorjanc *et al.*, 2017). Bernardo and Yu (2007) have suggested that genotyping would be more feasible than phenotyping when the price of each marker be 5000 times lower than the price of a phenotypic data information. In this example, the value of each SNP marker is 52000 times less costly than the value of obtaining a phenotypic record. In addition, Krchov and Bernardo (2015) pointed out that due seasonal variation and labor required for field trials, genotyping is more convenient than phenotyping. Recent studies on GS based on mixed

models have presented predictive accuracies ranging from 0.40 to 0.90 for grain yield in maize (Technow *et al.*, 2014; Crossa *et al.*, 2017), and due to the high cost of phenotyping, its potential to reshape maize breeding programs is highlighted.

Phenotyping in MET plays an important role to access the performance of lines across target breeding zones (Resende *et al.*, 2012). The inclusion of GxE interaction into GS models can boost predictions up and contribute to a better understanding of GxE interaction (Burgueño *et al.*, 2012). Mixed models offers a great flexibility for considering different VCOV structures for both genetic and residuals effects in MET, allowing genetic correlations between environments and the best fit structure for residuals effects (Pastina *et al.*, 2012; Piepho *et al.*, 2012; Margarido *et al.*, 2015).

The GxE interaction is an important component of genetic variability and indicates that the performance of genotypes are directly affected by the environment. Environments with high positive genetic correlations will have fewer crossover interactions, that is no substantial changes in the rank of genotypes across environments, and more significant crossover interactions is expected when environments have low or negative correlations, causing reranking of genotypes across environments (Yan, 2016; Van Eeuwijk *et al.*, 2016). In maize breeding, GxE interaction can be used to select target breeding regions (Yan, 2016), to select hybrids for broad or specific environments (Fehr, 1993; Balestre *et al.*, 2009), to select inbred lines to be used as parents of newly synthetic populations (Yan *et al.*, 2000; Balestre *et al.*, 2009; Dias *et al.*, 2018), and recently, included into GS models for hybrids prediction (Bernardo and Yu, 2007; Burgueño *et al.*, 2012; Dos Santos *et al.*, 2016; Dias *et al.*, 2018).

In genomic selection, most results of GBLUP models were fitted for single-environment predictions (Zhang *et al.*, 2015; Cuevas *et al.*, 2016), which may lead for a lost of information. Modeling VCOV structures in the framework of mixed models allows to consider heterogeneity of variances and genetic correlation between environments, being a good approach to predict unobserved genotypes (Van Eeuwijk *et al.*, 2016). The factor analytic (Piepho, 1997, 1998; Smith *et al.*, 2001) VCOV structure leads to a good approximation over the unstructured matrix with less parameters to be estimated, being a parsimonious way to account GxE interaction (Smith *et al.*, 2015).

Recently, some studies have reported an increase in the predictive accuracy of maize hybrids when GxE is accounted (Crossa *et al.*, 2010; Technow *et al.*, 2014; Beyene *et al.*, 2015; Acosta-Pech *et al.*, 2017; Crossa *et al.*, 2017). Recalling the example of the costs of genotyping and phenotyping mentioned above, the budget needed for phenotyping 100 maize single-cross hybrids in 15 environments without replicates would be estimated at US$ 19,500. This value would be the same budget needed to genotype 780 inbred lines and with an appropriate GS model, to predict 303,810 single-cross hybrids. Therefore, GS models that account information from MET by allowing heterogeneous variances and covariances between genotypes and environments are promising for predictions of unobserved single-cross hybrids.

References

Acosta-Pech, R., J. Crossa, G. de los Campos, S. Teyssèdre, B. Claustres, S. Pérez-Elizalde, and P. Pérez-Rodríguez, 2017 Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theoretical and Applied Genetics **130**: 1431–1440.

Albrecht, T., H. J. Auinger, V. Wimmer, J. O. Ogutu, C. Knaak, M. Ouzunova, H. P. Piepho, and C. C. Schön, 2014 Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theoretical and Applied Genetics **127**: 1375–1386.

Arnhold, E., F. Mora, R. G. Silva, P. I. V. Good-God, and M. A. Rodovalho, 2009 Evaluation of top-cross popcorn hybrids using mixed linear model methodology . Chilean Journal of Agricultural Research **69**: 46–53.

Balestre, M., R. G. von Pinho, and J. C. Souza, 2010 Prediction of maize single-cross performance by mixed linear models with microsatellite marker information. Genetics and Molecular Research **9**: 1054–1068.

Balestre, M., R. G. Von Pinho, J. C. Souza, and R. L. Oliveira, 2009 Genotypic stability and adaptability in tropical maize based on AMMI and GGE biplot analysis. Genetics and Molecular Research **8**: 1311–1322.

Bernardo, R., 1996 Testcross additive and dominance effects in best linear unbiased prediction of maize single-cross performance. Theoretical and Applied Genetics **93**: 1098–1102.

Bernardo, R., 2008 Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Science **48**: 1649–1664.

Bernardo, R., 2014 Genomewide selection when major genes are known. Crop Science **54**: 68–75.

Bernardo, R. and J. Yu, 2007 Prospects for genomewide selection for quantitative traits in maize. Crop Science **47**: 1082–1090.

Beyene, Y., K. Semagn, S. Mugo, A. Tarekegne, R. Babu, B. Meisel, 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 in eight bi-parental maize populations under drought stress. Crop Science **55**: 154–163.

Bienefeld, K., K. Ehrhardt, and F. Reinhardt, 2007 Genetic evaluation in the honey bee considering queen and worker effects - A BLUP-animal model approach. Apidologie **38**: 77–85.

Burgueño, J., J. Crossa, J. M. Cotes, F. S. Vicente, and B. Das, 2011 Prediction assessment of linear mixed models for multienvironment trials. Crop Science **51**: 944–954.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa, 2012 Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Science **52**: 707–719.

Conab, C. N. d. A., 2017a Acompanhamento da safra brasileira de grãos.

Conab, C. N. d. A., 2017b Séries Históricas.

Crossa, J., J. Burgueño, P. L. Cornelius, G. McLaren, R. Trethowan, and A. Krishnamachari, 2006 Modeling genotype x environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Science **46**: 1722–1733.

Crossa, J., G. De Los Campos, P. Pérez, D. Gianola, J. Burgueño, J. L. Araus, D. Makumbi, R. P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H. J. Braun, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics **186**: 713–724.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. de los Campos, J. Burgueño, J. M. González-Camacho, S. Pérez-Elizalde, Y. Beyene, S. Dreisigacker, R. Singh, X. Zhang, M. Gowda, M. Roorkiwal, J. Rutkoski, and R. K. Varshney, 2017 Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science **22**: 961–975.

Cuevas, J., J. Crossa, V. Soberanis, S. Pérez-Elizalde, P. Pérez-Rodríguez, G. de los Campos, O. A. Montesinos-López, and J. Burgueño, 2016 Genomic Prediction of Genotype × Environment Interaction Kernel Regression Models. The Plant Genome **9**: 1–20.

Desta, Z. A. and R. Ortiz, 2014 Genomic selection: Genome-wide prediction in plant improvement. Trends in Plant Science **19**: 592–601.

Dias, K. O. D. G., S. A. Gezan, C. T. Guimarães, A. Nazarian, L. da Costa e Silva, S. N. Parentoni, P. E. de Oliveira Guimarães, C. de Oliveira Anoni, J. M. V. Pádua, M. de Oliveira Pinto, R. W. Noda, C. A. G. Ribeiro, J. V. de Magalhães, A. A. F. Garcia, J. C. de Souza, L. J. M. Guimarães, and M. M. Pastina, 2018 Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity .

Dos Santos, J. P. R., R. C. De Castro Vasconcellos, L. P. M. Pires, M. Balestre, and R. G. Von Pinho, 2016 Inclusion of dominance effects in the multivariate GBLUP model. PLoS ONE **11**: 1–21.

DoVale, J. C. and R. Fritsche-Neto, 2013 Genetic control of traits associated with phosphorus use efficiency in maize by REML/BLUP. Revista Ciência Agronômica **44**: 554–563.

East, E. M., 1908 Inbreeding in corn. Connecticut Agricultural Experimental Station Report pp. 419–28.

Eller, M. S., J. B. Holland, and G. A. Payne, 2008 Breeding for Improved Resistance To Fumonisin Contamination in Maize. Toxin Reviews **27**: 371–389.

Fehr, W. R., 1993 *Principles of Cultivar Development: Theory and Technique*. Macmillian Publishing Company, first edition.

Ferrão, L. F. V., R. Ortiz, and A. A. F. Garcia, 2017 Genomic Selection: State of the Art. In *Genetic Improvement of Tropical Crops*, edited by H. Campos and P. D. S. Caligari, chapter 2, pp. 19–54, Springer International Publishing AG.

Filho, I. A. P. and E. Borgui, 2016 Mercado de Sementes de Milho no Brasil Safra 2016/2017. Technical Report January, Embrapa Milho e Sorgo, Sete Lagoas.

Fornasieri Filho, D., 2007 *Manual da Cultura do Milho*. Funep, Jaboticabal.

Galwey, N. W., 2006 *Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance*. Wiley, first edition.

Gezan, S. A., M. P. de Carvalho, and J. Sherrill, 2016 Statistical methods to explore genotype-by-environment interaction for loblolly pine clonal trials. Tree Genetics and Genomes **13**: 1.

Gorjanc, G., J.-F. Dumasy, S. Gonen, R. C. Gaynor, R. Antolin, and J. M. Hickey, 2017 Potential of Low-Coverage Genotyping-by-Sequencing and Imputation for Cost-Effective Genomic Selection in Biparental Segregating Populations. Crop Science **57**: 1–17.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2009 Genomic selection using low-density marker panels. Genetics **182**: 343–353.

Hallauer, A. R., M. J. Carena, and J. B. Miranda Filho, 2010 *Quantitative Genetics in Maize Breeding*. Springer New York Dordrecht Heidelberg London, Ames.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research **182**: 343–353.

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009 Genomic Selection for Crop Improvement. Crop Science **49**: 1–12.

Henderson, C. R., 1950 Estimation of Genetic Parameters. Annals of Mathematical Statistics **21**: 309–310.

Henderson, C. R., 1963 Selection index and expected genetic advance. In *Statistical genetics and plant breeding*, p. 623, National Academy of Genetic Advance - National Research Council, Washington DC.

Henderson, C. R., 1976 A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics **32**: 69–83.

Heslot, N., J.-L. Jannink, and M. E. Sorrells, 2015 Perspectives for Genomic Selection Applications and Research in Plants. Crop Science **55**: 1–12.

Jannink, J.-L. L., A. J. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics **9**: 166–177.

Kelly, A. M., A. B. Smith, J. A. Eccleston, and B. R. Cullis, 2007 The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Science **47**: 1063–1070.

Krchov, L.-M. and R. Bernardo, 2015 Relative Efficiency of Genomewide Selection for Testcross Performance of Doubled Haploid Lines in a Maize Breeding Program. Crop Science **55**: 2091–2099.

Malosetti, M., J. M. Ribaut, and F. A. van Eeuwijk, 2013 The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. Frontiers in Physiology **4**: 1–17.

Margarido, G. R. A., M. M. Pastina, A. P. Souza, and A. A. F. Garcia, 2015 Multi-trait multi-environment quantitative trait loci mapping for a sugarcane commercial cross provides insights on the inheritance of important traits. Molecular Breeding **35**: 1–15.

Massman, J. M., A. Gordillo, R. E. Lorenzana, and R. Bernardo, 2013 Genomewide predictions from maize single-cross data. Theoretical and Applied Genetics **126**: 13–22.

Mendes, F. F., L. José, M. Guimarães, J. C. Souza, P. Evaristo, O. Guimarães, C. Antônio, P. Pacheco, J. Rodrigues, D. A. Machado, W. F. Meirelles, and A. Resende, 2012 Adaptability and stability of maize varieties using mixed model methodology. Crop Breeding and Applied Biotechnology **12**: 111–117.

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.

Mi, X., T. Wegenast, H. F. Utz, B. S. Dhillon, and A. E. Melchinger, 2011 Best linear unbiased prediction and optimum allocation of test resources in maize breeding with doubled haploids. Theoretical and Applied Genetics **123**: 1–10.

Mrode, R. a., 2014 *Linear Models for the Prediction of Animal Breeding Values*. CABI, third edition.

Nakaya, A. and S. N. Isobe, 2012 Will genomic selection be a practical method for plant breeding? Annals of Botany **110**: 1303–1316.

Pastina, M. M., M. Malosetti, R. Gazaffi, M. Mollinari, G. R. A. Margarido, K. M. Oliveira, L. R. Pinto, A. P. Souza, F. A. van Eeuwijk, and A. A. F. Garcia, 2012 A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. Theoretical and Applied Genetics **124**: 835–849.

Piepho, H.-P., 1997 Analyzing Genotype-Environment Data by Mixed Models with Multiplicative Terms. Biometrics **53**: 761–766.

Piepho, H. P., 1998 Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theoretical and Applied Genetics **97**: 195–201.

Piepho, H. P., 2009 Ridge regression and extensions for genomewide selection in maize. Crop Science **49**: 1165–1176.

Piepho, H. P., J. Möhring, A. E. Melchinger, and A. Büchse, 2008 BLUP for phenotypic selection in plant breeding and variety testing. Euphytica **161**: 209–228.

Piepho, H. P., J. Möhring, T. Schulz-Streeck, and J. O. Ogutu, 2012 A stage-wise approach for the analysis of multi-environment trials. Biometrical Journal **54**: 844–860.

Poland, J. A., P. J. Bradbury, E. S. Buckler, and R. J. Nelson, 2011 Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proceedings of the National Academy of Sciences of the United States of America **108**: 6893–6898.

Quaas, R. L. and E. J. Pollak, 1980 Mixed Model Methodology for Farm and Ranch Beef Cattle Testing Programs. Journal of Animal Science **51**: 1277–1287.

Resende, M. F. R., P. Muñoz, J. J. Acosta, G. F. Peter, J. M. Davis, D. Grattapaglia, M. D. V. Resende, and M. Kirst, 2012 Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. New Phytologist **193**: 617–624.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow, R. Sulpice, T. Altmann, M. Stitt, L. Willmitzer, and A. E. Melchinger, 2012 Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nature Genetics **44**: 217–220.

Schrag, T. A., J. Möhring, H. P. Maurer, B. S. Dhillon, A. E. Melchinger, H. P. Piepho, A. P. Sørensen, and M. Frisch, 2009 Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. Theoretical and Applied Genetics **118**: 741–751.

Searle, S. R., G. Casella, and C. E. McCulloch, 1992 *Variance Components*. John Wiley & Sons, Hoboken, New Jersey.

Shull, G. H., 1908 The Composition of a Field of Maize. American Breeders' Association pp. 296–301.

Shull, G. H., 1909 A Pure-Line Method in Corn Breeding. American Breeders' Association pp. 51–58.

Shull, G. H., 1911 Hybridization methods in corn breeding. American Breeders' Association pp. 63–72.

Smith, A., B. R. Cullis, and R. Thompson, 2001 Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics **57**: 1138–1147.

Smith, A. B., A. Ganesalingam, H. Kuchel, and B. R. Cullis, 2015 Factor analytic mixed models for the provision of grower information from national crop variety testing programs. Theoretical and Applied Genetics **128**: 55–72.

Tech Services Inc., 2018 Pricing Brochure TSI 2018 Test Sites.

Technow, F., T. A. Schrag, W. Schipprack, E. Bauer, H. Simianer, and A. E. Melchinger, 2014 Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics **197**: 1343–1355.

Van Eeuwijk, F. A., D. V. Bustos-Korts, and M. Malosetti, 2016 What should students in plant breeding know about the statistical aspects of genotype x Environment interactions? Crop Science **56**: 2119–2140.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414–4423.

Wright, S., 1921 Systems of Mating. I. the Biometric Relations between Parent and Offspring. Genetics pp. 111–126.

Yan, W., 2016 Analysis and handling of G x E in a practical breeding program. Crop Science **56**: 2106–2118.

Yan, W., L. A. Hunt, Q. Sheng, and Z. Szlavnics, 2000 Cultivar Evaluation and Mega-Environment Investigation Based on the GGE Biplot. Crop Science **40**: 597–605.

Zhang, X., P. Pérez-Rodríguez, K. Semagn, Y. Beyene, R. Babu, M. A. López-Cruz, F. San Vicente, M. Olsen, E. Buckler, J. L. Jannink, B. M. Prasanna, and J. Crossa, 2015 Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity **114**: 291–299.

3 PREDICTING THE PERFORMANCE OF UNTESTED MAIZE SINGLE-CROSS HYBRIDS BASED ON INFORMATION FROM GENOMIC RELATIONSHIP MATRIX AND GENOTYPE BY ENVIRONMENT INTERACTION

Keywords: Genomic Selection (GS); GBLUP; Multi-Environment Trials (MET); Variance-Covariance (VCOV).

3.1 Abstract

Genomic selection (GS) has been implemented in several commercial plant and animal breeding programs and it has proven to improve efficiency and maximize genetic gains. The inclusion of GxE interaction into GS models allows to borrow information through correlated environments and could boost hybrids prediction accuracy. The goals of this study were (*i*) to evaluate the predictive accuracies of GBLUP (Genomic Best Linear Unbiased Prediction) models to predict grain yield performance of unobserved tropical maize single-cross hybrids, using models that take into account GxE interaction by fitting a factor analytic (FA) variance-covariance (VCOV) structure, and (*ii*) to investigate the usefulness of genomic relationship information in combination with different VCOV for genetics and residuals effects, under different levels of unbalanced environments. Predictions were performed for two situations: (CV1) untested hybrids, and (CV2) hybrids evaluated in some environments but missing in others. Phenotypic data of grain yield was measured in 156 maize single-cross hybrids at 12 environments. Hybrids genotypes were inferred based on their parents (inbred lines) via SNP (single nucleotide polymorphism) markers obtained from GBS (genotyping-by-sequencing). Models that borrowed information from correlated environments presented higher predictive accuracy over those that ignored it. Models with genomic relationship information and GxE interaction were able to keep predictive accuracies up 0.400 with less than 66% of missing environments (eight environments randomly selected), being superior than FA models that not accounted genomic information. Modeling genetic effects was more important than residuals effects. These results highlights the importance of including genomic relationship information and variance-covariance structures that allows models to borrow information from relatives and related environments for predictions of unobserved tropical maize single-cross hybrids.

3.2 Introduction

Future prospects of population growth and rising demand for biologically derived products increases even more the role of releasing worldwide stable cultivars. In order to increase food security, commercial maize (*Zea mays* L.) breeding programs try to release as fast as possible promising cultivars to the market (Federizzi *et al.*, 2012), evaluating hybrids candidates across representative environments and years. Phenotyping in multi-environment trials (MET) plays an important role to access the performance of lines across target breeding regions (Oakey *et al.*, 2016) and is one of the most resource-demanding stage in the breeding program (Fritsche-Neto *et al.*, 2010). The differential responses of hybrids across environments is known as genotype-by-environment (GxE) interaction.

Breeders have commonly to identify superior hybrids across environments (Elias *et al.*, 2016). This is challenging especially due to the difficulties to understand if the progeny performance are related to just genetical effects or GxE interaction. Environments with high genetic correlations will have fewer crossover interactions, which can result in no substantial changes in the rank of genotypes across environments. More significant crossover interactions is expected when environments have low correlations, causing reranking of genotypes across environments (Van Eeuwijk *et al.*, 2016; Yan, 2016). Therefore, GxE interaction is an important component for hybrids evaluation.

The GxE interaction can be incorporated into a genomic selection (GS) model to predict the performance of untested genotypes in one or more target environments (Burgueño *et al.*, 2012; Lopez-Cruz *et al.*, 2015; Crossa *et al.*, 2016). The concept of GS was introduced by Meuwissen *et al.* (2001) for livestock and can be defined as a form of marker-assisted selection in which markers covering the whole genome are used to estimate genomic breeding values (GEBV) of the lines. GS has been implemented in a range of breeding programs (Jonas and de Koning, 2016) and has proved to facilitate the rapid selection of superior lines and to accelerate the breeding cycles (Crossa *et al.*, 2017), becoming an important tool to increase the annual rate of genetic gains (Heffner *et al.*, 2010; Hickey *et al.*, 2017). A key point for the success of GS is the large availability of cost-effective high-throughput sequencing technologies (Crossa *et al.*, 2017), resulting in the availability of large-scale genomic information for most crops, while the cost of phenotyping trends to increase (Bernardo, 2008; Krchov and Bernardo, 2015).

The first model proposed for maize single-cross hybrids prediction based on Best Linear Unbiased Prediction (BLUP) was implemented by Bernardo (1994) and did not account the GxE interaction. After, Bernardo (1995, 1996a,b) included pedigree information into the models obtained from the coancestry coefficient by the use of the expected relationship matrix **A**. Pedigree-based additive infinitesimal models (Fisher, 1918; Wright, 1921) rely on the concept of identity-by-descent (IBD), which refers to alleles that are descended from a common ancestor in a base population. Marker-based relationship matrix **A** estimates the identity-by-state relationship between pair of individuals (VanRaden, 2008; Powell *et al.*, 2010), which does not require any known pedigree. The marker-based **A** matrix inputs the main source of relatedness among individuals in linear mixed models in a modeling framework known as GBLUP (Genomic Best Linear Unbiased). Both matrices can be incorporated into mixed models for hybrids prediction, but the realized matrix has the advantage of capturing the Mendelian

sampling under the absence of inbreeding depression and assortative mating (Powell *et al.*, 2010; Burgueño *et al.*, 2012).

High levels of predictive accuracies have been found for genomic selection models that incorporate both genomic information from the realized genomic relationship matrix and GxE interaction (Jarquín *et al.*, 2014; Acosta-Pech *et al.*, 2017). One of the strategies to include GxE interaction is to modeling the genetic variance-covariance (VCOV) matrix **G** across environments. A common approach to model **G** is the unstructured matrix, that allows a genetic variance for each environment and different covariances among pairs of environments. However, this structure contains $p = t(t + 1)/2$ parameters (p) to be estimated, and as the number of environments t increase, fitting the model became computational prohibitively and impractical (Kelly *et al.*, 2007). An alternative way to overcome this difficult is to use the factor analytic structure (Piepho, 1997, 1998; Smith *et al.*, 2001) to fit the MET model, which requires a reduced number of parameters to be estimated. The FA structure has been used in several breeding programs and has shown a good applicability over the unstructured VCOV structure (Kelly *et al.*, 2007; Oakey *et al.*, 2016).

Unbalanced data from MET is a routine in plant breeding programs, resulting in difficulties for data analysis (Dawson *et al.*, 2013). The process of selection naturally discards lines with poor performance and, on the other hand, new entries are added every year (Piepho *et al.*, 2008). Historically, joint analysis of variance (ANOVA) and linear regression models were used to analyse and quantify GxE interaction (Elias *et al.*, 2016), in which genetic effects are assumed to come from the same distribution and share a homogeneous variance component (Ferrão *et al.*, 2017). Therefore, for hybrids yield prediction, GS models that properly deal with unbalanced data in MET would lead to better predictive accuracies across environments.

In genomic selection, most results of GBLUP models were fitted for single-environment predictions (Zhang *et al.*, 2015; Cuevas *et al.*, 2016). However, the inclusion of MET and modeling both genetics (**G**) and residuals (**R**) effects with an appropriated VCOV structure could boost genomic predictions accuracies. Therefore, the goals of this work were (*i*) to predict the performance of untested tropical maize single-cross hybrids for grain yield using GBLUP model in the framework of multi-environmental trials analyses, and (*ii*) to investigate the usefulness of genomic relationship information in combination with different variance-covariance structures for genetics and residuals effects, under different levels of unbalanced environments.

3.3 Materials and Methods

3.3.1 Experimental Data

The dataset was obtained by the maize breeding program of Embrapa Maize and Sorghum Brazilian public institution. Yield data was collected at two different crop seasons in Brazil in 2012, the first ranging from September to November, and the second from January to March, or field management (high yield inputs or following the standard recommendation for maize field). The population of 152 maize hybrids were split into three trials (T1, T2, and T3) evaluated side-by-side at eight different contrasting sites.

In the first crop season, plants have favorable growing conditions as the increase of temperature and rainfall, plus a reduced intensity of plant disease and insect pests. In the second crop season, these conditions were the opposite. From the end of January to the following months, the intensity of rainfall and averages temperatures are decreasing, and beyond this, field crops have to face the spore load plus pest infestations not efficiently controlled from the first crop season. The combination of locations, crop seasons and field management were designated as "environment", giving a total of 12 environments (Figure A.8). The trait under consideration is grain yield, in tons per hectare (t ha^{-1}), corrected to 13% of grain moisture.

The first two trials (T1 and T2) and the third (T3) evaluated 60 and 32 hybrids each, respectively. In the field, each trial was augmented by four common checks (commercial maize cultivars) and arranged as a balanced lattice square of 8x8 (T1 and T2) and 6x6 (T3), with 2 replications. These trials represent three different steps of the maize breeding program. The first two trials (T1 and T2) consists in 120 hybrids from an intermediate stage, and the third trial (T3) in 32 hybrids from an advanced stage of the Maize Breeding Program.

The amount of 156 hybrids comprises 149 single-crosses, two three-way crosses, one double cross and four commercial checks, being only the single-crosses under consideration for genomic selection. The single-crosses were obtained from 144 inbred lines, classified as dent (64 lines) and flint (77 lines) heterotic groups, and also another group C (3 lines), which combines well with both dent and flint sources. Four lines were used as testers from the opposite heterotic group to synthesize the most part of hybrids.

3.3.2 Genotypic Data

A panel of 680 inbred lines from Maize Breeding Program of Embrapa Maize and Sorghum were genotyped with the standard genotyping-by-sequencing (GBS) protocol (Elshire *et al.*, 2011) by the Genomic Diversity Facility at Cornell University (Ithaca, NY, USA). Tags were aligned to the B73 reference genome (AGPv3) (Schnable *et al.*, 2009). Standard quality controls were applied to the data, removing all non-bi-allelic markers, and single nucleotide polymorphisms (SNPs) were discarded if at least one of is true: the minor allele frequency (MAF) was lower than 5%; more than 20% of missing genotypes were found; and the inbreeding coefficient was lower than 0.8. The SNPs were called using the GBS pipeline available in the software TASSEL v.5 (Glaubitz *et al.*, 2014). After filtering, missing data were imputed using Beagle 4.1 (Browning and Browning, 2016). The number of SNPs per chromosome ranged from 1,951 (chromosome 10) to 5,024 (chromosome 1) and the final number of SNPs was

29,515. From these 680 lines, the lines used as parents of the single-cross hybrids of this study were selected. Then, for each SNP, the genotypes of the single-crosses hybrids were inferred based on the genotype of their parents (inbred line) in the software R version 3.4.3 (R Core Team, 2017). One of the 144 inbred lines used as parents was not genotyped, resulting in the availability of the genotypic information of 147 hybrids instead of 149 hybrids. Principal components analysis (PCA) of SNP matrix of the 143 inbred lines was performed in the software TASSEL v.5 (Glaubitz *et al.*, 2014) to verified the consistency of heterotic groups.

3.3.3 Statistical Models

All statistical-genetics models were fitted using the package ASReml-R (Butler *et al.*, 2009) by solving the mixed-model equations proposed by Henderson (1950). To solve the equations, variance components were estimated using the residual maximum likelihood (REML) (Patterson and Thompson, 1971) estimation method, by minimizing the residual likelihood function using the Average Information algorithm (Gilmour *et al.*, 1995).

3.3.3.1 Single-Environment Trial Analyses

Single-environment trial analyses within each environment was performed with the model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_1 b} + \boldsymbol{Z_2 g} + \boldsymbol{\epsilon} \tag{3.1}$$

where:
$\boldsymbol{y}$: is a n x 1 vector of phenotypes for m hybrids and j replicates
$\boldsymbol{X}$: is the incidence matrix of fixed effects coefficients with dimension n x j
$\boldsymbol{\beta}$: is a j x 1 vector of fixed effects of replicates
$\boldsymbol{Z_1}$: is the incidence matrix for random effects of blocks ($\boldsymbol{b}$), nested in replications, with dimension n x $r.j$
$\boldsymbol{b}$: is a $r.j$ x 1 vector of random effects of blocks within replications, where $\boldsymbol{b} \sim$ NMV($\mathbf{0}$, $\sigma_b^2\mathbf{I}$)
$\boldsymbol{Z_2}$: is the incidence matrix for random effects of hybrids ($\boldsymbol{g}$) with dimension n x g
$\boldsymbol{g}$: is a g x 1 vector of random effects of hybrids, where $\boldsymbol{g} \sim$ NMV($\mathbf{0}$, $\sigma_g^2\mathbf{I}$)
$\boldsymbol{\epsilon}$: is a n x 1 vector of residuals, where $\boldsymbol{\epsilon} \sim$ NMV($\mathbf{0}$, $\sigma_\epsilon^2\mathbf{I}$)

The generalized measure of heritability was estimated using $\hat{H}^2$ = 1 - [PEV/($2\sigma_g^2$)], where PEV (prediction error variance) is the mean variance of the difference between two genetic effects and σ_g^2 is the genetic variance (Cullis *et al.*, 2006). The coefficient of variation (CV %) was also estimated using CV % = $\frac{\sigma}{\mu}$ x 100, where σ is the square root of residual variance component (σ_ϵ^2) and μ is the average of grain yield of each trial within environment.

3.3.3.2 Genomic Selection and Multi-Environment Trials Analyses (MET)

In order to predict single-cross hybrids yield performance under multi-environment trials (MET), we fitted different models formulation, differing by their genetic (Σ_g) and residual (Σ_r) variance-covariance structures. Single-cross hybrids that did not have genotypic informa-

tion (2 of 156), tree-way and double crosses (3 of 156) and commercial checks (4 of 156) were considered as checks and modeled as fixed effects. The following model was fitted:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_1 b} + \boldsymbol{Z_2 g} + \boldsymbol{\epsilon} \tag{3.2}$$

where:
$\boldsymbol{y}$: is a n x 1 vector of phenotypes for m hybrids across s environments and q trials, where $n = \Sigma_{i=1}^{s} n_i$, in which n_i is the number of plots in environment s
$\boldsymbol{X}$: is the incidence matrix of fixed effects coefficients with dimension n x k
$\boldsymbol{\beta}$: is a k x 1 vector of fixed effects of environments, trials within environments, replicates within trials within environments, checks, and checks in environments
$\boldsymbol{Z_1}$: is the incidence matrix for random effects of blocks within replications within trials within environments, with dimension n x v
$\boldsymbol{b}$: is a v x 1 vector of random effects of blocks within replications within trials within environments, where $\boldsymbol{b} \sim \text{NMV}(\mathbf{0}, \sigma_b^2\mathbf{I})$
$\boldsymbol{Z_2}$: is the incidence matrix for random effects of hybrids within environments with dimension n x $g.s$
$\boldsymbol{g}$: is a g x 1 vector of random effects of hybrids within environments, where $\boldsymbol{g} \sim \text{NMV}(\mathbf{0}, \mathbf{G} \otimes \mathbf{A})$
$\boldsymbol{\epsilon}$: is a n x 1 vector of residuals within environments, where $\boldsymbol{\epsilon} \sim \text{NMV}(\mathbf{0}, \oplus_{i=1}^{s} \mathbf{I}_{n_i} \otimes \mathbf{R})$

The Kronecker product is denoted by $\otimes$ and $\mathbf{I}_{n_i}$ is an identity matrix mapping the presence and absence of some variance or covariance component in the variance-covariance matrix structure. The $\mathbf{A}$ is the realized genomic relationship matrix and $\mathbf{G}$ and $\mathbf{R}$ are VCOV structures of its effects and dimensions. Genomic selection models were evaluated for the presence of the genomic relationship matrix $\mathbf{A}$.

3.3.3.3 Variance-Covariance Structures in MET

The variance-covariance structures for random effects of hybrids and residuals within environments were defined as $\Sigma_g = \mathbf{G} \otimes \mathbf{I}$ or $\Sigma_r = \mathbf{I} \otimes \mathbf{R}$, respectively. Matrices $\mathbf{G}$ and $\mathbf{R}$ were both modeled with identity ($\mathbf{I}$) and diagonal ($\mathbf{D}$) structures. To account the correlation structure of genetic and GxE effects across environments, displayed by genes commonly expressed between pairs of environments, the $\mathbf{G}$ matrix was modeled using the factor analytic ($\mathbf{FA(k)}$) structure of order k. It was also tried to fitted models with the unstructured matrix for $\mathbf{G}$, however due to the number of environments, these models presented convergence problems.

Models were divided into three classes, (1) FA models without incorporating the genomic relationship matrix - models 1 to 8, (2) FA models with genomic relationship information - models 9 to 16, and (3) models assuming no correlation across environments of the GxE effects but including genomic relationship information - models 17 to 20 (Table 1). In the first class of models (models 1 to 8) hybrids were modeled as independents, being $\Sigma_g = (\mathbf{I} \otimes \mathbf{FA}(k))$, and in the second and third class of models (models 9 to 20) hybrids were modeled with the additive genomic relationship matrix $\mathbf{A}$, being $\Sigma_g = (\mathbf{A} \otimes \mathbf{FA}(k)\mathbf{,\ I,\ or\ D})$. For each k factor, the $\mathbf{R}$ matrix was modeled with identity or diagonal matrices, being $\Sigma_r = (\mathbf{I} \otimes \mathbf{I})$ or $\Sigma_r = (\mathbf{I} \otimes \mathbf{D})$, respectively (Table 1).

The second and third class of models considered the realized genomic relationship matrix **A**, computed using SNP markers from GBS, following the methodology described by VanRaden (2008) as,

$$\mathbf{A} = \frac{\mathbf{ZZ'}}{2\Sigma p_i(1-p_i)} \tag{3.3}$$

where **Z** = **M** - **P**, where **M** is the incidence matrix for markers considering two alleles (A and a) for a given ith marker locus, coded as 0, 1 and 2 for AA, Aa and aa, respectively, and **P** is derived from observed allele frequencies expressed as $\boldsymbol{P} = 2p_i$, where p_i is the MAF of locus i. The additive genomic relationship matrix **A** was estimated using the package AGHmatrix (Amadeu *et al.*, 2016) in software R version 3.4.3 (R Core Team, 2017).

The third class of models ignored the MET modeling, not borrowing informations from correlated environments. For these models, VCOV structures as identity or diagonal matrices were evaluated. The first and the second classes of models considered the MET modeling, taking account genetic and additive correlations between environments, respectively, using factor analytic structure of order k **(FA(k))** proposed by Piepho (1997, 1998) and Smith *et al.* (2001). Estimations of genetic variance and correlation matrices between environments, for **FA(k)** models, were obtained by $\hat{\mathbf{G}} = (\boldsymbol{\Lambda\Lambda'} + \boldsymbol{\Psi})$ and $\hat{\mathbf{C}} = \mathbf{DGD}$, respectively, where $\boldsymbol{\Lambda}$ is a s x k matrix of loading for all environments, Ψ is s x s diagonal matrix of specific variances of each environment and $\hat{\mathbf{D}}$ is a diagonal matrix of the inverse of the square roots of the diagonal values of $\hat{\mathbf{G}}$.

3.3.4 Models Selection Criteria

Two criteria were used to compare the models: (i) the goodness of fit via Akaike information criterion - AIC (Akaike, 1974) and Bayesian information criterion - BIC (Schwarz, 1978), and (ii) for **FA(k)** models, the overall percentage of genetic variance ($\bar{v}$) accounted, defined as $\bar{v} = 100\text{tr}\ (\boldsymbol{\Lambda\Lambda'})/\text{tr}\ (\boldsymbol{\Lambda\Lambda'} + \boldsymbol{\Psi})$, where "tr" is the trace of the matrix and the other terms were previously defined (Smith *et al.*, 2015). For first and second class of models, it was selected the best order (k) of FA structure to go forward with genomic predictions. For the third class of models, regardless the best-fit model, predictive accuracy was accounted for all models to quantify the influence of modeling VCOV matrices on genomic predictions. For AIC and BIC criteria, models that have the lowest values were considered. To verify the advantage of FA structure under **G**, the predictive accuracies of these models were compared with those of models that did not take into account information from correlated environments, hybrids within environments, or both.

3.3.5 Cross-Validation Schemes

Models were also compared based on their predictive accuracy, computed via Pearson correlation between genetic estimated breeding value (GEBV) and observed adjusted means (y_i), obtained from single-environment trial analysis without molecular marker information. Two distinct cross-validation strategy, CV1 and CV2, were implemented as proposed by Burgueño *et al.* (2012). In the first case (CV1), hybrids from validation set were deleted in all

environments and predictions were performed based on the phenotypic information from other hybrids, through the realized genomic relationship matrix (**A**). The second strategy (CV2) highlight the situation where hybrids are phenotyped in some environments but missing in others. Predictions in this scenario take into account information from correlated environments if GxE interaction is modeled, and if a relationship matrix is included, information from relatives evaluated in multiple environments. Predictions in the CV2 scenario, in an increasing level of missing environment, were made as follow: in the first case, one environment selected at random was considered as missing data and hybrids prediction were performed for this environment; then, two environments selected at random were considered as missing data and hybrids prediction were performed for these environments; and so on until 11 missing environments at random, the last level of CV2 scenario.

For both CV1 (in all environments) and CV2 (in each selected environment), ten times replicated five-fold cross-validation procedure was implemented to achieve the predictive accuracies, in which all single-cross hybrids with genotypes (147) were randomly split into five non-overlapping groups, being four of them training sets (80%) and one validation set (20%), considered as not phenotyped on each environment. Therefore, all results are based on 20% of missing hybrids. Permutation of these five groups led to five possible training and validation data sets. For all levels of missing environment, predictions were grouped and correlated with observed adjusted means.

3.3.6 Hybrids Rank

It was also computed the coincidence index of hybrids rank across and within environments for the top 20% hybrids and for the bottom 20% hybrids, produced by GEBV against observed performance (balanced data), for each model and level of missing environment. Additionally, linear regression coefficient of the observed performance on GEBV across environments was taken into account to analyse the efficiency of predictions relative to the levels of missing environment.

3.4 Results

3.4.1 Models Selection

The AIC criteria for models from first class ranged from 4704.20 (model $G_{FA(1)-I}$) to 4322.03 (model $G_{FA(3)-D}$) and for second class of models from 4727.79 (model $A_{FA(1)-I}$) to 4291.82 (model $A_{FA(4)-D}$). Within each class, the inclusion of diagonal structure for residuals effects reduced the values of both AIC and BIC criteria for the same k factor. The same pattern was observed when the additive genomic relationship matrix **A** was included for models 5 to 8 and 13 to 16, in which $\mathbf{R} = \mathbf{D}$. However, the inclusion of the **A** matrix increased the values of both criteria for models 1 to 4 and 9 to 12, in which $\mathbf{R} = \mathbf{I}$ (Table 1).

The percentage of genetic variance accounted for FA models (%vaf) ranged from 48.2% to 81.7% for the first class of models and from 57.4% to 88.5% for the second class of models. As expected, in both classes, as k became greater, the higher was the %vaf. Likewise for the AIC and BIC criteria, when $\mathbf{R} = \mathbf{D}$, the %vaf always increased for the same k factor. The inclusion of the **A** matrix also increased the %vaf, regardless residuals modeling. In the first class of models, the best AIC value was found for model 3, modeled with $\mathbf{G} = \mathbf{FA(3)}$ and $\mathbf{R} = \mathbf{D}$. The %vaf of this model was 76%, superior to the cut-off value of 75% adopted. Smith *et al.* (2015) used 80% as a cut-off value for %vaf, but their data-base comprised 200 cultivars evaluated across 196 trials. For the second class of models that included genomic information, model 11 with $k = 3$ also explained more than 75% of %vaf. The BIC criteria for both classes of models always selected models with $k = 1$ and $\mathbf{R} = \mathbf{D}$, being not helpful to select factor analytic models. Therefore, models 3 and 7 from first class of models and models 11 and 15 from second class of models, all with $k = 3$ and varying identity and diagonal structures for residuals, were selected to go forward with genomic selection (Table 1).

For the third class of models, the best AIC value was found for model A_{D-D} (20) and the best BIC value for model A_{I-D} (18). Regardless the best-fit model for this class, all second class models were used in genomic selection to analyze the influence of modeling genetics and residuals effects in the predictive accuracy.

3.4.2 Estimates of Genetic Parameters

Genetic variances were significantly greater than zero ($\sigma_g^2 > 0$) for most of the trials within environments based on the likelihood ratio test (LRT) with $\alpha = 0.05$, with the exception of T1 within environment 5 and T3 within environments 1 and 4. For the later ones, the coefficients of variation (CV %) were greater than 13%. The generalized measure of heritability ranged from 0.38 to 0.90, being zero for trial T3 within environment 1 where the genetic variance component was estimated as zero. Single-environment trial analysis also revealed that, at the same location, environments which fields were sown in the first crop season were more productive than environments sown in the second crop seasons (Table 2).

Additive and genetic correlations varied considerably between pairs of environments for models that comprise GxE interaction using a FA structure. Models $A_{FA(3)-D}$ and $A_{FA(3)-I}$, that included the genomic additive relationship matrix, presented in general higher correlations than models $I_{FA(3)-D}$ and $I_{FA(3)-I}$, in which hybrids were considered non genetically related

to each other. For example, the lowest value of pairwise correlation among environments found for models $A_{FA(3)-D}$ and $A_{FA(3)-I}$ was 0.21, and for models $I_{FA(3)-D}$ and $I_{FA(3)-I}$ was 0.06 and 0.08, respectively. Residuals modeling changed the magnitude of correlations, being slightly higher for models $A_{FA(3)-I}$ and $I_{FA(3)-I}$, in which $\Sigma_R = (\mathbf{I} \otimes \mathbf{I})$. The correlations induced by modeling residuals effects were smaller than those induced by modeling genetic effects. Overall, the estimated additive and genetic correlations between environments were reasonably high, with an average pairwise correlation of 0.58, 0.61, 0.47 and 0.51 for models $A_{FA(3)-D}$, $A_{FA(3)-I}$, $I_{FA(3)-D}$ and $I_{FA(3)-I}$ respectively. Correlations lower than 0.37 were found for 9, 6, 19 and 14 pairs of environments for these models, respectively. Based on the average of correlation of one environment to the others for each model, in general, environments 6, 7 and 11 had the lowest values of correlation and environments 5 and 9 the greatest ones (Figure 1).

Principal component analysis (PCA) showed good heterotic group consistency of the 143 inbred lines used as parents of the single-cross hybrids (Figure 2). Using SNP markers information, the genotypes of the single-cross hybrids were inferred and, due the good consistency of inbred lines heterotic groups, most hybrids were not close related (Figure 3). The four lines used as testers produced 48, 38, 23 and 20 single-cross hybrids. Within each of these groups, hybrids are half-sibs and their expected relatedness coefficient is 0.25 (Lynch and Walsh, 1998). From genomic relationship matrix, on average, these coefficients were 0.27, 0.29, 0.36 and 0.30, respectively.

3.4.3 Predictive Accuracy

When hybrids from validation set were considered as not phenotyped in all environments (CV1), all models that included genomic information presented similar results. Models $A_{FA(3)-D}$, A_{D-D}, and A_{I-D}, in which $\mathbf{R} = \mathbf{D}$, had predictive accuracies of 0.273, 0.262 and 0.274, respectively. Models $A_{FA(3)-I}$, A_{D-I}, and A_{I-I}, in which $\mathbf{R} = \mathbf{I}$, had predictive accuracies of 0.261, 0.232 and 0.264, respectively (Figure 4). Modelling the residuals with diagonal structure performed slightly better. Models from the first class were not able to predict in the CV1 scenario, since these models does not borrow information from relatives within environments through the additive genomic relationship matrix $\mathbf{A}$.

The inclusion of GxE interaction in the CV2 scenario with FA structure, until the level of five missing environments at random, almost double the predictive accuracy independently of residuals modeling. Taking model A_{D-D} from third class as a reference, models from first class had predictive accuracies on average of 70.30%, 62.92%, 62.38%, 61.05%, 52.11%, 58.66%, 40.77%, 36.35%, 37.78%, 12.63% and -19.67% superior/inferior over the reference model, from one to 11 missing environments at random, respectively. The only exception was for the level of 11 missing environments at random, in which the reference model performed better. For models from second class, the predictive accuracies were 71.38%, 70.46%, 68.49%, 70.87%, 53.11%, 71.07%, 57.15%, 50.89%, 56.88%, 34.13% and 19.53% superior over the reference model, respectively. Therefore, for first class models that did not account genomic information, borrowing information from correlated environments increased the predictive accuracy up to 50% over the reference model until six missing environments at random. Six environments represents a reduction of 50% of data. For the second class models that accounted GxE interaction and

genomic information, predictions were up to 50% over the reference model until nine missing environments at random, representing 75% of the data (Figure 4 and Tables A.3, A.4, A.5 and A.6).

Models from first and second classes had similar performance from one to five missing environments at random, and rising the number of missing environments, models from second class that explored genomic information had better performance. From six to 11 missing environments at random, models from second class performed on average 7.83%, 11.63%, 10.66%, 13.87%, 19.01% and 48.8% better than models from first class. Models $A_{FA(3)-D}$ and $A_{FA(3)-I}$ were able to keep predictions up to 0.400 until eight missing environments at random, and models $I_{FA(3)-D}$ and $I_{FA(3)-I}$, until five environments at random, highlighting the influence of missing environments for models that do not accounted genomic information. In general, as the number of missing environments became larger, the predictive accuracies got smaller for FA models, mainly for first class of models. Overall, the predictive accuracy of models $A_{FA(3)-D}$, $A_{FA(3)-I}$, $I_{FA(3)-D}$ and $I_{FA(3)-I}$, across all levels of missing environment, ranged from 0.444 to 0.325, 0.462 to 0.305, 0.449 to 0.200 and 0.460 to 0.220, respectively (Figure 4 and Tables A.3, A.4, A.5 and A.6).

The third class of models, not modeled with FA structure but with the genomic relationship matrix **A**, had similar predictive accuracies across all levels of missing environments, including the CV1 scenario. For these models, heterogeneous residuals variances performed slightly better in all levels of missing environment. For example, in a pairwise comparison between models A_{D-D} and A_{D-I}, the first model had on average an advantage of 14.3% in the predictive accuracy. For models A_{I-D} and A_{I-I}, this advantage decreased to 3.6%. For this class, models A_{D-D} and A_{I-D} had the lowest values of AIC and BIC, respectively, and in terms of prediction, no differences were found. Overall, models A_{I-I}, A_{I-D}, A_{D-I} and A_{D-D} had predictive accuracies ranging from 0.284 to 0.240, 0.280 to 0.258, 0.253 to 0.211 and 0.276 to 0.242, respectively (Figure 4 and Tables A.3, A.4, A.5 and A.6).

The efficiency of predictions relative to the levels of missing environments, based on the linear regression coefficient of the observed performance of hybrids on GEBV across environments, reflects the gradual reduction of predictive accuracy relative to the increase of missing environments for models that explore correlations between environments (models $I_{FA(3)-D}$, $I_{FA(3)-I}$, $A_{FA(3)-D}$ and $A_{FA(3)-I}$) (Figure A.9). For models $A_{FA(3)-D}$ and $A_{FA(3)-I}$, in the CV1 scenario, the coefficients of determination were similar to the results of models from third class, not modeled with FA structure. For the third class of models, the coefficients of determination were almost the same regardless the level of missing environments (Figure A.10).

The predictive accuracy of models that included the GxE interaction via FA structure were superior within all environments. This superiority was less evident for environments 6, 7 and 11, that had the lowest values of average correlation among themselves and the others environments. On the other hand, for environments 5 and 9, which had the greatest values of average correlation, the difference of predictive accuracy between models with and without FA structure was more evident. Although environments 3, 4, 8, 10 and 12 not presented elevated levels of correlations, the superiority of the models that included GxE interaction was also evident (Figure 5).

3.4.4 Changes in Ranking

For selection across environments, models $I_{FA(3)-D}$ and $I_{FA(3)-I}$ had the best values of coincidence index of all models and levels of CV2. For these models, until 10 missing environments at random, with one exception, all values were higher than 80%. Models $A_{FA(3)-D}$ and $A_{FA(3)-I}$, likewise with the exception of one value, presented values of coincidence higher than 80% until eight missing environments at random. In the more extreme scenario (CV1), predictions of these models were just based on the additive matrix **A**, and the coincidence index ranged from 37% to 47% for both top and bottom hybrids, respectively. Models from third class, that ignored correlation between environments, presented on average 20% and 16% of coincidence of the top and bottom 20% hybrids, respectively, across all levels of missing environment (Figure 6).

Models that included GxE interaction also showed advantage over models that ignored it for selection within environments. When one environment was missing at random, the first level of CV2, models $A_{FA(3)-D}$ and $A_{FA(3)-I}$ had coincidence index of 64% and 58% for the top hybrids, and of 60% and 54% for the bottom ones, respectively. For these models, values of coincidence were above 50% until the level of seven missing environments at random. Models $I_{FA(3)-D}$ and $I_{FA(3)-I}$, from first class, performed slightly better than models $A_{FA(3)-D}$ and $A_{FA(3)-I}$. For these models, values of coincidence index were above 50% until the missing of nine environments at random. Additive models from third class that ignored correlation between environments did not improve their performance, being their results similar to the coincidence index across environments (Figure 7).

3.5 Discussion

The performance of maize single-cross hybrids or any crop is directly affected by the environment. The relative performance and rank of genotypes may vary according to the environment, and from a plant breeding view, this is known as genotype by environment (GxE) interaction. The GxE interaction occurs due to differential expression of genes across environments (Zhang *et al.*, 2015; Ferrão *et al.*, 2017), which means that environmental conditions or even the level of technology for field production could change phenotypic performance through different patterns of gene-expression profiling. In practice, plant breeders have the options of utilize, avoid, or ignore it (Eisemann *et al.*, 1990; Yan, 2016).

In a first approach, our goal here is to evaluate if the inclusion of GxE interaction in a phenotypic model can boost predictions of untested tropical maize single-cross hybrids, in an intermediate stage of hybrids evaluation, based on 20% of missing hybrids under an increasing level of missing environments. In this case, the inclusion of factor analytic structure for genetic effects allowed that predictions taken into account information from phenotypic records of these hybrids in correlated environments. In a second moment and for the same models, we investigated if the inclusion of the genomic relationship matrix could improve predictions due that genomic estimated breeding values were estimated through information between hybrids within the missing environment(s), between hybrids across environments and among correlated environments.

The experiments were conduced at different locations or conditions (crop-seasons or field management), therefore we also investigated the influence of residuals modeling in the predictive accuracy of all models. Two variance-covariance (VCOV) structures were tested: the identity and diagonal matrices. The former structure determined that all environments shared the same variance component, and the later that each environment had an unique variance component. Mixed models are flexible in terms of including different assumptions of genetics ($\mathbf{G}$) and residuals ($\mathbf{R}$) modeling, allowing suitable VCOV structures that best-fit an specific biological situation (Van Eeuwijk *et al.*, 2016).

Three statistical criteria were used to select the best-fit FA model and therefore its k order to go forward with genomic selection. Recall that three classes of models were evaluated: (1) FA models do not incorporating the genomic relationship matrix, (2) FA models with genomic relationship information, and (3) models assuming correlation absence across environments of the GxE effects but including genomic relationship information. In a contrast of models with and without FA structure, account GxE interaction resulted in a better goodness of fit of models, highlighting the importance to take into account information from related environments. Zhang *et al.* (2015) and Cuevas *et al.* (2016) pointed out the actual practice of applying univariate or single-environment GBLUP models, ignoring correlation across environments and causing loss of information. Improvements in the goodness of fit were also observed when the genomic relationship matrix was included for FA models at the same k factor, for models that considered heterogeneous variances of residuals. Under a biological reasoning it is expected some correlation between the yield performance of hybrids across environments in a breeding program rather than homogeneous variances and the absence of correlations. On the other hand, models under a FA structure in $\mathbf{G}$ and with homogeneous structure in $\mathbf{R}$ showed an increasing value of both AIC

and BIC when incorporating the **A** relationship matrix. This results indicates that modeling the **G** under an FA structure, and the parsimonious strategy of unique residual variance components in **R**, tends to result in a better representation of the full sample of unobserved set of hybrids.

Considering heterogeneous variance components for residuals resulted in improvements in the goodness of fit for all models. One possible explanation is that each environment present its own source of variation that can not be explained by the model, due to climate conditions, plant diseases, or any other source not considered by the model. The BIC always penalized FA models of high orders. FA models are nested (FA1, FA2, FA3, and so on) (Smith *et al.*, 2015), and as explained by Sorensen and Gianola (2002), BIC is well defined for non-nested models. The percentage of genetic variance accounted by the k factors, as expected, increased as k became greater. Smith *et al.* (2015) observed the same pattern of BIC to select FA models. Therefore, in the GS context, more than one criteria should be used to select the best fit model (Ferrão *et al.*, 2017).

The predictive accuracies and coincidence for hybrids ranking of models that borrowed information from correlated environments, clearly showed the importance of including GxE interaction into GBLUP models, even in high levels of missing environments. The matrices of genetic and additive correlation across environments for models with and without genomic information, respectively, confirmed the high association between environments. Genetic effects are expected to vary according to the environment (Zhang *et al.*, 2015; Ferrão *et al.*, 2017) due to the GxE interaction. Hence, models that allow different variance components across environments are more realistic and capable of capturing these patterns. Residuals modeling with heterogeneous variances for models under FA in **G** did not improve predictive accuracies over models whith homogeneous variances for residuals. Similar results were found by Burgueño *et al.* (2012), where genetic effects were more important than residuals effects.

The gains in prediction accuracy obtained in CV2 over CV1 scenario, with models that account GxE interaction, were directly related to the ability of these models of borrowing information from correlated environments. Hence, the magnitude of correlations is an important parameter to be considered. Similar results were found by Burgueño *et al.* (2012), Crossa *et al.* (2014) and Lopez-Cruz *et al.* (2015). Therefore, predictions of newly lines, a situation statistically created by the CV1 scenario, were more challenging than predicting single-cross hybrids evaluated in some environments but missing in others (CV2).

Several studies included FA structure to account the GxE interaction (Kelly *et al.*, 2007; Burgueño *et al.*, 2011; Cullis *et al.*, 2014; Smith *et al.*, 2015; Dias *et al.*, 2018b), and specifically for genomic selection, it have been used for wheat (Burgueño *et al.*, 2012; Dawson *et al.*, 2013; Rutkoski *et al.*, 2015), barley (Malosetti *et al.*, 2016; Oakey *et al.*, 2016) and maize (Schulz-Streeck *et al.*, 2013; Dias *et al.*, 2018a). Our results have shown a great flexibility of FA structure to handle with low, moderate and high levels of missing data in the framework of MET, as also pointed out by Elias *et al.* (2016) in a revision about GxE interaction in plant breeding experiments. Phenotyping in multi-environment trials is routine in plant breeding programs, and although the FA structure is an approximation to the unstructured VCOV matrix, it provides reliable information to acess the performance of single-cross maize hybrids within environments (Smith *et al.*, 2001; Kelly *et al.*, 2007).

A better understand of GxE interaction is important to any breeding program. Accurate information of GxE interaction can be used as a guideline to (*i*) define breeding strategies looking at a specific future market share or for a regular breeding zone, (*ii*) evaluate stability of genotypes for a specific environment/condition or for a mega-environment, (*iii*) maximize genetic gain, among others (Gezan *et al.*, 2016; Dias *et al.*, 2018a). Another advantage of using FA models and genomic information is that latent regression plots can be obtained to analyse the stability of breeding values of hybrids across environments (Smith *et al.*, 2015; Dias *et al.*, 2018a) and to select the inbred lines parents of hybrids that presented great stability of additive effects to generate synthetic populations for the next breeding cycle (Dias *et al.*, 2018a).

Cost reduction and improved selection are examples of how GS can reshape breeding programs (Hickey *et al.*, 2017), but its application depends on the ability of models to predict real situations faced in the breeding programs (Ferrão *et al.*, 2017). Our results emphasize that even in high levels of missing data, models that account correlation between environments and genomic information can be a valuable tool to predict breeding values. Just as an example, considering the cost of GBS genotyping in a sequencing coverage (x) of 2x as US$ 25,00 (Gorjanc *et al.*, 2017) per line and that the cost of one maize yield-trial plot as US$ 13,00 (Tech Services Inc., 2018). Then, the budget needed for a breeding program similar to the data presented in this study would be: US$ 3,575 for genotyping the 143 inbred lines used as parents, and US$ 45,864 for phenotyping the 147 single-cross hybrids considered for genomic selection at 12 environments. Using a genomic selection model that embrace GxE interaction plus genomic information, it was shown that until eight missing environments at random or 66% of missing data, predictions of untested single-cross hybrids were up 0.400 with an average coincidence index of at least of 80% and of 50% for selections across and within environments, respectively. Hence, a reduction of breeding costs by 8.33% or US$ 3,822 can be achieved if hybrids were predict in one environment. This amount is sufficient to cover the costs of inbred lines genotyping (US$ 3,575). For the following levels of missing environment, the amount of costs reduction is linear; if hybrids were predicted in two environments, the reduction would be by 16.67% or US$ 7,644; for three environments by 25,00%, and so on, until a reduction by 66,67% or US$ 30,576 for prediction at eight environments.

Regardless the level of missing data for genomic prediction, any reduction of the total budged of hybrids phenotyping could be allocated to optimize the breeding program. An interesting way to allocate the saved budget is the production of newly synthetic populations for inbred lines extractions, which is the source of cultivars which meet specific breeding objectives (Bernardo, 2010, p.15). As example, the cost for producing a newly synthetic population obtained from 10 inbred lines - including the cost of labor, time demanded and nursery space for crosses - is on average US$ 1,200 (Dr. David Benson - Global Consultant, CEO and founder of Cornhusker Hybrids - personal communication, February 21st, 2018). Then, if hybrids were predicted in one environment, the saved budget could be used to produce three newly synthetic population or to cover the costs of inbred lines genotyping, as mentioned above. Other possibilities to allocate the saved budget is the evaluation of more hybrids at the intermediate stage and therefore increase the intensity of selection, to obtain genotypic data for newly inbred lines and hence predict the performance of newly developed single-cross hybrids, or even to reduce

costs. It was also noted by Krchov and Bernardo (2015) that once genomic selection is implemented in the breeding process, the reduction in the amount of phenotyping leads to a better quality of the field data, enhancing the effectiveness of selection.

So far, all results of FA models that included genomic information were based on the additive relationship matrix. However, as maize is an allogamous species, it is also worthwhile to investigate the inclusion of dominance effects into the models and therefore make hybrids predictions with a GBLUP model that account additive plus dominance effects. Results of this approach showed no improvements in hybrids predictions neither in hybrids ranking (data not shown), although some exciting results have been reported in the literature (Dos Santos *et al.*, 2016; Dias *et al.*, 2018a). Due to small dominance effect relationships between hybrids, the dominance relationship matrix was less informative than the additive relationship matrix and hence did not improve prediction accuracy. Similar results were found by Ertl *et al.* (2014). Another point could be that an increase in the population size could have better estimated the dominance effects.

Finally, we obtained encouraging accuracies of tropical maize single-cross hybrids for genomic selection implementation by accounting genomic additive relationship information and the effects of genetic heterogeneity and genotype by environment interaction. Our methodology can also be expanded to other crops in which MET plays an important role in the breeding process. Future research in the integration of optimized experimental designs and crop growth models (Heslot *et al.*, 2015; Rincent *et al.*, 2017), that combine ecophysiological and genetics modeling, seems to be a promising way for genomic selection predictions.

4 CONCLUSIONS

(*i*) The inclusion of factor analytic structure boosted the predictive accuracy of untested maize single-cross hybrids, regardless residuals modeling;

(*ii*) Models that included genomic relationship information and GxE interaction by factor analytic structure achieved higher predictive accuracy in elevated levels of missing environments; and

(*iii*) High levels of predictive accuracy of untested maize single-cross hybrids were found with moderated to low levels of missing environments.

References

Acosta-Pech, R., J. Crossa, G. de los Campos, S. Teyssèdre, B. Claustres, S. Pérez-Elizalde, and P. Pérez-Rodríguez, 2017 Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theoretical and Applied Genetics **130**: 1431–1440.

Akaike, H., 1974 A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control **19**: 716–723.

Amadeu, R. R., C. Cellon, J. W. Olmstead, A. A. F. Garcia, M. F. R. Resende, and P. R. Muñoz, 2016 AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. The Plant Genome **9**: 1–10.

Bernardo, R., 1994 Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Science1 **34**: 20–25.

Bernardo, R., 1995 Genetic models for predicting maize single-cross performance in unbalanced yield trial data. Crop Science **36**: 50–56.

Bernardo, R., 1996a Best linear unbiased prediction of the performance of crosses between untested maize inbreds. Crop Science **36**: 872–876.

Bernardo, R., 1996b Testcross additive and dominance effects in best linear unbiased prediction of maize single-cross performance. Theoretical and Applied Genetics **93**: 1098–1102.

Bernardo, R., 2008 Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Science **48**: 1649–1664.

Bernardo, R., 2010 *Breeding for Quantitative Traits in Plants*. Stemma Press, Woodbury, Minnesota, second edition.

Browning, B. L. and S. R. Browning, 2016 Genotype Imputation with Millions of Reference Samples. American Journal of Human Genetics **98**: 116–126.

Burgueño, J., J. Crossa, J. M. Cotes, F. S. Vicente, and B. Das, 2011 Prediction assessment of linear mixed models for multienvironment trials. Crop Science **51**: 944–954.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa, 2012 Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Science **52**: 707–719.

Butler, D., B. R. Cullis, and R. Gilmour, 2009 ASReml-R reference manual.

Crossa, J., G. De Los Campos, M. Maccaferri, R. Tuberosa, J. Burgueño, and P. Pérez-Rodríguez, 2016 Extending the marker x Environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Science **56**: 2193–2209.

Crossa, J., P. Pérez, J. Hickey, J. Burgueño, L. Ornella, J. Cerón-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, D. Bonnett, and K. Mathews, 2014 Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity **112**: 48–60.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. de los Campos, J. Burgueño, J. M. González-Camacho, S. Pérez-Elizalde, Y. Beyene, S. Dreisigacker, R. Singh, X. Zhang, M. Gowda, M. Roorkiwal, J. Rutkoski, and R. K. Varshney, 2017 Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science **22**: 961–975.

Cuevas, J., J. Crossa, V. Soberanis, S. Pérez-Elizalde, P. Pérez-Rodríguez, G. de los Campos, O. A. Montesinos-López, and J. Burgueño, 2016 Genomic Prediction of Genotype × Environment Interaction Kernel Regression Models. The Plant Genome **9**: 1–20.

Cullis, B. R., P. Jefferson, R. Thompson, and A. B. Smith, 2014 Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a Pinus radiata breeding programme. Theoretical and Applied Genetics **127**: 2193–2210.

Cullis, B. R., A. B. Smith, and N. E. Coombes, 2006 On the design of early generation variety trials with correlated data. Journal of Agricultural, Biological, and Environmental Statistics **11**: 381–393.

Dawson, J. C., J. B. Endelman, N. Heslot, J. Crossa, J. Poland, S. Dreisigacker, Y. Manès, M. E. Sorrells, and J. L. Jannink, 2013 The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crops Research **154**: 12–22.

Dias, K. O. D. G., S. A. Gezan, C. T. Guimarães, A. Nazarian, L. da Costa e Silva, S. N. Parentoni, P. E. de Oliveira Guimarães, C. de Oliveira Anoni, J. M. V. Pádua, M. de Oliveira Pinto, R. W. Noda, C. A. G. Ribeiro, J. V. de Magalhães, A. A. F. Garcia, J. C. de Souza, L. J. M. Guimarães, and M. M. Pastina, 2018a Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity .

Dias, K. O. D. G., S. A. Gezan, C. T. Guimarães, S. N. Parentoni, P. E. d. O. Guimarães, N. P. Carneiro, A. F. Portugal, E. A. Bastos, M. J. Cardoso, C. d. O. Anoni, J. V. de Magalhães, J. C. de Souza, L. J. M. Guimarães, and M. M. Pastina, 2018b Estimating Genotype × Environment Interaction for and Genetic Correlations among Drought Tolerance Traits in Maize via Factor Analytic Multiplicative Mixed Models. Crop Science **58**: 72.

Dos Santos, J. P. R., R. C. De Castro Vasconcellos, L. P. M. Pires, M. Balestre, and R. G. Von Pinho, 2016 Inclusion of dominance effects in the multivariate GBLUP model. PLoS ONE **11**: 1–21.

Eisemann, R., M. Cooper, and D. Woodruff, 1990 Beyond the analytical methodology - Better interpretation and exploitation of genotype-by-environment interaction in breeding. In *Genotype-by-environment interaction and plant breeding*, edited by M. Kang, pp. 108–117, Louisiana State University, Baton Rouge.

Elias, A. A., K. R. Robbins, R. W. Doerge, and M. R. Tuinstra, 2016 Half a century of studying genotype × Environment interactions in plant breeding experiments. Crop Science **56**: 2090–2105.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE **6**: 1–10.

Ertl, J., A. Legarra, Z. G. Vitezica, L. Varona, C. Edel, R. Emmerling, and K. U. Götz, 2014 Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genetics Selection Evolution **46**: 1–10.

Federizzi, L. C., S. A. M. Carbonell, M. T. Pacheco, and I. C. Nava, 2012 Breeders ' work after cultivar development - the stage of recommendation. Crop Breeding and Applied Biotechnology **S2**: 67–74.

Ferrão, L. F. V., R. G. Ferrão, M. A. G. Ferrão, A. Francisco, and A. A. F. Garcia, 2017 A mixed model to multiple harvest-location trials applied to genomic prediction in Coffea canephora. Tree Genetics and Genomes **13**: 95.

Fisher, R. A., 1918 The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the RoyaZ Society of Edinburgh **52**: 399–433.

Fritsche-Neto, R., M. C. Gonçalves, R. Vencovsky, and C. L. D. Souza Junior, 2010 Prediction of genotypic values of maize hybrids in unbalanced experiments. Crop Breeding and Applied Biotechnology **10**: 32–39.

Gezan, S. A., M. P. de Carvalho, and J. Sherrill, 2016 Statistical methods to explore genotype-by-environment interaction for loblolly pine clonal trials. Tree Genetics and Genomes **13**: 1.

Gilmour, A. R., R. Thompson, and B. R. Cullis, 1995 Average Infomation REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. Biometrics **51**: 1440–1450.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire, Q. Sun, and E. S. Buckler, 2014 TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE **9**.

Gorjanc, G., J.-F. Dumasy, S. Gonen, R. C. Gaynor, R. Antolin, and J. M. Hickey, 2017 Potential of Low-Coverage Genotyping-by-Sequencing and Imputation for Cost-Effective Genomic Selection in Biparental Segregating Populations. Crop Science **57**: 1–17.

Heffner, E. L., A. J. Lorenz, J. L. Jannink, and M. E. Sorrells, 2010 Plant breeding with Genomic selection: Gain per unit time and cost. Crop Science **50**: 1681–1690.

Henderson, C. R., 1950 Estimation of Genetic Parameters. Annals of Mathematical Statistics **21**: 309–310.

Heslot, N., J.-L. Jannink, and M. E. Sorrells, 2015 Perspectives for Genomic Selection Applications and Research in Plants. Crop Science **55**: 1–12.

Hickey, J. M., T. Chiurugwi, I. Mackay, and W. Powell, 2017 Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nature Genetics **49**: 1297–1303.

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt, J. Lorgeou, F. Piraux, L. Guerreiro, P. Pérez, M. Calus, J. Burgueño, and G. de los Campos, 2014 A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics **127**: 595–607.

Jonas, E. and D. J. de Koning, 2016 Goals and hurdles for a successful implementation of genomic selection in breeding programme for selected annual and perennial crops. Biotechnology and Genetic Engineering Reviews **32**: 18–42.

Kelly, A. M., A. B. Smith, J. A. Eccleston, and B. R. Cullis, 2007 The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Science **47**: 1063–1070.

Krchov, L.-M. and R. Bernardo, 2015 Relative Efficiency of Genomewide Selection for Testcross Performance of Doubled Haploid Lines in a Maize Breeding Program. Crop Science **55**: 2091–2099.

Lopez-Cruz, M., J. Crossa, D. Bonnett, S. Dreisigacker, J. Poland, J.-L. Jannink, R. P. Singh, E. Autrique, and G. de los Campos, 2015 Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker × Environment Interaction Genomic Selection Model. G3: Genes, Genomes, Genetics **5**: 569–582.

Lynch, M. and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, first edition.

Malosetti, M., D. Bustos-Korts, M. P. Boer, and F. A. Van Eeuwijk, 2016 Predicting responses in multiple environments: Issues in relation to genotype x Environment interactions. Crop Science **56**: 2210–2222.

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.

Oakey, H., B. Cullis, R. Thompson, J. Comadran, C. Halpin, and R. Waugh, 2016 Genomic Selection in Multi-environment Crop Trials. G3: Genes, Genomes, Genetics **6**: 1313–1326.

Patterson, H. D. and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika **58**: 545–554.

Piepho, H.-P., 1997 Analyzing Genotype-Environment Data by Mixed Models with Multiplicative Terms. Biometrics **53**: 761–766.

Piepho, H. P., 1998 Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theoretical and Applied Genetics **97**: 195–201.

Piepho, H. P., J. Möhring, A. E. Melchinger, and A. Büchse, 2008 BLUP for phenotypic selection in plant breeding and variety testing. Euphytica **161**: 209–228.

Powell, J. E., P. M. Visscher, and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics **11**: 800–805.

R Core Team, 2017 R: A Language and Environment for Statistical Computing.

Rincent, R., E. Kuhn, H. Monod, F. X. Oury, M. Rousset, V. Allard, and J. Le Gouis, 2017 Optimization of multi-environment trials for genomic selection based on crop models. Theoretical and Applied Genetics **130**: 1735–1752.

Rutkoski, J., R. P. Singh, J. Huerta-Espino, S. Bhavani, J. Poland, J. L. Jannink, and M. E. Sorrells, 2015 Efficient Use of Historical Data for Genomic Selection: A Case Study of Stem Rust Resistance in Wheat. The Plant Genome **8**: 0.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C.-T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A.-P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J.-M. Chia, J.-M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang,

M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, and R. K. Wilson, 2009 The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science **326**: 1112–1115.

Schulz-Streeck, T., J. O. Ogutu, A. Gordillo, Z. Karaman, C. Knaak, and H. P. Piepho, 2013 Genomic selection allowing for marker-by-environment interaction. Plant Breeding **132**: 532–538.

Schwarz, G., 1978 Estimating the Dimension of a Model. The Annals of Statistics **6**: 461–464.

Smith, A., B. R. Cullis, and R. Thompson, 2001 Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics **57**: 1138–1147.

Smith, A. B., A. Ganesalingam, H. Kuchel, and B. R. Cullis, 2015 Factor analytic mixed models for the provision of grower information from national crop variety testing programs. Theoretical and Applied Genetics **128**: 55–72.

Sorensen, D. and D. Gianola, 2002 *Likelihood of Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York.

Tech Services Inc., 2018 Pricing Brochure TSI 2018 Test Sites.

Van Eeuwijk, F. A., D. V. Bustos-Korts, and M. Malosetti, 2016 What should students in plant breeding know about the statistical aspects of genotype x Environment interactions? Crop Science **56**: 2119–2140.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414–4423.

Wright, S., 1921 Systems of Mating. I. the Biometric Relations between Parent and Offspring. Genetics pp. 111–126.

Yan, W., 2016 Analysis and handling of G x E in a practical breeding program. Crop Science **56**: 2106–2118.

Zhang, X., P. Pérez-Rodríguez, K. Semagn, Y. Beyene, R. Babu, M. A. López-Cruz, F. San Vicente, M. Olsen, E. Buckler, J. L. Jannink, B. M. Prasanna, and J. Crossa, 2015 Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity **114**: 291–299.

Table 1: Goodness of fit for models divided into three classes, (1) FA models without incorporating the genomic relationship matrix - models 1 to 8, (2) FA models with genomic relationship information - models 9 to 16, and (3) models assuming no correlation across environments of the GxE effects but including genomic relationship information - models 17 to 20. **I**: identity matrix, **FA**(k): factor analytic matrix of order k, **D**: diagonal matrix, and **A**: additive relationship matrix from molecular markers.

Model		**Covariance structure**			**Selection criteria**		
Number	**Code**	$\boldsymbol{\Sigma_g}$	$\boldsymbol{\Sigma_r}$	**Nu. Par.**[a]	**AIC**	**BIC**	$\%\boldsymbol{var}^{\boldsymbol{b}}$
			BLUP				
1	$\mathrm{I}_{FA(1)-I}$	$\mathbf{I} \otimes \mathbf{FA}(1)$	$\mathbf{I} \otimes \mathbf{I}$	26	4704.08	4866.09	48.2
2	$\mathrm{I}_{FA(2)-I}$	$\mathbf{I} \otimes \mathbf{FA}(2)$	$\mathbf{I} \otimes \mathbf{I}$	37	4667.45	4879.30	61.7
3	$\mathrm{I}_{FA(3)-I}$	$\mathbf{I} \otimes \mathbf{FA}(3)$	$\mathbf{I} \otimes \mathbf{I}$	47	4663.74	4937.91	70.8
4	$\mathrm{I}_{FA(4)-I}$	$\mathbf{I} \otimes \mathbf{FA}(4)$	$\mathbf{I} \otimes \mathbf{I}$	56	4669.78	4981.34	80.4
5	$\mathrm{I}_{FA(1)-D}$	$\mathbf{I} \otimes \mathbf{FA}(1)$	$\mathbf{I} \otimes \mathbf{D}$	37	4376.63	4607.18	51.7
6	$\mathrm{I}_{FA(2)-D}$	$\mathbf{I} \otimes \mathbf{FA}(2)$	$\mathbf{I} \otimes \mathbf{D}$	48	4340.81	4639.91	64.3
7	$\mathrm{I}_{FA(3)-D}$	$\mathbf{I} \otimes \mathbf{FA}(3)$	$\mathbf{I} \otimes \mathbf{D}$	58	4322.03	4664.74	76.0
8	$\mathrm{I}_{FA(4)-D}$	$\mathbf{I} \otimes \mathbf{FA}(4)$	$\mathbf{I} \otimes \mathbf{D}$	67	4322.88	4709.21	81.7
			GBLUP				
9	$\mathrm{A}_{FA(1)-I}$	$\mathbf{A} \otimes \mathbf{FA}(1)$	$\mathbf{I} \otimes \mathbf{I}$	26	4727.79	4889.79	57.4
10	$\mathrm{A}_{FA(2)-I}$	$\mathbf{A} \otimes \mathbf{FA}(2)$	$\mathbf{I} \otimes \mathbf{I}$	37	4697.69	4903.31	70.0
11	$\mathrm{A}_{FA(3)-I}$	$\mathbf{A} \otimes \mathbf{FA}(3)$	$\mathbf{I} \otimes \mathbf{I}$	47	4698.65	4954.12	76.6
12	$\mathrm{A}_{FA(4)-I}$	$\mathbf{A} \otimes \mathbf{FA}(4)$	$\mathbf{I} \otimes \mathbf{I}$	56	4704.20	5009.52	81.6
13	$\mathrm{A}_{FA(1)-D}$	$\mathbf{A} \otimes \mathbf{FA}(1)$	$\mathbf{I} \otimes \mathbf{D}$	37	4331.08	4561.63	63.5
14	$\mathrm{A}_{FA(2)-D}$	$\mathbf{A} \otimes \mathbf{FA}(2)$	$\mathbf{I} \otimes \mathbf{D}$	48	4306.55	4593.18	74.0
15	$\mathrm{A}_{FA(3)-D}$	$\mathbf{A} \otimes \mathbf{FA}(3)$	$\mathbf{I} \otimes \mathbf{D}$	58	4302.04	4644.76	83.4
16	$\mathrm{A}_{FA(4)-D}$	$\mathbf{A} \otimes \mathbf{FA}(4)$	$\mathbf{I} \otimes \mathbf{D}$	67	4291.82	4671.91	88.5
17	A_{I-I}	$\mathbf{A} \otimes \mathbf{I}$	$\mathbf{I} \otimes \mathbf{I}$	3	5103.38	5122.07	-
18	A_{I-D}	$\mathbf{A} \otimes \mathbf{I}$	$\mathbf{I} \otimes \mathbf{D}$	14	4595.73	4682.96	-
19	A_{D-I}	$\mathbf{A} \otimes \mathbf{D}$	$\mathbf{I} \otimes \mathbf{I}$	14	4953.22	5040.45	-
20	A_{D-D}	$\mathbf{A} \otimes \mathbf{D}$	$\mathbf{I} \otimes \mathbf{D}$	25	4583.74	4739.52	-

[a] Number of parameters estimated for each model. [b] Percentage of genetic variance accounted for FA models.

Table 2: Results from single-environment trial analysis for average grain yield (GY, in t ha^{-1}), coefficient of variation (CV %), generalized measure of heritability ($\hat{H}^2$), genetic (σ^2_g), block (σ^2_b) and residuals (σ^2_ϵ) variance components.

Environments	**Env 1**			**Env 2**			**Env 3**			**Env 4**			**Env 5**			**Env 6**		
Trials	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
GY (t ha^{-1})	10.82	11.14	11.25	3.78	3.71	4.18	9.47	9.74	9.94	8.40	8.68	8.92	6.10	6.57	7.43	7.33	7.14	6.21
CV %	8.92	7.08	13.01	8.99	11.90	8.70	11.50	11.10	10.96	13.50	10.60	13.22	14.30	14.60	12.01	11.60	16.60	13.94
$\hat{H}^2$	0.64	0.82	0.00	0.81	0.78	0.72	0.57	0.72	0.64	0.44	0.75	0.35	0.32	0.38	0.46	0.72	0.51	0.61
σ^2_g	0.83	1.43	0.00^{ns}	0.24	0.34	0.19	0.86	1.53	1.05	0.53	1.32	0.41^{ns}	0.19^{ns}	0.30	0.34	1.05	0.78	0.68
σ^2_b	0.00^{ns}	0.00^{ns}	0.04^{ns}	0.00^{ns}	0.00^{ns}	0.02^{ns}	0.19	0.00^{ns}	0.00^{ns}	0.04^{ns}	0.02^{ns}	0.40	0.02^{ns}	0.12^{ns}	0.00^{ns}	0.30	0.27	0.33
σ^2_ϵ	0.93	0.62	2.14	0.12	0.20	0.13	1.19	1.17	1.19	1.29	0.85	1.39	0.76	0.93	0.80	0.72	1.41	0.75
Environments	**Env 7**			**Env 8**			**Env 9**			**Env 10**			**Env 11**			**Env 12**		
Trials	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
GY (t ha^{-1})	8.35	8.30	7.04	11.84	10.83	12.26	8.46	8.37	8.11	7.69	7.32	8.36	4.57	4.27	6.26	4.08	4.00	7.32
CV %	12.90	13.50	13.11	8.20	9.73	7.97	12.30	12.30	10.26	12.20	11.90	9.60	11.50	9.17	10.77	9.16	12.80	7.51
$\hat{H}^2$	0.78	0.76	0.75	0.53	0.51	0.53	0.72	0.75	0.74	0.67	0.86	0.90	0.46	0.61	0.66	0.72	0.62	0.73
σ^2_g	2.19	1.97	1.37	0.56	0.60	0.54	1.27	1.29	1.03	0.35	0.86	1.56	0.40	0.49	0.75	0.20	0.22	0.45
σ^2_b	0.15^{ns}	0.00^{ns}	0.12^{ns}	0.08^{ns}	0.09^{ns}	0.01^{ns}	0.16	0.14^{ns}	0.00^{ns}	0.06	0.07^{ns}	0.01^{ns}	0.00^{ns}	0.04^{ns}	0.01^{ns}	0.03	0.02^{ns}	0.09^{ns}
σ^2_ϵ	1.16	1.26	0.85	0.94	1.11	0.96	0.89	0.82	0.74	0.31	0.26	0.36	0.94	0.59	0.76	0.14	0.26	0.30

^{ns} Variance component statistically equal to zero (σ^2_b or $\sigma^2_g = 0$), based in the likelihood ratio test (LRT) with $\alpha = 0.05$.

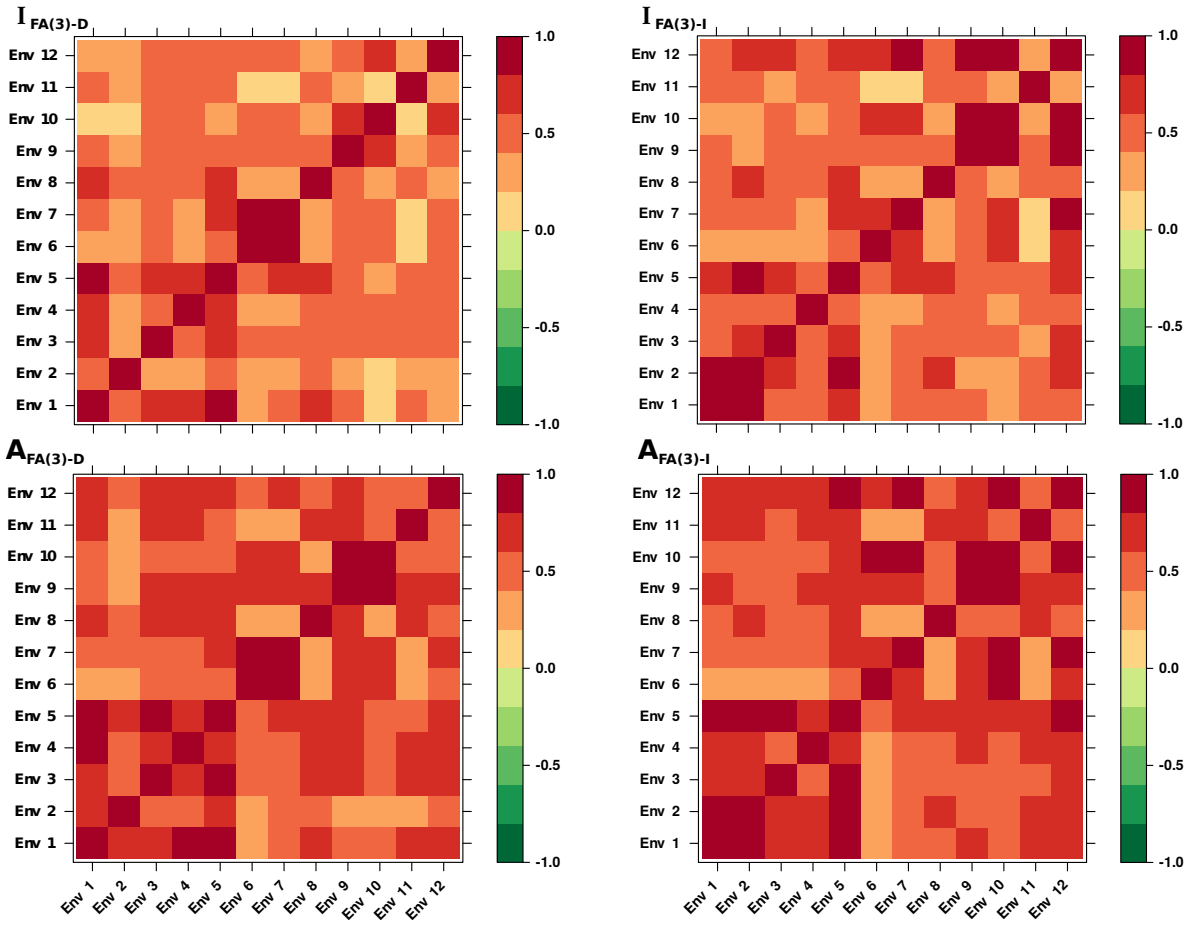

Figure 1: Heatmap of genetic and additive correlations between environments, for FA models without incorporating the genomic relationship matrix ($I_{FA(3)-D}$ and $I_{FA(3)-I}$), and for FA models with genomic relationship information ($A_{FA(3)-D}$ and $A_{FA(3)-I}$).

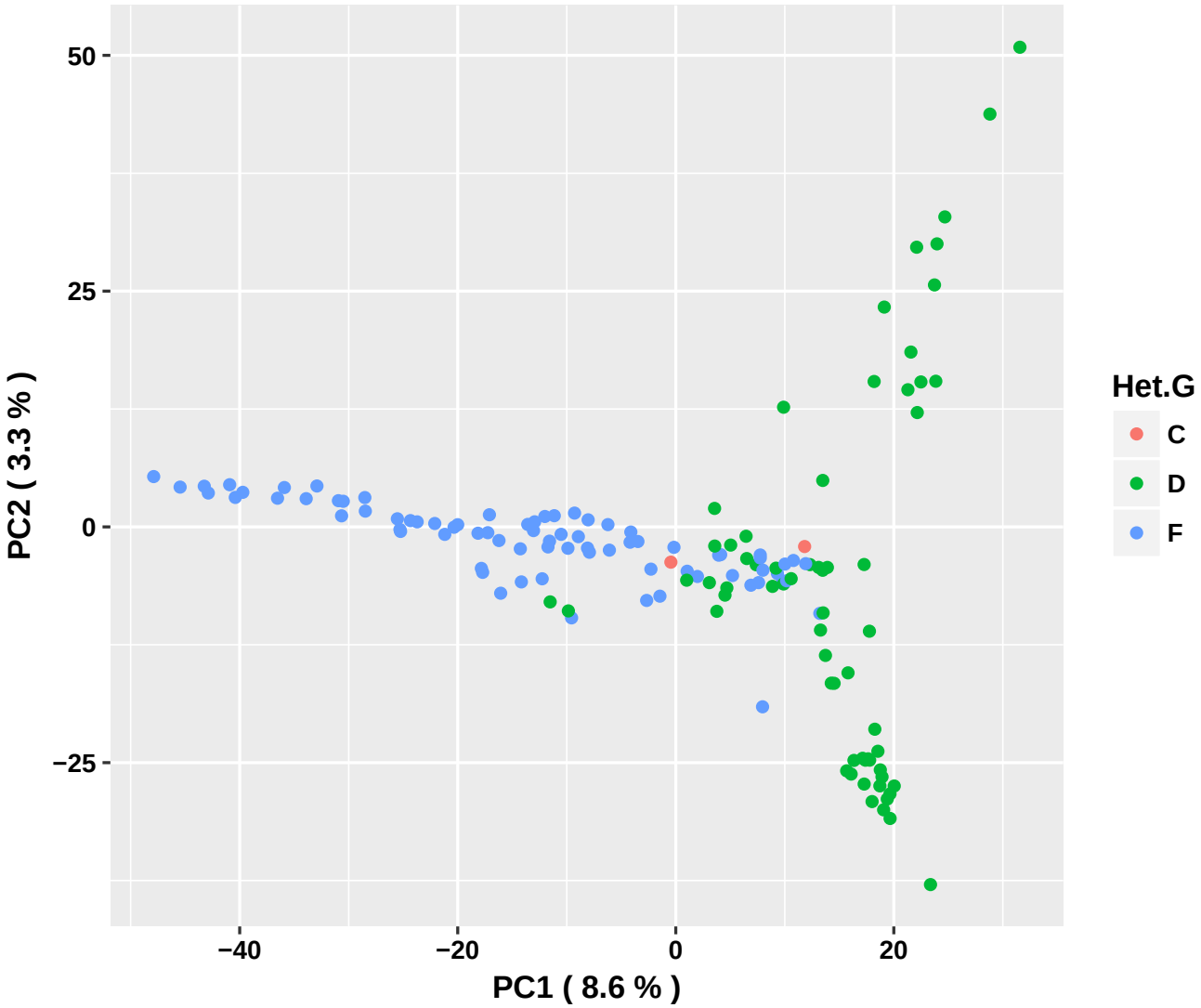

Figure 2: Plot of first two principal components from principal component analysis (PCA), based on 29,515 SNP markers for 143 inbreed lines used as parents of maize single-cross hybrids. In the right side, legend means heterotic groups C (group C), D (Dent) and F (Flint).

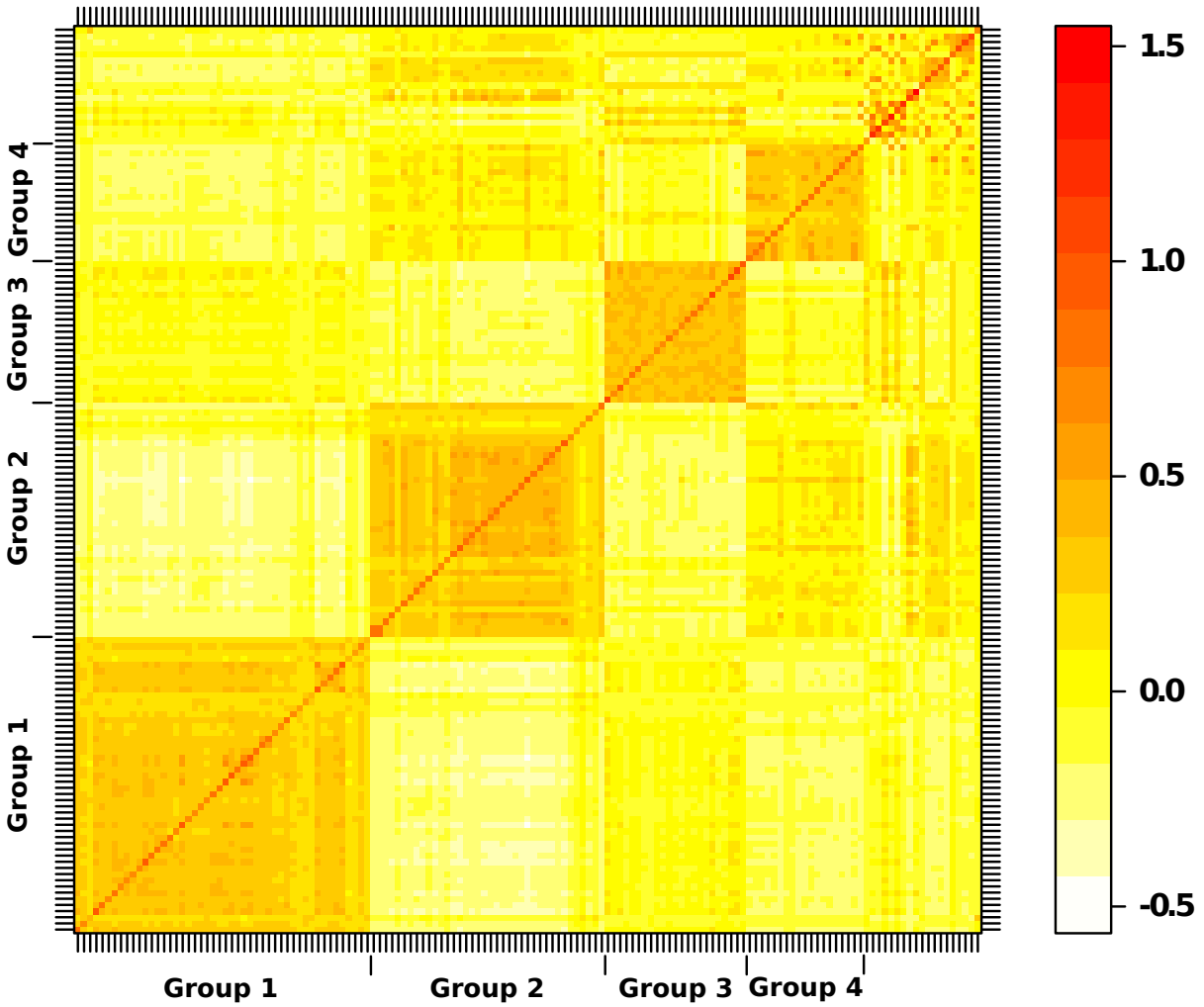

Figure 3: Heatmap of realized genomic relationship matrix **A** for 147 maize single-cross hybrids ordered by four groups of half-sibs, each group synthesized by the same tester. From group one to four, the size of the groups are 48, 38, 23 and 20 single-cross hybrids, with an average relatedness coefficient of 0.27, 0.29, 0.36 and 0.30, respectively. The remaining hybrids were synthesized by others testers.

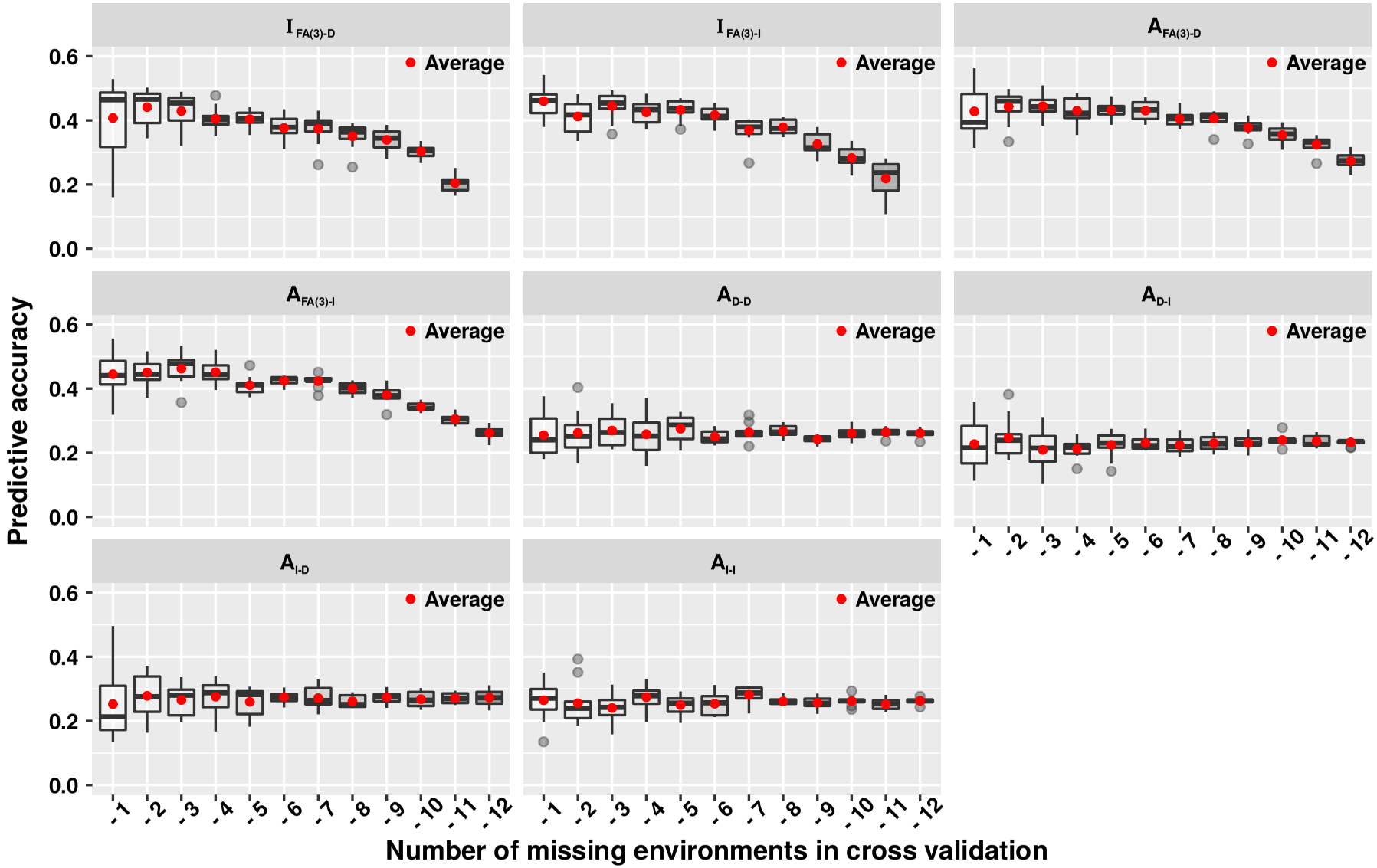

Figure 4: Average predictive accuracies across environments, for FA models without incorporating the genomic relationship matrix ($I_{FA(3)-D}$ and $I_{FA(3)-I}$), for FA models with genomic relationship information ($A_{FA(3)-D}$ and $A_{FA(3)-I}$), and for models assuming no correlation across environments of the GxE effects but including genomic relationship information (A_{D-D}, A_{D-I}, A_{I-D} and A_{I-I}). In the x axis, levels of missing environments ranging from one (-1) to all (-12).

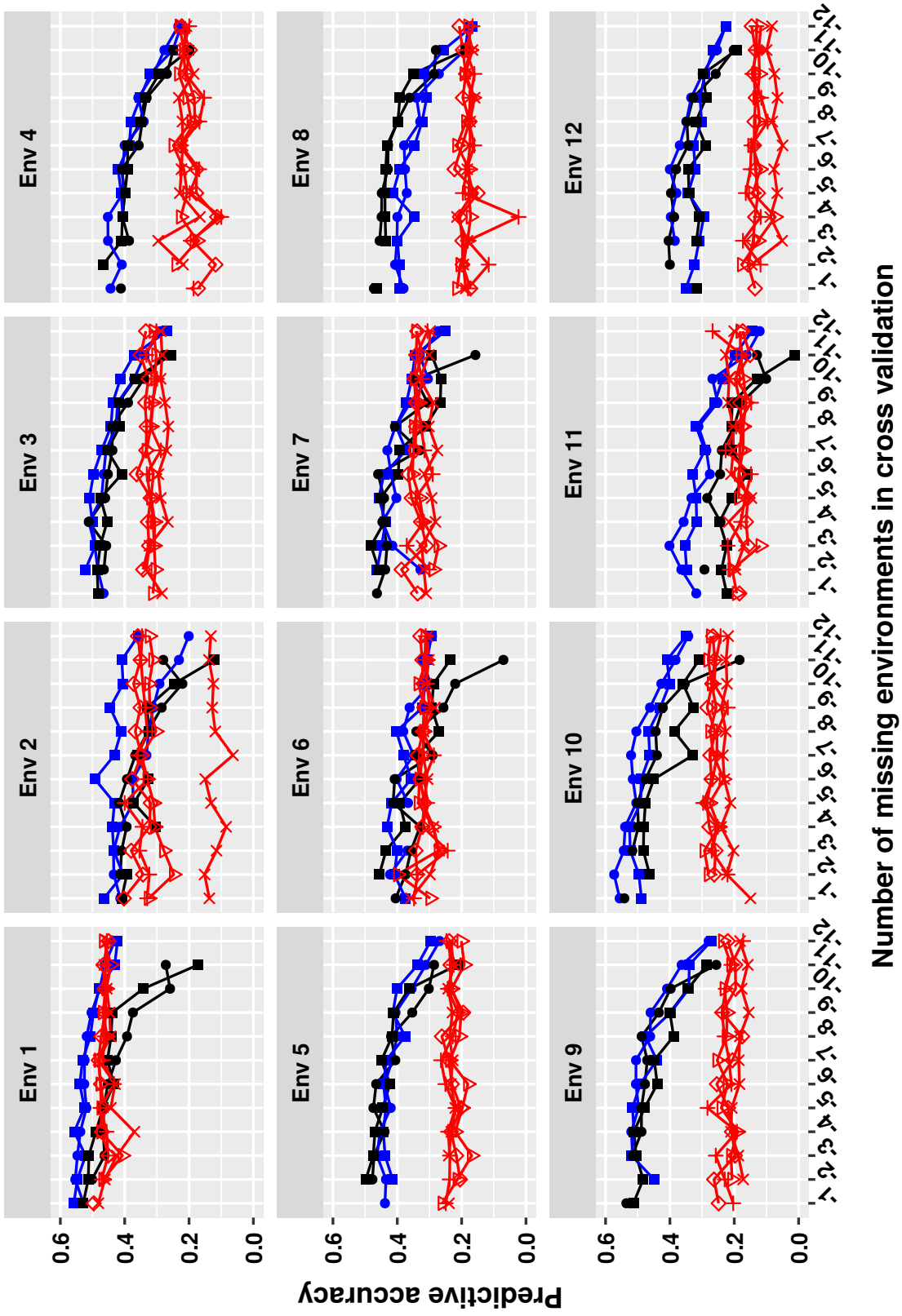

Figure 5: Average predictive accuracies for all models within environments. From left to right, "Env 1" to "Env 12" are abbreviations for environments 1 to 12, respectively. In the x axis, levels of missing environments ranging from one (-1) to all (-12).

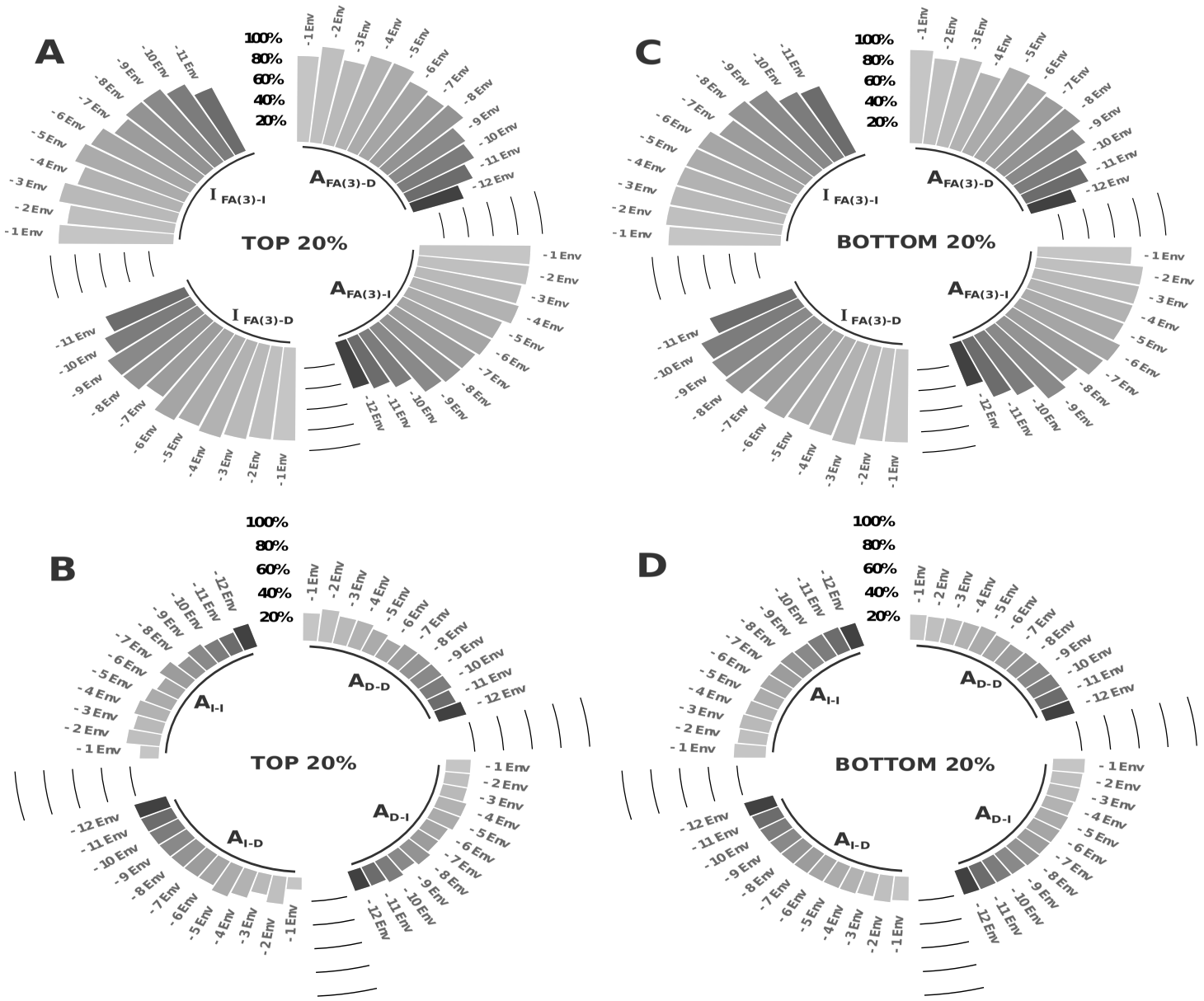

Figure 6: Average coincidence index for selections across environments. In the left side, results of top 20% hybrids, and in the right side, of the bottom 20% hybrids. A and C: FA models with and without incorporating the genomic relationship matrix; B and D: models assuming no correlation across environments of the GxE effects but including genomic relationship information. The legends on each bar shows the level of missing environments at random, from one (-1 Env) to 12 (-12 Env).

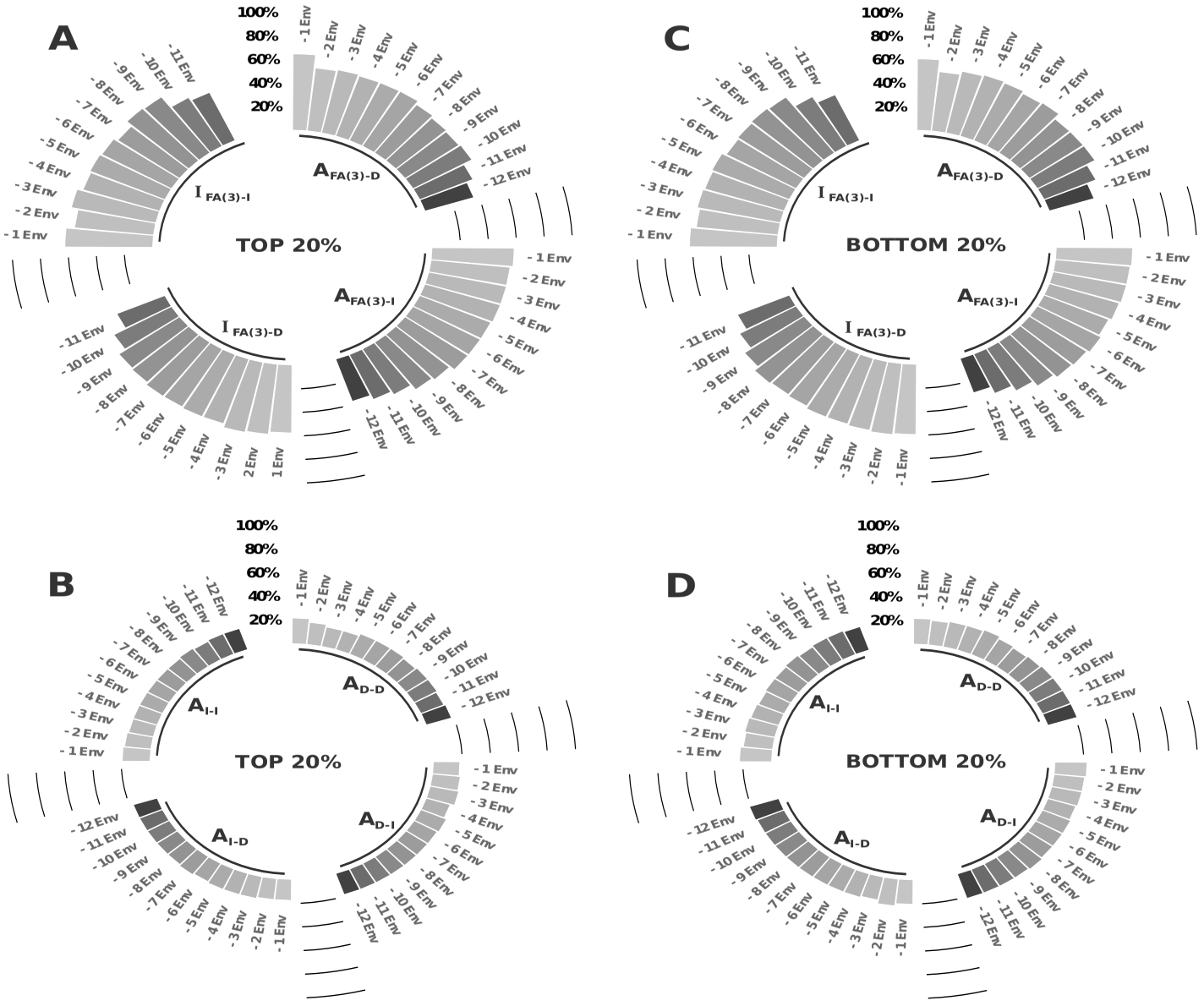

Figure 7: Average coincidence index for selections within environments. In the left side, results of top 20% hybrids, and in the right side, of the bottom 20% hybrids. A and C: FA models with and without incorporating the genomic relationship matrix; B and D: models assuming no correlation across environments of the GxE effects but including genomic relationship information. The legends on each bar shows the level of missing environments at random, from one (-1 Env) to 12 (-12 Env).

APPENDIX

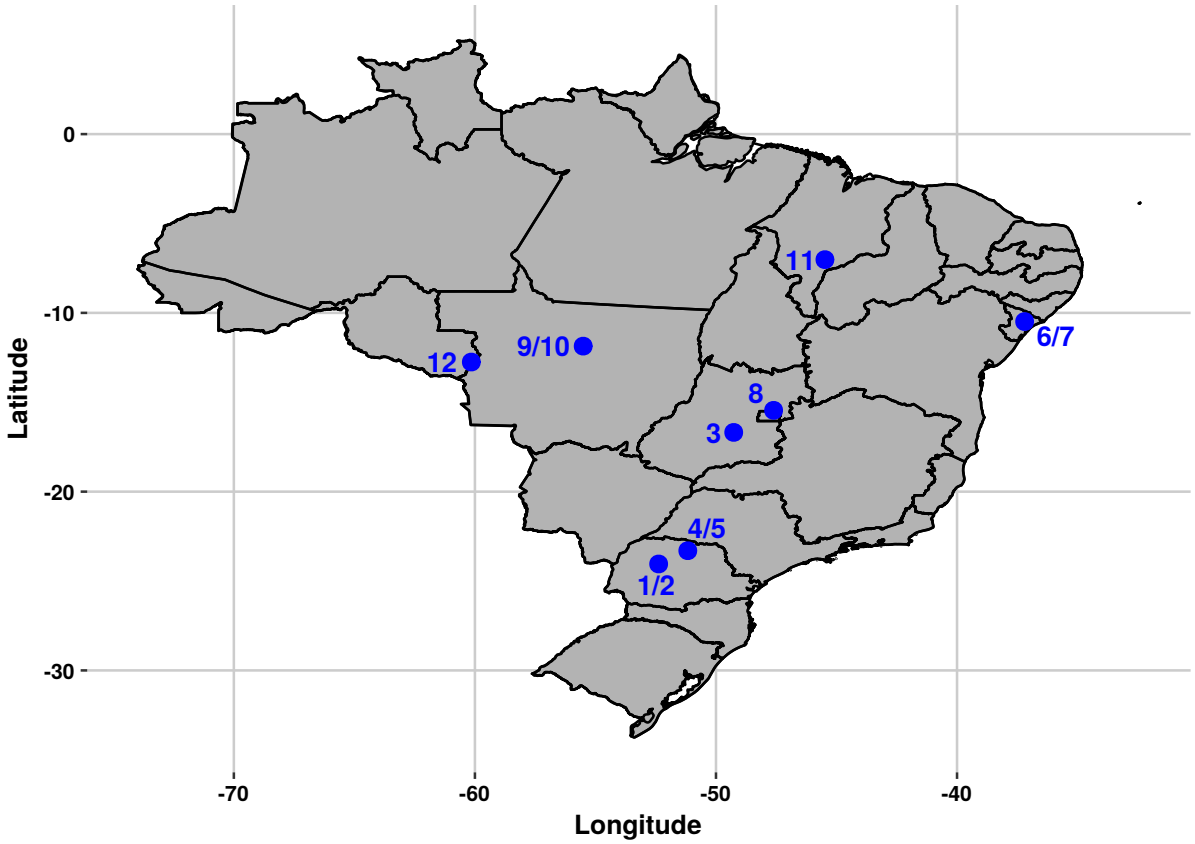

Figure A.8: Geographic coordinates of environments in the map of Brazil. Env 1: first crop season at Campo Mourão - PR, Env 2: second crop season at Campo Mourão - PR; Env 3: first crop season at Goiânia - GO; Env 4: first crop season at Londrina - PR, Env 5: second crop season at Londrina - PR; Env 6: high yield enhancing inputs at Nossa Senhora das Dores - SE, Env 7: standard yield inputs at Nossa Senhora das Dores - SE; Env 8: first crop season at Planaltina - DF; Env 9: first crop season at Sinop - MT, Env 10: second crop season at Sinop - MT; Env 11: first crop season at São Raimundo das Mangabeiras - MA; and Env 12: first crop season at Vilhena - RO.

Table A.3: Average correlations from 5-fold cross-validation among predicted and observed values of single-cross hybrids yield within environments, in all levels of missing environments (CV1 and CV2), for FA models without incorporating the genomic relationship matrix.

Levels[a]**:**	**- 1 Env**		**- 2 Env**		**- 3 Env**		**- 4 Env**		**- 5 Env**		**- 6 Env**	
M. Code[b]**:**	$\mathbf{I_{FA(3)-D}}$	$\mathbf{I_{FA(3)-I}}$	$\mathbf{I_{FA(3)-D}}$	$\mathbf{I_{FA(3)-I}}$	$\mathbf{I_{FA(3)-D}}$	$\mathbf{I_{FA(3)-I}}$	$\mathbf{I_{FA(3)-D}}$	$\mathbf{I_{FA(3)-I}}$	$\mathbf{I_{FA(3)-D}}$	$\mathbf{I_{FA(3)-I}}$	$\mathbf{I_{FA(3)-D}}$	$\mathbf{I_{FA(3)-I}}$
Env 1	0.529	-	0.513	0.501	0.511	0.463	0.490	0.467	0.467	0.465	0.434	0.452
Env 2	0.409	-	0.393	0.414	-	0.410	0.305	0.394	0.372	0.417	0.326	0.393
Env 3	0.483	-	0.485	0.465	0.476	0.457	0.455	0.512	0.475	0.461	0.409	0.452
Env 4	-	0.412	0.467	-	0.410	0.386	0.407	0.407	0.399	0.400	0.391	0.409
Env 5	-	-	0.498	0.477	0.475	0.472	0.470	0.440	0.445	0.474	0.423	0.466
Env 6	-	0.405	0.456	0.374	0.435	0.358	0.375	0.329	0.392	0.413	0.330	0.408
Env 7	-	0.463	0.458	0.438	0.482	0.432	0.440	0.446	0.452	0.441	0.397	0.460
Env 8	0.464	0.473	-	-	0.436	0.455	0.438	0.449	0.438	0.448	0.436	0.429
Env 9	0.513	0.535	0.483	-	0.505	0.516	0.512	0.488	0.479	0.489	0.439	0.478
Env 10	-	0.542	0.463	-	0.483	0.517	0.481	0.498	0.476	0.501	0.450	0.476
Env 11	0.223	-	0.240	0.293	0.224	-	0.246	0.243	0.208	0.284	0.161	0.244
Env 12	0.316	-	-	0.400	0.318	0.404	0.308	0.387	0.341	0.397	0.343	0.381
	- 7 Env		**- 8 Env**		**- 9 Env**		**- 10 Env**		**- 11 Env**			
Env 1	0.454	0.427	0.442	0.393	0.439	0.374	0.342	0.259	0.172	0.272		
Env 2	0.361	0.367	0.326	0.328	0.330	0.285	0.246	0.221	0.121	0.280		
Env 3	0.454	0.438	0.415	0.440	0.419	0.390	0.367	0.340	0.257	0.278		
Env 4	0.391	0.356	0.351	0.345	0.337	0.332	0.293	0.269	0.199	0.249		
Env 5	0.449	0.406	0.416	0.419	0.412	0.354	0.363	0.302	0.212	0.286		
Env 6	0.340	0.290	0.270	0.341	0.293	0.257	0.287	0.220	0.235	0.070		
Env 7	0.393	0.331	0.313	0.407	0.266	0.306	0.264	0.344	0.293	0.157		
Env 8	0.432	0.429	0.398	0.399	0.392	0.362	0.351	0.286	0.188	0.280		
Env 9	0.449	0.471	0.389	0.488	0.400	0.434	0.343	0.398	0.285	0.256		
Env 10	0.328	0.440	0.386	0.444	0.327	0.422	0.359	0.354	0.309	0.184		
Env 11	0.208	0.239	0.208	0.201	0.207	0.178	0.130	0.101	0.012	0.129		
Env 12	0.289	0.343	0.319	0.347	0.287	0.328	0.298	0.258	0.193	0.202		

[a] Levels of missing environments at random ranging from one (-1 Env) to all (-12 Env). [b] Abbreviation for models $I_{FA(3)-D}$ and $I_{FA(3)-I}$.

Table A.4: Average correlations from 5-fold cross-validation among predicted and observed values of single-cross hybrids yield within environments, in all levels of missing environments (CV1 and CV2), for FA models with genomic relationship information.

Levels[a]	**- 1 Env**		**- 2 Env**		**- 3 Env**		**- 4 Env**		**- 5 Env**		**- 6 Env**	
M. Nu.[b]	$\mathbf{A_{FA(3)-D}}$	$\mathbf{A_{FA(3)-I}}$	$\mathbf{A_{FA(3)-D}}$	$\mathbf{A_{FA(3)-I}}$	$\mathbf{A_{FA(3)-D}}$	$\mathbf{A_{FA(3)-I}}$	$\mathbf{A_{FA(3)-D}}$	$\mathbf{A_{FA(3)-I}}$	$\mathbf{A_{FA(3)-D}}$	$\mathbf{A_{FA(3)-I}}$	$\mathbf{A_{FA(3)-D}}$	$\mathbf{A_{FA(3)-I}}$
Env 1	0.559	0.528	0.547	0.554	0.522	0.547	0.557	0.538	0.525	0.519	0.539	0.526
Env 2	0.464	0.414	0.405	0.434	0.434	0.435	0.439	0.414	0.432	0.374	0.493	0.375
Env 3	-	0.466	0.522	0.483	0.491	0.493	0.501	0.511	0.509	0.472	0.497	0.465
Env 4	-	0.444	-	0.409	-	0.452	0.405	0.452	0.411	0.405	0.420	0.407
Env 5	-	0.438	0.415	0.435	0.438	0.463	0.443	0.454	0.436	0.421	0.443	0.440
Env 6	0.374	-	0.409	0.423	0.400	0.369	0.432	-	0.417	0.367	0.357	0.402
Env 7	-	-	0.463	0.328	0.449	0.416	0.436	0.446	0.458	0.403	0.442	0.419
Env 8	0.392	0.381	0.394	0.407	0.399	0.401	0.347	0.400	0.417	0.371	0.393	0.376
Env 9	-	-	0.450	-	0.519	0.516	0.512	0.520	0.516	0.498	0.492	0.506
Env 10	0.488	0.556	0.497	0.573	0.523	0.543	0.526	0.539	0.491	0.504	0.489	0.515
Env 11	-	0.318	0.348	0.364	0.353	0.401	0.317	0.357	0.319	0.333	0.329	0.276
Env 12	0.351	-	0.324	-	0.309	0.385	0.295	0.397	0.344	0.380	0.321	0.400
	- 7 Env		**- 8 Env**		**- 9 Env**		**- 10 Env**		**- 11 Env**		**- 12 Env (CV1)**	
Env 1	0.531	0.524	0.506	0.518	0.497	0.503	0.480	0.477	0.431	0.462	0.422	0.431
Env 2	0.432	0.333	0.411	0.325	0.445	0.311	0.405	0.292	0.409	0.231	0.359	0.201
Env 3	0.472	0.458	0.443	0.427	0.437	0.422	0.413	0.371	0.371	0.338	0.269	0.289
Env 4	0.386	0.400	0.380	0.341	0.352	0.358	0.321	0.310	0.251	0.276	0.225	0.233
Env 5	0.434	0.416	0.376	0.401	0.414	0.405	0.401	0.355	0.337	0.311	0.297	0.268
Env 6	0.380	0.349	0.403	0.381	0.322	0.362	0.314	0.297	0.308	0.322	0.294	0.292
Env 7	0.373	0.431	0.404	0.406	0.372	0.360	0.355	0.305	0.345	0.330	0.251	0.273
Env 8	0.346	0.379	0.323	0.329	0.309	0.340	0.317	0.271	0.257	0.189	0.167	0.179
Env 9	0.441	0.505	0.476	0.461	0.398	0.460	0.341	0.409	0.340	0.364	0.272	0.280
Env 10	0.463	0.521	0.467	0.504	0.430	0.462	0.401	0.427	0.409	0.383	0.350	0.343
Env 11	0.291	0.287	0.318	0.311	0.261	0.252	0.236	0.268	0.199	0.162	0.144	0.121
Env 12	0.327	0.369	0.302	0.349	0.306	0.334	0.291	0.298	0.265	0.254	0.226	0.227

[a] Levels of missing environments at random ranging from one (-1 Env) to all (-12 Env). [b] Abbreviation for models $A_{FA(3)-D}$ and $A_{FA(3)-I}$.

Table A.5: Average correlations from 5-fold cross-validation among predicted and observed values of single-cross hybrids yield within environments, in all levels of missing environments (CV1 and CV2), for models assuming no correlation across environments of the GxE effects but including genomic relationship information.

Levels[a]	**- 1 Env**		**- 2 Env**		**- 3 Env**		**- 4 Env**		**- 5 Env**		**- 6 Env**	
M. Code[b]	$\mathbf{A_{D-D}}$	$\mathbf{A_{D-I}}$	$\mathbf{A_{D-D}}$	$\mathbf{A_{D-I}}$	$\mathbf{A_{D-D}}$	$\mathbf{A_{D-I}}$	$\mathbf{A_{D-D}}$	$\mathbf{A_{D-I}}$	$\mathbf{A_{D-D}}$	$\mathbf{A_{D-I}}$	$\mathbf{A_{D-D}}$	$\mathbf{A_{D-I}}$
Env 1	-	0.482	0.471	0.462	0.452	0.424	0.456	0.370	0.476	0.443	0.433	0.430
Env 2	0.330	0.137	0.323	0.152	0.354	0.115	0.345	0.085	0.400	0.132	0.328	0.150
Env 3	-	0.284	-	0.333	0.308	0.323	0.312	0.266	0.320	0.288	0.331	0.296
Env 4	0.186	-	-	0.219	0.195	0.296	0.100	0.166	0.198	0.229	0.168	0.223
Env 5	0.249	0.239	0.237	-	0.237	0.240	0.242	0.232	0.219	0.220	0.250	0.234
Env 6	0.348	0.358	0.337	0.298	0.243	0.270	0.321	0.288	0.307	0.320	0.315	0.306
Env 7	-	0.310	0.301	0.317	0.371	0.324	0.317	0.282	0.343	0.293	0.290	0.316
Env 8	0.180	0.187	0.116	0.195	0.181	0.182	0.023	0.214	0.196	0.164	0.184	0.183
Env 9	0.203	-	0.230	0.172	0.258	0.187	0.192	0.210	0.283	0.208	0.209	0.184
Env 10	-	0.150	0.221	0.223	0.270	0.202	0.251	0.241	0.297	0.212	0.267	0.231
Env 11	-	-	0.213	0.196	0.220	0.170	0.179	0.216	0.167	0.147	0.148	0.210
Env 12	-	-	0.118	0.146	0.173	0.051	0.118	0.088	0.164	0.066	0.150	0.076
	- 7 Env		**- 8 Env**		**- 9 Env**		**- 10 Env**		**- 11 Env**		**- 12 Env (CV1)**	
Env 1	0.468	0.481	0.462	0.444	0.460	0.447	0.459	0.456	0.462	0.463	0.455	0.455
Env 2	0.359	0.064	0.330	0.118	0.336	0.128	0.344	0.124	0.350	0.138	0.346	0.132
Env 3	0.288	0.271	0.321	0.264	0.314	0.275	0.299	0.288	0.332	0.284	0.301	0.288
Env 4	0.225	0.220	0.168	0.220	0.153	0.232	0.211	0.186	0.218	0.218	0.199	0.208
Env 5	0.264	0.228	0.228	0.237	0.211	0.229	0.243	0.242	0.225	0.225	0.247	0.233
Env 6	0.286	0.320	0.321	0.318	0.300	0.283	0.308	0.310	0.315	0.302	0.312	0.307
Env 7	0.316	0.275	0.326	0.302	0.337	0.289	0.323	0.330	0.341	0.302	0.306	0.298
Env 8	0.161	0.174	0.178	0.174	0.173	0.160	0.160	0.189	0.188	0.166	0.166	0.185
Env 9	0.216	0.186	0.231	0.183	0.238	0.155	0.227	0.177	0.209	0.158	0.172	0.182
Env 10	0.254	0.235	0.254	0.226	0.220	0.240	0.263	0.223	0.257	0.225	0.242	0.219
Env 11	0.227	0.203	0.195	0.204	0.147	0.220	0.194	0.216	0.176	0.226	0.267	0.200
Env 12	0.147	0.050	0.096	0.082	0.116	0.066	0.144	0.075	0.131	0.099	0.131	0.083

[a] Levels of missing environments at random ranging from one (-1 Env) to all (-12 Env). [b] Abbreviation for models A_{D-D} and A_{D-I}.

Table A.6: Average correlations from 5-fold cross-validation among predicted and observed values of single-cross hybrids yield within environments, in all levels of missing environments (CV1 and CV2), for models assuming no correlation across environments of the GxE effects but including genomic relationship information.

Levels[a]	**- 1 Env**		**- 2 Env**		**- 3 Env**		**- 4 Env**		**- 5 Env**		**- 6 Env**	
M. Code[b]	$\mathbf{A_{I-D}}$	$\mathbf{A_{I-I}}$	$\mathbf{A_{I-D}}$	$\mathbf{A_{I-I}}$	$\mathbf{A_{I-D}}$	$\mathbf{A_{I-I}}$	$\mathbf{A_{I-D}}$	$\mathbf{A_{I-I}}$	$\mathbf{A_{I-D}}$	$\mathbf{A_{I-I}}$	$\mathbf{A_{I-D}}$	$\mathbf{A_{I-I}}$
Env 1	0.496	-	-	0.467	0.451	0.407	0.471	0.473	0.465	0.472	0.474	0.476
Env 2	0.402	0.325	0.343	0.247	0.380	0.278	0.323	0.306	0.319	0.310	0.377	0.330
Env 3	-	0.312	0.343	0.305	0.325	0.308	0.328	0.307	0.319	0.318	0.363	0.311
Env 4	0.172	-	0.117	0.240	0.189	0.174	0.113	0.226	0.177	0.206	0.177	0.216
Env 5	-	0.260	0.205	0.213	0.217	0.169	0.233	0.220	0.202	0.198	0.230	0.181
Env 6	-	0.298	0.339	0.398	0.343	0.271	0.323	0.289	0.316	0.334	0.329	0.331
Env 7	0.337	-	0.387	0.288	0.310	0.272	0.327	0.319	0.356	0.330	0.366	0.354
Env 8	0.173	0.214	0.197	0.204	0.198	0.181	0.206	0.172	0.150	0.177	0.222	0.176
Env 9	0.249	-	0.263	0.207	0.198	0.210	0.208	0.190	0.219	0.240	0.254	0.221
Env 10	-	-	0.262	0.280	0.258	0.293	0.278	0.248	0.285	0.276	0.271	0.233
Env 11	0.184	0.192	-	0.206	0.157	0.119	0.163	-	0.158	0.186	0.184	0.187
Env 12	0.135	-	0.157	0.176	0.141	0.126	0.136	0.073	0.126	0.142	0.134	0.119
	- 7 Env		**- 8 Env**		**- 9 Env**		**- 10 Env**		**- 11 Env**		**- 12 Env (CV1)**	
Env 1	0.481	0.481	0.473	0.454	0.462	0.468	0.463	0.457	0.467	0.442	0.450	0.465
Env 2	0.348	0.348	0.367	0.303	0.351	0.322	0.372	0.324	0.351	0.312	0.359	0.323
Env 3	0.333	0.332	0.332	0.318	0.337	0.312	0.326	0.305	0.351	0.302	0.334	0.310
Env 4	0.228	0.249	0.181	0.193	0.184	0.206	0.209	0.232	0.197	0.216	0.223	0.230
Env 5	0.240	0.237	0.262	0.207	0.195	0.202	0.234	0.231	0.235	0.192	0.239	0.201
Env 6	0.337	0.320	0.320	0.317	0.317	0.316	0.319	0.335	0.322	0.308	0.329	0.313
Env 7	0.345	0.349	0.342	0.349	0.343	0.332	0.347	0.346	0.342	0.322	0.338	0.342
Env 8	0.199	0.214	0.175	0.186	0.197	0.164	0.187	0.191	0.197	0.172	0.207	0.180
Env 9	0.206	0.253	0.176	0.224	0.237	0.222	0.196	0.230	0.197	0.210	0.220	0.235
Env 10	0.275	0.264	0.259	0.275	0.284	0.256	0.269	0.270	0.261	0.281	0.266	0.273
Env 11	0.182	0.173	0.177	0.172	0.182	0.164	0.198	0.171	0.151	0.181	0.174	0.181
Env 12	0.138	0.144	0.138	0.121	0.134	0.128	0.140	0.121	0.134	0.124	0.146	0.117

[a] Levels of missing environments at random ranging from one (-1 Env) to all (-12 Env). [b] Abbreviation for models A_{I-D} and A_{I-I}.

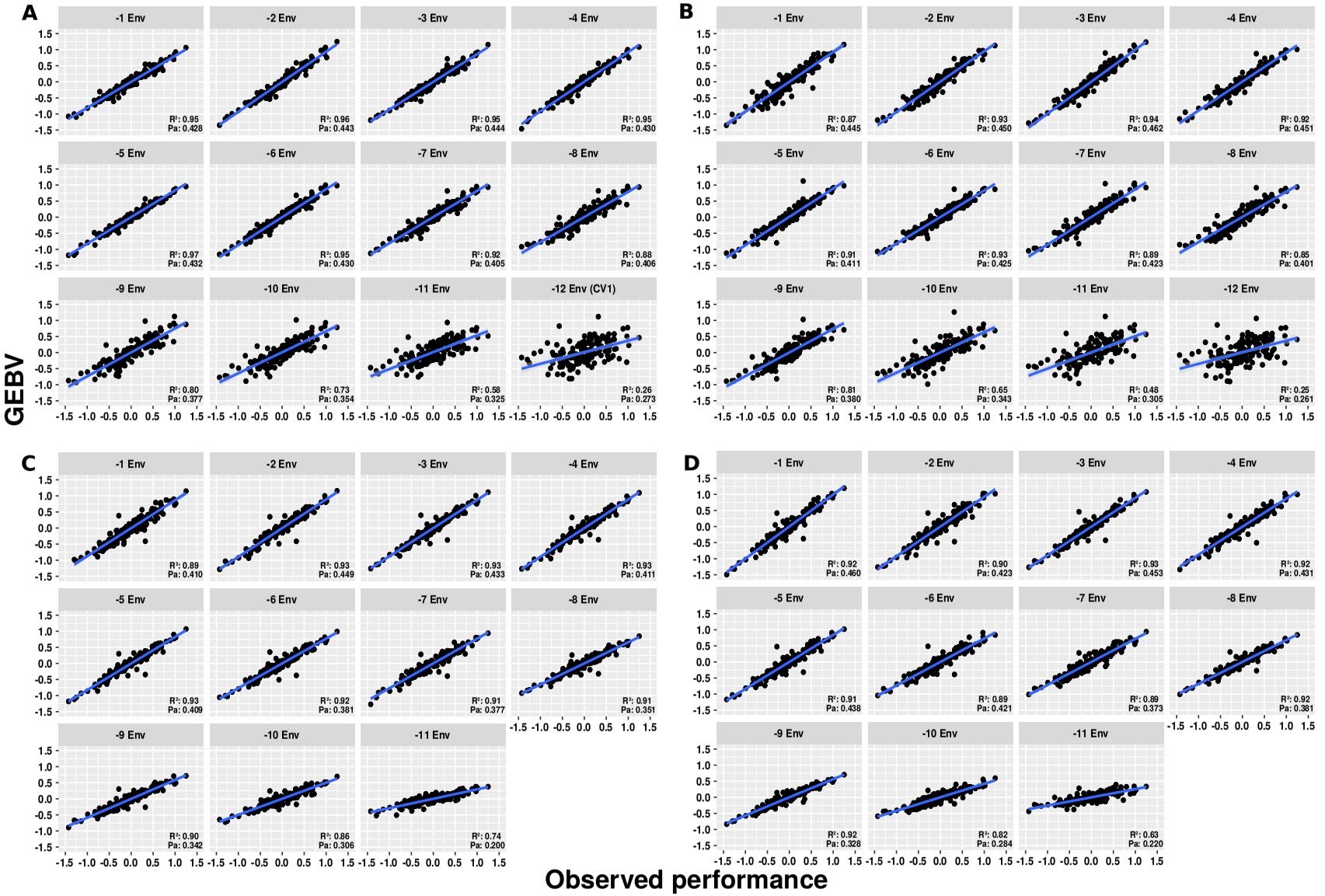

Figure A.9: Genomic estimated breeding value (GEBV) against observed performance plot across environments, for FA models with (A: $A_{FA(3)-D}$, B: $A_{FA(3)-I}$) and without (C: $I_{FA(3)-D}$, D: $I_{FA(3)-I}$) incorporating the genomic relationship matrix. The legends in each scatter plot shows the level of missing environments, ranging from one (-1 Env) to 12 (-12 Env) missing environments at random, the coefficient of determination (R^2) of regression (blue line) and the average predictive accuracy across environments (Pa.).

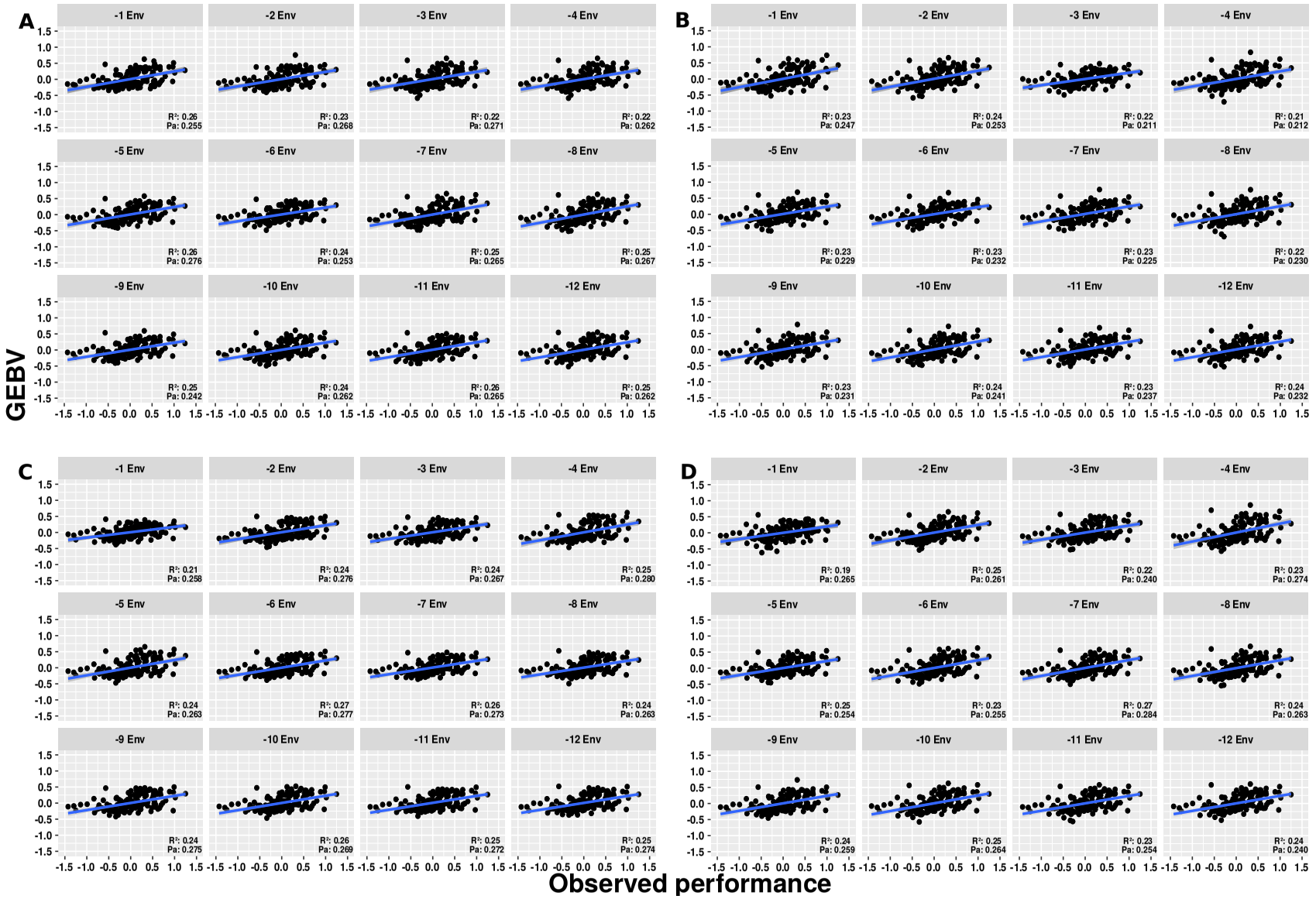

Figure A.10: Genomic estimated breeding value (GEBV) against observed performance plot across environments, for models assuming no correlation across environments of the GxE effects but including genomic relationship information (A: A_{D-D}, B: A_{D-I}, C: A_{I-D} and D: A_{I-I}). The legends in each scatter plot shows the level of missing environments, ranging from one (-1 Env) to 12 (-12 Env) missing environments at random, the coefficient of determination (R^2) of regression (blue line) and the average predictive accuracy across environments (Pa.).