Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

SINGLE-STEP GENOMIC AND PEDIGREE GENOTYPE × ENVIRONMENT INTERACTION MODELS FOR PREDICTING WHEAT LINES IN INTERNATIONAL ENVIRONMENTS

Paulino Pérez-Rodríguez[1], José Crossa[2*], Jessica Rutkoski[2], Jesse Poland[3], Ravi Singh[2], Andrés Legarra[4], Enrique Autrique[2], Gustavo de los Campos[5], Juan Burgueño[2], and Susanne Dreisigacker[2]

[1] Colegio de Postgraduados, CP 56230, Montecillos, Estado de México, México.

[2] International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 México City, México.

[3] J. Poland, USDA-ARS and Dep. of Agronomy, Kansas State Univ. (KSU), 4011 Throckmorton Hall, Manhattan KS, 66506.

[4] INRA, UR631 SAGA, BP 52627, 32326 Castanet-T, France.

[5] Department of Epidemiology & Biostatistics, Michigan State University, 909 Fee Road, Room B601, East Lansing, MI 48824, USA.

[*] Corresponding author (j.crossa@cgiar.org)

ABSTRACT

Genomic prediction models have been commonly employed in plant breeding but only in reduced data sets comprising a few hundred genotyped individual plants. However, pedigree information for an entire breeding population is frequently available, as are historical data on the performance of a large number of selection candidates. The single-step method extends the genomic relationship information of genotyped individuals to pedigree information from a larger number of phenotyped individuals so as to combine relationship information on all members of the breeding population. Furthermore, genomic prediction models that incorporate genotype × environment interaction (G×E) have produced substantial increases in prediction accuracy compared to single-environment genomic prediction models. The main objective of this study was to show how to use single-step genomic and pedigree models to assess the prediction accuracy of a large number of CIMMYT wheat lines (58,798) evaluated in several simulated

environments in Cd. Obregon (Mexico), and predict the grain yield performance of some of them in several sites of South Asia (India, Pakistan and Bangladesh) using a reaction norm model that incorporates G×E. Another objective of this study was to describe the statistical and computational challenges encountered when developing the pedigree and single-step models in such large data sets. Results indicate that genomic prediction accuracy achieved by models using only pedigree, only markers, or both pedigree and markers to predict varying environments in India, Pakistan and Bangladesh is higher (0.25-0.38) than prediction accuracy of models that use only phenotypic prediction (0.20) or that do not include the G×E term.

Keywords: wheat pedigree prediction, wheat genomic prediction, single-step prediction accuracy, genotype × environment interaction, international environments.

INTRODUCTION

Global wheat production is increasing less than 1% annually, and recently wheat yields have stagnated in many regions of South Asia (Ray et al., 2012). In South Asia, the wheat crop is already being grown under high temperature conditions; however, due to climate change, temperatures could increase well beyond the optimal for growing wheat, which would further reduce grain yield. As a result, South Asian countries may not be able to meet the region's already growing demand for wheat grain.

Well managed crop improvement programs are necessary to increase food production in different parts of the world. Several molecular marker methods have proven their relevance in different cereal crops. Genomic selection (GS) is becoming a standard approach to achieve genetic progress in plants because it reduces the generation interval by reducing the need to have progeny field-tested every cycle. Breeding values can be predicted as the sum of the effects of all markers by regressing the values of the phenotypes on all markers (Meuwissen et al., 2001). Several authors have successfully implemented GS in plant breeding with intermediate-to-high density marker coverage for traits such as grain yield, biomass yield, resistance to several diseases, and flowering evaluated under different environmental conditions. Studies have demonstrated that some of the factors determining prediction accuracy in GS are the heritability of the trait, the number of markers, the size of the training population, the relationship between

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

the training and the testing sets, and genotype × environment interaction (G×E) (de los Campos et al., 2009; Crossa et al., 2010, 2011; Pérez-Rodríguez et al., 2012; Burgueño et al., 2012; Hickey et al., 2012; González-Camacho et al., 2012; Riedelsheimer et al., 2012; Weber et al., 2012). Furthermore, including high-density marker platforms with G×E interactions adds power to GS models (Burgueño et al., 2012; Jarquín et al., 2014; López-Cruz et al., 2015; Heslot et al., 2012).

Recently, genomic predictions have been extensively studied in bread wheat using elite germplasm sets (de los Campos et al., 2009, 2010; Crossa et al., 2010; Gónzalez-Camacho et al., 2012; Heslot et al., 2012; Pérez-Rodríguez et al., 2012; López-Cruz et al., 2015). The results have proven that the use of dense molecular markers coupled with pedigree information increases the prediction accuracy of unobserved phenotypes. One of the problems usually encountered by GS in animal and plant breeding is that the number of evaluated lines exceeds the number of genotyped lines, due to genotypic costs. Nejati-Javaremi et al. (1997) were the first to propose incorporating genotypic information for predicting the breeding values of animals in a manner similar to the way pedigree information is used in the Best Linear Unbiased Predictor (BLUP) method. When the pedigrees of all phenotyped individuals were available but only some were genotyped, dairy cattle researchers (Misztal et al., 2009; Legarra et al. 2009; Aguilar et al. 2010, 2011; Christensen et al., 2012) derived a unified (single-step) computation approach for Genomic Best Linear Unbiased Predictor (ssGBLUP) for combining phenotypic, pedigree and genomic information based on Henderson's (1975, 1976) standard mixed model equations. These authors augmented pedigree-based relationship matrix (***A***) with contributions from genomic relationship matrix (***G***) of the genotyped individuals. They showed how to modify the original ***A*** matrix to obtain an ***H*** matrix that includes not only the pedigree-based relationship matrix but also a matrix that contains the differences between genomic-based and pedigree-based matrices. These authors also developed efficient computer algorithms for inverting matrix ***H*** computed from large numbers (millions) of animals in the data.

Although augmenting matrix ***A*** by using only a fraction of the individuals that were genotyped would reduce genotyping costs, the ssGBLUP method has not been extensively applied in plant breeding. Just recently, Ashraf et al. (2016) were the first to investigate the impact on prediction accuracy when some wheat lines were not genotyped and only pedigree and phenotype information was available; the authors concluded that the ssGBLUP method for

deriving matrix ***H*** can provide higher prediction accuracy than either genomic or pedigree-based prediction. In plants, the ssGBLUP approach proposed by Ashraf et al. (2016) has been used with a limited number of lines. The approach has not been tested on large datasets, e.g., every cropping cycle, CIMMYT's Global Wheat Program (GWP) generates thousands of new breeding lines that are candidates for field evaluation. Applying GS in the GWP is economically feasible (1) when advancing breeding lines in the first preliminary yield trials to predict the performance of the selected lines in multi-environment trials or (2) for predicting a selected set of lines in different international target environments using as a training set the parents evaluated in Mexico and the progeny to be predicted in international environments such as those in South Asia.

In recent years, the GWP aimed at forming a large reference data set comprising 58,798 breeding lines, including their phenotypic and pedigree data from the last seven cropping cycles in Cd. Obregon (Mexico) and South Asia. This large reference set contains complete phenotypic data and pedigree information; however, only 29,484 of the lines have been genotyped. Therefore, an ***H*** matrix that combines wheat lines having only molecular markers with those having pedigree and phenotype must be generated.

The main objectives of this study were (1) to use the large reference set for predicting the performance of wheat lines in several environments in South Asia; and (2) to perform the predictions using phenotypic, pedigree and genomic information to genetically evaluate the wheat lines using a single-step model that combines pedigree and marker information into a unified ***H*** matrix. Here we used information for genotyped and non-genotyped individuals combined, by applying the method proposed by Legarra et al. (2009) and Aguilar et al. (2010). Prediction accuracy was studied using a G×E interaction multiplicative model (the reaction norm model of Jarquín et al., 2014) with pedigree information (***A***), genomic information (***G***), or both (***H***) and comparing its prediction accuracy results to those of a genomic model that does not include G×E interaction. This reaction norm model uses highly random dimensional matrices for the genomic and pedigree matrices. We also describe the statistical and computational challenges encountered when developing the pedigree and single-step models in such large data sets.

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

MATERIALS AND METHODS

Experimental data

The data set included a total of 58,798 wheat lines that were evaluated at the Norman E. Borlaug Experiment Research Station in Ciudad Obregon, Mexico, under various field management conditions (Optimal, Drought, Late Heat, Severe Drought and Early Heat) during seven cycles (2009-2016). Some of the lines were also evaluated under the same conditions in South Asia (Jalbapur, Ludhiana and Pusa in India; Faisalabad in Pakistan; and Jamalpur in Bangladesh) during 2013-2016. Original data from each year comprise a large number of trials, each established using an alpha-lattice design with three replicates. The field management conditions under which each trial was established in each year are described in Table 1. The conditions-location combinations will be referred to as environments. Table 2 shows the number of lines evaluated in each environment.

The basic model fitted to each of the 12 environments described in Table 2 comprises the random effects of the trials, the random effects of the replicates within trials, the random effects of the incomplete blocks within trials and replicates, and the random effects of the breeding lines.

A pedigree relationship matrix (***A***) for the 58,798 individuals was computed using a modified version of the software 'pedigreemm' (Bates and Vazquez, 2009) that accounts for self-pollination; the latest version of the routines can be found at https://github.com/Rpedigree/pedigreeR. Given the dimensions of ***A***, it is difficult to hold it in RAM memory and compute it. Appendix A shows the small R script (R Core Team, 2016) that was used to obtain and store the relationship matrix. It uses results from partitioned matrices to obtain the result and speed up the computations; R was recompiled from source and linked against OpenBLAS (http://www.openblas.net). For further details on the computations, see Appendix A. In total, 29,484 individuals were genotyped using Genotyping by Sequencing (GBS) (e.g., Elshire et al., 2011). We kept all the SNP markers and imputed the missing values using observed data. Markers with minor allele frequency (MAF) of less than 0.05 were removed; after this process, 9,045 markers were available for prediction.

Statistical models

Recently, Jarquín et al. (2014) and López-Cruz et al. (2015) proposed statistical models for performing genomic predictions taking into account G×E. The models were originally developed to incorporate genetic information from molecular markers and, in the case of Jarquín's model, it is also possible to incorporate environmental covariates. The Jarquín model has also shown to be useful when the genetic information is obtained from a pedigree (Pérez-Rodríguez et al., 2015). Here we describe Jarquín's model based on genomic and pedigree information. To speed up the computations and make them feasible, we re-parametrized the original model by using very well-known results from Cholesky decomposition and mixed models (e.g., Henderson, 1976; Harville and Callanan, 1989).

Model 1: G×E interaction using pedigree

The parametric G×E interaction model takes into account the main effect of E environments, the main effect of genotypes and the interaction between genotypes and the environment. In matrix notation, the model can be written as:

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_E\boldsymbol{\beta}_E + \boldsymbol{Z}_g\boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{e}, \quad (1)$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_E)'$ is the response vector, and $\boldsymbol{y}_j$ represents the observations in the j-th environment ($j = 1, \ldots, E$). The general mean is μ; $\boldsymbol{Z}_E$ is an incidence matrix for environments, which is assumed multivariate with $\boldsymbol{\beta}_E \sim MN(\mathbf{0}, \sigma_E^2\boldsymbol{I})$; $\boldsymbol{Z}_g$ is an incidence matrix that connects genotypes with phenotypes; $\boldsymbol{u}_1$ represents the random effect of genotypes; it is assumed multivariate with $\boldsymbol{u}_1 \sim MN(\mathbf{0}, \sigma_u^2\boldsymbol{A})$; and $\boldsymbol{u}_2$ represents the effect of G×E interaction. We assume $\boldsymbol{u}_2 \sim MN(\mathbf{0}, \sigma_{ge}^2(\boldsymbol{Z}_g\boldsymbol{G}\boldsymbol{Z}_g')\#(\boldsymbol{Z}_E\boldsymbol{Z}_E'))$, where # denotes the Hadamard product (cell-by-cell) of the two matrices in parentheses (see Jarquín et al., 2014; Pérez-Rodríguez et al., 2015). Finally, we assume that the residuals are distributed as follows: $\boldsymbol{e} \sim MN(\mathbf{0}, \sigma_e^2\boldsymbol{I})$.

Since $\boldsymbol{A}$ is positive definite and symmetric, it can be factored as $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}'$ by using Cholesky decomposition. Therefore, from (1):

$$\boldsymbol{Z}_g\boldsymbol{u}_1 \overset{d}{=} \boldsymbol{Z}_g\boldsymbol{L}\boldsymbol{u}_1^* \quad (2)$$

where $\boldsymbol{u}_1^* \sim MN(\mathbf{0}, \sigma_u^2\boldsymbol{I})$. Furthermore, it is not necessary to perform the $\boldsymbol{Z}_g\boldsymbol{L}$ product because for each row of the resulting matrix, we just need to copy the k-th row of $\boldsymbol{L}$, where k is the column in the i-th row of $\boldsymbol{Z}_g$ that is different from zero, that is, $\boldsymbol{Z}_g(i, k)$ is equal to one. The matrix $\boldsymbol{Z}_E\boldsymbol{Z}_E'$ is block diagonal; blocks different from zero correspond to matrices with ones, i.e.,

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

$$\boldsymbol{Z}_E\boldsymbol{Z}_E' = \begin{pmatrix} \boldsymbol{J}_1 & & & \\ & \boldsymbol{J}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{J}_E \end{pmatrix}, \quad (3)$$

where $\boldsymbol{J}_j$ $(j = 1, \ldots, E)$ is a square matrix with ones whose dimensions correspond to the number of genotypes evaluated in environment j. Since $\boldsymbol{Z}_E\boldsymbol{Z}_E'$ is block diagonal, in order to compute $\boldsymbol{Z}_g\boldsymbol{A}\boldsymbol{Z}_g'\#\boldsymbol{Z}_E\boldsymbol{Z}_E'$, we just need to compute the corresponding block elements in the diagonal of $\boldsymbol{Z}_g\boldsymbol{A}\boldsymbol{Z}_g' = \boldsymbol{Z}_g\boldsymbol{L}\boldsymbol{L}'\boldsymbol{Z}_g'$. Let $\boldsymbol{Z}_g\boldsymbol{L} = \widetilde{\boldsymbol{Z}}$; then

$$\widetilde{\boldsymbol{A}} = \widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{Z}}' = \begin{pmatrix} \widetilde{\boldsymbol{Z}}_{11} & \widetilde{\boldsymbol{Z}}_{12} & \cdots & \widetilde{\boldsymbol{Z}}_{1E} \\ \widetilde{\boldsymbol{Z}}_{21} & \widetilde{\boldsymbol{Z}}_{22} & \cdots & \widetilde{\boldsymbol{Z}}_{2E} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\boldsymbol{Z}}_{E1} & \widetilde{\boldsymbol{Z}}_{E2} & \cdots & \widetilde{\boldsymbol{Z}}_{EE} \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{Z}}_{11} & \widetilde{\boldsymbol{Z}}_{12} & \cdots & \widetilde{\boldsymbol{Z}}_{1E} \\ \widetilde{\boldsymbol{Z}}_{21} & \widetilde{\boldsymbol{Z}}_{22} & \cdots & \widetilde{\boldsymbol{Z}}_{2E} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\boldsymbol{Z}}_{E1} & \widetilde{\boldsymbol{Z}}_{E2} & \cdots & \widetilde{\boldsymbol{Z}}_{EE} \end{pmatrix}',$$

The block diagonal elements of $\widetilde{\boldsymbol{A}}$, can be computed as follows:

$$\begin{aligned} \widetilde{\boldsymbol{A}}_{11} &= \sum_{Environments} \widetilde{\boldsymbol{Z}}_{1j}\widetilde{\boldsymbol{Z}}_{1j}' = \boldsymbol{A}_{11} \\ &\vdots \\ \widetilde{\boldsymbol{A}}_{EE} &= \sum_{Environments} \widetilde{\boldsymbol{Z}}_{Ej}\widetilde{\boldsymbol{Z}}_{Ej}' = \boldsymbol{A}_{EE} \end{aligned} \quad (4)$$

where $\boldsymbol{A}_{jj}$, corresponds to the relationship matrix for individuals evaluated in environment j. From (3) and (4) and by using Cholesky decomposition, the term $\boldsymbol{Z}_g\boldsymbol{A}\boldsymbol{Z}_g'\#\boldsymbol{Z}_E\boldsymbol{Z}_E'$ can be obtained as follows:

$$\boldsymbol{Z}_g\boldsymbol{A}\boldsymbol{Z}_g'\#\boldsymbol{Z}_E\boldsymbol{Z}_E' = BDiag(\boldsymbol{A}_{11}, \ldots, \boldsymbol{A}_{EE}) = BDiag(\boldsymbol{L}_1\boldsymbol{L}_1', \ldots, \boldsymbol{L}_E\boldsymbol{L}_E') = \boldsymbol{L}_{ge}\boldsymbol{L}_{ge}', \quad (5)$$

where $\boldsymbol{L}_{ge} = BDiag(\boldsymbol{L}_1, \ldots, \boldsymbol{L}_E)$. Therefore, from (5):

$$\boldsymbol{u}_2 \overset{d}{=} \boldsymbol{L}_{ge}\boldsymbol{u}_2^*, \quad (6)$$

where $\boldsymbol{u}_2^* \sim MN(\boldsymbol{0}, \sigma_{ge}^2\boldsymbol{I})$ and "$\overset{d}{=}$" stands for equality in distribution.

Therefore, using results from (2) and (6), model (1) can be written as:

$$\boldsymbol{y} = \mu\boldsymbol{1} + \boldsymbol{Z}_E\boldsymbol{\beta}_E + \boldsymbol{Z}_g\boldsymbol{L}\boldsymbol{u}_1^* + \boldsymbol{L}_{ge}\boldsymbol{u}_2^* + \boldsymbol{e} \quad (7)$$

Models (1) and (7) are equivalent, but model (7) has at least two advantages over model (1): (i) it avoids many matrix products, and (ii) it can be implemented relatively easily using the well known Gibbs sampler (Geman and Geman, 1984) in the Bayesian framework.

Model 2: G×E interaction using molecular markers

Let $\boldsymbol{W}$ be a $g \times p$ matrix of standardized markers, where g is the number of genotyped individuals and p is the number of markers; let $\boldsymbol{G} = \boldsymbol{WW}'/p$ be the genomic relationship matrix (López-Cruz et al., 2015). A model similar to (1) can be obtained by replacing $\boldsymbol{A}$ with $\boldsymbol{G}$.

Model 3: G×E interaction using molecular markers and pedigree (single-step approach)

In this model, the information for genotyped and non-genotyped individuals is combined using the approach proposed by Legarra et al. (2009) and Aguilar et al. (2010). A relationship matrix that includes full pedigree and genomic information is given by:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{A}_{nn} + \boldsymbol{A}'_{gn}\boldsymbol{A}_{gg}^{-1}\left(\boldsymbol{G}_a - \boldsymbol{A}_{gg}\right)\boldsymbol{A}_{gg}^{-1}\boldsymbol{A}_{gn} & \boldsymbol{A}'_{gn}\boldsymbol{A}_{gg}^{-1}\boldsymbol{G}_a \\ \boldsymbol{G}_a\boldsymbol{A}_{gg}^{-1}\boldsymbol{A}_{gn} & \boldsymbol{G}_a \end{bmatrix},$$

where the matrix is divided based on whether the individuals have been genotyped or not. Submatrices $\boldsymbol{A}_{gg}$, $\boldsymbol{A}_{nn}$ and $\boldsymbol{A}_{gn}$ are submatrices of $\boldsymbol{A}$ containing relationships among genotyped individuals, among non-genotyped individuals and between genotyped and non-genotyped individuals, respectively (Legarra et al., 2009; Christensen et al., 2012). $\boldsymbol{G}_a$ is an adjusted relationship matrix obtained from the genomic relationship matrix given by López-Cruz et al. (2015), i.e., $\boldsymbol{G} = \boldsymbol{WW}'/p$ and $\boldsymbol{A}_{gg}$, that is:

$$\boldsymbol{G}_a = \beta\boldsymbol{G} + \alpha,$$

where β and α are obtained by solving the following system of equations:

$$Avg\left(diag(\boldsymbol{G})\right)\beta + \alpha = Avg\left(diag\left(\boldsymbol{A}_{gg}\right)\right),$$

$$Avg(\boldsymbol{G})\beta + \alpha = Avg(\boldsymbol{A}_{gg}),$$

where $\boldsymbol{G}_a$ is a rescaled matrix such that: (i) the average of the diagonal elements is equal to the average of the diagonal elements of $\boldsymbol{A}_{gg}$, and (ii) the average of all the elements is equal to the average elements of $\boldsymbol{A}_{gg}$. See Christensen et al. (2012) for further details. Note that in this formulation based on $\boldsymbol{H}$ (and not its inverse), $\boldsymbol{H}$ does not need to be full rank.

Appendix B shows the R code that allows us to build the $\boldsymbol{H}$ matrix. A parametric G×E interaction model takes into account the effect of E environments, the main effect of genotypes and the interaction between genotypes and the environment. A model that uses information

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

obtained from markers and pedigree can be obtained by replacing the $\boldsymbol{A}$ matrix in model (1) with the $\boldsymbol{H}$ matrix described above.

Model without G×E interaction

Note that models that do not include the G×E term can be derived from models 1-3 just by removing the corresponding random G×E term. For example, by removing the term $\boldsymbol{u}_2$ representing the effect of G×E from model (1), it becomes

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Z}_E\boldsymbol{\beta}_E + \boldsymbol{Z}_g\boldsymbol{u}_1 + \boldsymbol{e}$$

In this case, the resulting models are equivalent to the across-environment GBLUP model of López-Cruz et al. (2015). We include models without the G×E term in order to compare the prediction accuracy of models with and without G×E interaction. The single-environment model was not included because all the wheat lines included in the prediction of South Asian environments had complete pedigree and markers; thus developing matrix $\boldsymbol{H}$ for the single-step model does not make sense.

Assessing the model's predictive ability

The main interest of breeders is to predict the performance of non-evaluated lines in South Asian sites (Jalbapur, Ludhiana, Pusa in India; Faisalabad in Pakistan; and Jamalpur in Bangladesh). To mimic that situation, we designed a cross-validation scheme where we fitted the G×E models (models 1-3) as well as models without G×E using as the training set all available records under Drought, Late Heat, Optimal and Severe Drought conditions obtained in Cd. Obregon (Mexico), and 20% of available records in each of the South Asian sites assigned at random. In the prediction process, 80% of lines in the corresponding sites in the South Asian countries (India, Pakistan, and Bangladesh) were predicted using the rest of the records. A total of 20 random partitions (such as the ones described above) were generated.

The models' predictive ability was compared by using Pearson's correlation coefficient. The models that use the $\boldsymbol{A}$ and $\boldsymbol{H}$ matrices included the phenotypic information of the 58,798 wheat lines, whereas the model that is based on markers only included information for 29,484 wheat lines that correspond to the individuals that were genotyped. The genotyped individuals are a subset of the individuals with pedigree information; therefore, lines in the testing set have pedigree and marker information. The numbers of individuals in the testing sets in South Asian

sites are shown in Table 3, so in each random partition, the same individuals are predicted with three different models based on the ***A***, ***G*** and ***H*** matrices.

Software

The models described above were fitted using a modified version of the BGLR package (de los Campos and Pérez-Rodríguez, 2015). The package was modified in order to accept as inputs big.matrix objects created using the bigmemory package (Kane et al., 2013). The bigmemory package was used to handle the huge matrices that have to be used during the analysis and also to take advantage of what in computer science is known as *shared memory*. Once loaded into RAM memory, the data can be accessed from several processors, making it possible to perform a cross-validation relatively easily.

Data availability

The complete phenotypic and marker data can be found at http://genomics.cimmyt.org/wheat_50k/PG/.

RESULTS

Descriptive statistics

Figure 1 shows a boxplot of grain yield by location and median yield per location. From the plot it can be seen that the Optimal conditions had the highest grain yield, while the Late Heat and Severe Drought conditions had the worst grain yield. Yields in South Asian environments, especially in Pakistan and Bangladesh, were usually lower than in Mexican environments. Table 4 shows the number of lines evaluated in each environment, and lines in common between pairs of environments. It also shows sample correlations for grain yield for each pair of environments. The number of lines evaluated in common between pairs of environments ranged from 537 to 4,735. The phenotypic sample correlation ranged from -0.05 to 0.53, which suggests large G×E.

Figure 2 displays the distribution of the diagonal entries for matrices ***A***, ***H*** and ***G***. Note that in the ***A*** matrix, the diagonal entries are around ~1.5; in this case, $a(i,i) = 1 + F_i$, where F_i is the inbreeding coefficient of the *i*-th individual. The diagonal entries of the ***G*** matrix are around 1.0, reflecting the fact that the markers were centered and standardized. The diagonal

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

entries of the ***H*** matrix are around 1.5, which stems from the standardization of ***G*** to be on the same scale as ***A***.

Prediction accuracy of models without G×E

Table 4 shows the phenotypic correlations between pairs of environments. For example, the phenotypic correlation of the 4062 common wheat lines between environments B5I_OBR and B2I_OBR is 0.156, whereas the phenotypic correlation of the 1537 common wheat lines in B5I_OBR and STN_PUS (Pusa, India) is 0.210. In general, phenotypic correlations were not high, ranging from -0.051 to 0.481.

Table 5 shows the average Pearson's correlations between observed and predicted phenotypes and their corresponding standard deviation for the model without including G×E. The average correlations come from 20 random partitions with all the data records available in Mexico and 20% of the data available in South Asia. Note that these are the predictions of 80% of the entries included in the six South Asia environments. Prediction accuracies are relatively low, with those based on pedigree being slightly higher than those based on markers or on both pedigree and markers.

Prediction accuracy of G×E models

Table 6 shows the average of Pearson's correlations between observed and predicted phenotypes and its corresponding standard obtained using the same cross-validation scheme described above, but now including the G×E term. The predictive ability of models based on Pedigree, Markers and Pedigree + Markers is about the same, with pedigree prediction accuracy being higher than genomic and pedigree + genomic prediction accuracy in four environments (DLP_FAS, STN_JAM, STN_JBL, and STN_PUS). Ludhiana and Faisalabad under standard management conditions (0.3785, 0.2455, respectively) were the best predictive models for the genomic and pedigree + genomic model, respectively.

Figure 3(a)-(c) shows scatterplots of predictive correlations for each of 20 cross-validations across the six environments in South Asia. Figure 3a depicts the correlations between predicted values based on markers (***G***) versus those based on matrix ***H*** and shows that prediction accuracy based on ***G*** was superior to that obtained based on ***H***. Figure 3b

displays the correlation based on markers (***G***) versus that obtained based on pedigree (***A***), where the prediction based on pedigree seems slightly better than that based on ***H*** (Fig. 3c).

Table 7 shows the percentage change in prediction accuracy of models with and without G×E. The % change was calculated as $(r_{G\times E} - r_{no\ G\times E})/r_{no\ G\times E} \times 100$, where $r_{G\times E}$ is the Pearson's correlation for a model with the G×E term and $r_{no\ G\times E}$ is the Pearson's correlation for a model without the G×E term. From results in Table 7, it is clear that models that include the G×E term predict better than those that do not include G×E. For example, the G×E model using matrix ***H*** gave a 66% increase in prediction accuracy compared to the model using matrix ***H*** but without G×E.

Figure 4 presents a bar plot of correlations for each predicted environment in South Asia using the ***H*** matrix. Bars in black represent the mean of the weighted phenotypic correlation of a given environment and the rest of the environments in Table 4. The phenotypic correlation for environment *j* in South Asia can be obtained as follows: $r_j = \sum_{k\neq j} \frac{n_{jk}}{n_j} r_{jk}$, where $j = 1, \ldots, 6$ (environments in South Asia) and $k = 1, \ldots, 11$ represents the set of environments in South Asia and Mexico excluding environment *j*, n_{jk} corresponds to the number of lines in common between environments *j* and *k*, $n_j = \sum_k n_{jk}$ and r_{jk} is the phenotypic correlation between environments *j* and *k*. As an example, Table 8 presents the information needed to compute the weighted correlation for environment DLP_FAS; the columns present the information needed to compute the weighted correlation (note that this information was obtained from Table 4). The rest of the correlations were obtained using the approach described above. The gray bars represent the means of the correlations between observed and predicted values obtained from cross-validations. Note that in general the G×E models give good predictions, usually better than the phenotypic correlations. Although we are predicting 80% of the records in each of the South Asian environments, the correlations are higher than the phenotypical correlations between a given environment and the rest of the environments.

DISCUSSION

In wheat breeding, the cost of genotyping thousands of plants in segregating populations or in advanced generations makes the application of GS unfeasible. One possibility for solving this

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

problem would be to augment the numerical relationship matrix (***A***) of all individuals with the genomic relationship matrix (***G***) of the genotyped individuals and perform predictions based on the resulting complete ***H***, which will allow performing prediction of non-genotyped individuals in the testing set. Augmenting matrix ***A*** by using only a fraction of the genotyped individuals would reduce genotyping costs. Furthermore, as described by Christensen et al. (2012), the single-step method allows the genomic relationship matrix of genotyped individuals to be extended using pedigree information to a combined relationship matrix ***H*** of all individual plants or lines. This allows using all phenotypic data and not only those phenotypes that have pedigree and marker information; this extra phenotypic information should also enhance prediction accuracy. This makes the models and methods developed by Misztal et al. (2009), Legarra et al. (2009) and Aguilar et al. (2010; 2011) very attractive for predicting unobserved and non-genotyped plants.

In a recent article, Fernando et al. (2014) proposed a single-step Bayesian regression strategy that allows using all genotyped and non-genotyped individuals by means of imputed marker covariates for non-genotyped individuals. The advantage of the Bayesian approach over the single-step BLUP is that it does not require computing the inverse of ***G***. However, this model has not yet been applied to realistic datasets.

The single-step approach of Misztal et al. (2009), Legarra et al. (2009) and Aguilar et al. (2010; 2011) was used in dairy cattle studies and first applied to plant breeding data by Ashraf et al. (2016) in a set of 1, 176 genotyped CIMMYT wheat lines and 11,131 non-genotyped wheat lines tested in five environments in Cd. Obregon, Mexico, during the 2012-2013 cycle. The authors developed optimized weighting factors for matrix ***H*** and applied a multivariate method for assessing G×E; they found that the prediction accuracy of the single-step ***H*** matrix was higher than the accuracies achieved using the ***A*** and ***G*** matrices. The present study used seven selection cycles of CIMMYT wheat breeding with a total of 58,798 wheat lines evaluated in Cd. Obregon and predicts several wheat lines in South Asian environments (India, Pakistan, and Bangladesh).

Genomic prediction accuracy for models with and without G×E

From the results in Tables 5, 6 and 7, it is clear that models that include the G×E term predict the environments in South Asia better than models that do not include the G×E term. The gain in

prediction accuracy of models that include G×E ranges from 16 to 90% with an average of 40%. However, models that do not incorporate G×E but use pedigree or high density molecular markers, or both, are still superior in terms of prediction accuracy than those that use only phenotypic data.

Genomic prediction accuracy versus phenotypic prediction accuracy of G×E models

In this study, we assessed the prediction accuracy of a large number of wheat lines evaluated in several environments and years in Cd. Obregon, Mexico, and predicted lines in several South Asian environments. For Ludhiana, Pusa, and Jabalpur, about 1227 wheat lines were predicted based on the performance of these lines in six environments in Cd. Obregon, plus the performance of about 57,000 wheat lines related to those to be predicted (1227) and evaluated in previous years in Cd. Obregon, Mexico.

Prediction accuracy was the correlation between the predicted values of the lines in Cd. Obregon plus a low proportion of them (20%) in six environments in South Asia using three G×E models (***A***, ***G***, and ***H***) with the observed values of 80% of the lines in the six environments in South Asia (that were not phenotyped). The correlations for all the environments were around 0.25-0.27, except for Ludhiana in India, which showed higher prediction accuracy (0.36-0.37). These genomic prediction accuracies were higher than the prediction accuracies computed from the common phenotypic correlations between all pairs of environments. These results indicated that the prediction accuracy with which breeders make selections in Cd. Obregon, Mexico, is lower than the accuracy they could obtain by performing genomic selection and prediction. Although wheat breeders expect that lines selected in Cd. Obregon will perform well in South Asian environments, the results of this study should prompt them to increase research on genomic selection in Cd. Obregon (a very stable site with high radiation) of candidates for selection that will perform well in several environments in different South Asian countries (India, Pakistan, and Bangladesh).

The prediction accuracy of models with ***A***, ***G***, and ***H*** for models with or without G×E did not change much. This is an important result that allows, through the use of matrix ***H***, using all phenotypic data to predict the genetic values of the unobserved wheat lines, thereby avoiding having to use only a subset of the phenotypes of those lines with pedigree and another subset of

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

phenotype data of lines with only marker data. Also the single-step method for computing $\boldsymbol{H}$ allows the inclusion of both components of the breeding value to be predicted, the parental average or between-family variability captured by the pedigree and Mendelian sample component (or with family variability) accounted by markers.

Big data used to derive pedigree and combine it with markers into the single-step prediction method with a G×E model

So far, no studies using plant breeding data on more than 50,000 lines have been reported in the genomic selection literature. This is the first study that shows that large training populations can provide genomic predictions that are more precise than phenotypic predictions. This is the first time that the theory used to develop and implement the pedigree system for such a large number of lines is reported in plant breeding. Although the models used for prediction are now well known, from the computational and statistical points of view, it is necessary to develop algorithms and data structures that allow researchers to handle the data and fit the models efficiently.

In this study, we used the G×E reaction norm model on a large data set in conjunction with pedigree, markers, or both, in genomic selection and prediction. We compared models including and excluding G×E and also in the genomic prediction literature there are plenty of examples where including those interactions significantly improved the prediction accuracy of untested individuals. The single-step method that combines the use of pedigree and markers through matrix $\boldsymbol{H}$ allows using all the available information. Also, the reaction norm G×E model allows borrowing information among positively correlated environments, although the predictive power of the model was similar to that of the model that includes markers only. Ashraf et al. (2016) used the single-step $\boldsymbol{H}$ approach on a set of only 11,131 non-genotyped and 1176 genotyped wheat lines.

Animal breeders make extensive use of the fact that relationship matrix $\boldsymbol{A}$ has a very sparse inverse that can be computed directly from the pedigree, if all individuals (including those with no phenotype) are included (Henderson, 1976, 1977). This results in a sparse $\boldsymbol{H}^{-1}$ structure

as well (Aguilar et al., 2010; Christensen and Lund, 2010), with a storage cost quadratic in the number of genotyped individuals and only linear in the number of non-genotyped individuals. These sparse inverses exist for any level of autopolyploid species (Kerr et al., 2012) and could potentially be used for prediction with large data sets. However, this would preclude the use of Cholesky decomposition used in (7).

CONCLUSIONS

This study shows how to solve statistical and computational challenges when incorporating and combining highly dimensional pedigree and genomic matrices into a single-step model for predicting unobserved individuals in other environments. We found that genomic prediction of genotyped and non-genotyped wheat lines produces higher prediction accuracy than that of lines predicted based on phenotypic data. The results provide evidence that the single-step approach that combines pedigree and marker information is useful for reducing genotyping costs while maintaining the prediction accuracy of unobserved individuals at relatively intermediate levels. The incorporation of G×E models using a combination of pedigree and genomic information is another way of increasing the prediction accuracy of unobserved candidates for selection and offers plant breeders an important alternative for predicting germplasm evaluated under different environmental conditions.

ACKNOWLEDGMENTS

We are grateful to field and lab workers of CIMMYT's Global Wheat Program and to national program researchers who collected the data used in this study. This work was funded by USAID Feed the Future Innovation Lab for applied Wheat Genomics with headquarters at Kansas State University in collaboration with CIMMYT and their partners in South Asia, specifically in India and Pakistan. We also thank the Bill and Melinda Gates Foundation for providing financial support for genotyping the materials used here.

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

REFERENCES

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93, 743–752.

Aguilar, I., Misztal, I., Legarra, A., and Tsuruta, S. 2011. Efficient computations of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128(6): 422-8. doi: 10.1111/j.1439-0388.2010.00912.x.

Ashraf, B., Edriss, V., Akdemir, D., Autrique, E., Bonnett, D., Crossa, J., Janss, L., Singh, R., and Jannink, J-L. 2016. Genomic prediction using phenotypes from pedigree lines with no markers. Crop Sci. 56: 957-964.

Bates, D., and Vazquez, A. 2009. pedigreemm: Pedigree-based mixed-effects models. Available at http://CRAN.R-project. org/package=pedigreemm (verified 8 July, 2016).

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. 2012. Genomic prediction of breeding values when modeling Genotype × Environment interaction using pedigree and dense molecular markers. Crop Sci. 52(2):707–719. doi:10.2135/crop- sci2011.06.0299.

Christensen, O. F., and Lund, M. S. 2010. Genomic prediction when some animals are not genotyped. Genetic Selection and Evolution 42: 2. Doi 10.1186/1297-9686-42-2.

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T., and Su, G. 2012. Single-step methods for genomic evaluation in pigs. Animal 6:1565–1571.

Crossa, J., de los Campos, G., Pérez-Rodríguez, P., Gianola, D., Burgueño, J., et al., 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Crossa, J., Pérez-Rodríguez, P., de los Campos, G., Mahuku, G., Dreisigacker, S., and Magorokosho, C. 2011. Genomic selection and prediction in plant breeding. J of Crop Improvement 25(3):239–61.

de los Campos, G., and Pérez-Rodríguez, P. 2015. BGLR: Bayesian Generalized Linear Regression. R package version 1.0.4, http://CRAN.R-project.org/package=BGLR.

de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., and Crossa, J. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92: 295–308.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., et al., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375–385.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E., et al., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6: e19379.

Fernando, R. L., Deckers, J. C., and Garrick, D. J. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole genome analyses. Genetics Selecion and Evolution 46(1), 50. http://doi.org/10.1186/1297-9686-46-50.

Geman, S., and Geman, D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6(6): 721–741.

González-Camacho, J. M., de los Campos, G., Pérez-Rodríguez, P., Gianola, D., et al. 2012. Genome-enabled prediction of genetic values using radial basis function. Theor. Appl. Genet. 125:759–771. doi:10.1007/s00122-012-1868-9.

Harville, D. A., and Callanan, T. P. 1989. Computational aspects of likelihood-based inference for variance components. Pages 136-176 in Advances in Statistical Methods for Genetic Improvement of Livestock. D. Gianola and K. Hammond, eds. Springer-Verlag, Berlin, Germany.

Henderson, C. R. 1975. Rapid method for computing the Inverse of a Relationship Matrix. Journal of Dairy Science 58(11): 1727-1730.

Henderson, C. R. 1976. A simple method for computing the inverse of the numerator relationship matrix used in prediction of breeding values. Biometrics 32:69-83.

Henderson, C. R. 1977. Best linear unbiased prediction of breeding values not in the model for records. Journal of Dairy Science 60(5), 783-787.

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. 2012. Genomic selection in plant breeding: a comparison of models. Crop Sci. 52:146–160.

Hickey, J. M., Crossa, J., Babu, R., and de los Campos, G. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52(2):654-663.

Jarquín, D., Crossa, J., Lacaze, X., Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., et al. 2014.

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics 127 (3): 595–607. doi:10.1007/s00122-013-2243-1.

Kane, M. J., Emerson, J., and Weston, S. 2013. Scalable Strategies for Computing with Massive Data. Journal of Statistical Software 55(14), 1-19. URL http://www.jstatsoft.org/v55/i14/.

Kerr, R. J., Li, L., Tier, B., Dutkowski, G. W., and McRae, T. A. 2012. Use of the numerator relationship matrix in genetic analysis of autopolyploid species. Theoretical and Applied Genetics 124(7), 1271-1282.

Legarra, A., Aguilar, I., and Misztal, I. 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92, 4656–4663.

López-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J-L., Singh, R.P., Autrique, E., and de los Campos, G. 2015. Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3: Genes, Genomes, Genet. 5:569–582. doi:10.1534/g3.114.016097.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. 2001. Prediction of total genetic values using genome-wide dense marker maps. Genetics 157: 1819–1829.

Misztal, I., Legarra, A., and Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648–4655.

Nejati-Javaremi, A., Smith, C., and Gibson, J. P. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. 75:1738–1745.

Pérez Rodríguez, P., Crossa, J., Bondalapati, K., De Meyer, G., Pita, F., and de los Campos, G. 2015. A pedigree reaction norm model for prediction of cotton (*Gossypium* sp.) yield in multi-environment trials. Crop Science 55, 1143-1151.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manes, Y., and Dreisigacker, S. 2012. Comparison between linear and non-parametric models for genome-enabled prediction in wheat. G3: Genes, Genomes, Genet. 2:1595–1605.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (verified 8 July 2016).

Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., and Foley, J. A. 2012. Recent patterns of crop yield growth and stagnation. Nat. Commun. 3: 1293.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., et al., 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44: 217-220. doi:10.1038/ng.1033.

VanRaden P.M. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91:4414–4423.

Weber, V. S., Melchinger, A. E., Magorokosho, C., Makumbi, D., Bänziger, M., and Atlin, G. N. 2012. Efficiency of managed-stress screening of elite maize hybrids under drought and low nitrogen for yield under rainfed conditions in Southern Africa. Crop Sci. 52: 1011-1020.

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

Table 1: Description of the conditions under which the 58,798 wheat lines were evaluated in different environments.

Description	Field Management Conditions
Standard management conditions (STN)	Optimal
Delayed planting (DLP)	Late Heat
Bed planting and 5 irrigations (B5I)	Optimal
Bed planting and 2 irrigations (B2I)	Drought
Zero till, bed planting and 5 irrigations (Z5I)	Optimal
Zero till, bed planting and 2 irrigations (Z2I)	Drought
Melgas flat planting and 5 irrigations (MEL)	Optimal
Melgas flat planting and drip irrigation (DRM)	Severe Drought
Bed planting and drip irrigation (BDR)	Severe Drought
Early heat (EHT)	Early Heat
Late heat (LHT)	Late Heat

Table 2: Number of lines evaluated in different environments during 2009-2016 by the GWP.

Environment	Number of lines evaluated
B5I_OBR	56,964
B2I_OBR	4,063
DRB_OBR	5,913
EHT_OBR	2,188
LHT_OBR	4,736
MEL_OBR	4,735
DLP_FAS	1,547
STN_FAS	1,547
STN_JAM	537
STN_JBL	1,548
STN_LDH	1,548
STN_PUS	1,548

FAS=Faisalabad, Pakistan; JAM=Jamalpur, Bangladesh; JBL=Jabalpur, India; LDH=Ludhiana, India, OBR=Cd, Obregon, Mexico; PUS=Pusa, India. Standard management conditions (STN); Delayed planting (DLP); Bed planting and 5 irrigations (B5I); Bed planting and 2 irrigations (B2I); Zero till, bed planting and 5 irrigations (Z5I); Zero till, bed planting and 2 irrigations (Z2I); Melgas flat planting (MEL); Drip irrigation to impose drought in flat (DRM); Drip irrigation to impose drought, beds (DRB); Early heat (EHT); Late heat (LHT).

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

Table 3: Number of individuals in the testing set in South Asian sites.

Environment	**Number of individuals in the testing set**
DLP_FAS	1237
STN_FAS	1237
STN_JAM	429
STN_JBL	1238
STN_LDH	1238
STN_PUS	1238

Table 4: Number of genotypes (diagonal and upper triangular) and sample phenotypic correlations (lower triangular) by environment.

Env	B5I_OBR	B2I_OBR	DRB_OBR	EHT_OBR	LHT_OBR	MEL_OBR	DLP_FAS	STN_FAS	STN_JAM	STN_JBL	STN_LDH	STN_PUS
B5I_OBR	56964	4062	4090	2187	4734	4735	1537	1537	532	1537	1537	1537
B2I_OBR	0.156	4063	4063	2186	4063	4062	1515	1515	530	1515	1515	1515
DRB_OBR	-0.050	0.534	5913	2186	4091	4090	1535	1535	530	1535	1535	1535
EHT_OBR	0.479	0.186	-0.051	2188	2187	2187	1062	1062	532	1062	1062	1062
LHT_OBR	0.203	0.262	0.167	0.199	4736	4734	1537	1537	532	1537	1537	1537
MEL_OBR	0.370	0.238	0.117	0.354	0.169	4735	1537	1537	532	1537	1537	1537
DLP_FAS	0.154	0.094	0.111	0.131	0.067	0.174	1547	1547	537	1547	1547	1547
STN_FAS	0.124	0.120	0.167	0.009	0.029	0.102	0.338	1547	537	1547	1547	1547
STN_JAM	0.228	0.146	0.130	0.160	0.079	0.113	0.170	0.206	537	537	537	537
STN_JBL	0.188	0.176	0.168	0.082	0.136	0.143	0.235	0.263	0.136	1548	1548	1548
STN_LDH	0.225	0.079	0.078	0.190	0.040	0.168	0.206	0.286	0.382	0.140	1548	1548
STN_PUS	0.210	0.137	0.099	0.117	0.025	0.173	0.280	0.241	0.481	0.255	0.222	1548

FAS=Faisalabad, Pakistan; JAM=Jamalpur, Bangladesh; JBL=Jabalpur, India; LDH=Ludhiana, India, OBR=Obregon, Mexico; PUS=Pusa, India. Standard management conditions (STN); Delayed planting (DLP); Bed planting and 5 irrigations (B5I); Bed planting and 2 irrigations (B2I); Zero till, bed planting and 5 irrigations (Z5I); Zero till, bed planting and 2 irrigations (Z2I); Melgas flat planting (MEL); Drip irrigation to impose drought in flat (DRM); Drip irrigation to impose drought, beds (DRB); Early heat (EHT); Late heat (LHT).

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

Table 5: Correlations between predicted and observed values obtained using the cross-validation where all the wheat lines from Cd. Obregon, Mexico, plus 20% of the wheat lines in each of the environments in India, Pakistan, and Bangladesh were used in the training set to predict 80% of the lines in the corresponding environments in India, Pakistan, and Bangladesh. The highest correlations in each environment are in boldface.

	Model without G×E		
Environment	Pedigree (***A***)	Markers (***G***)	Pedigree + Markers (***H***)
DLP_FAS	**0.2113** (0.0304)	0.1716 (0.0104)	0.1834 (0.0135)
STN_FAS	**0.1611** (0.0181)	0.1235 (0.0129)	0.1455 (0.0120)
STN_JAM	**0.2448** (0.0251)	0.1861 (0.0189)	0.1992 (0.0213)
STN_JBL	**0.2480** (0.0184)	0.1928 (0.0154)	0.2075 (0.0163)
STN_LDH	**0.2554** (0.0158)	0.2472 (0.0104)	0.2477 (0.0094)
STN_PUS	**0.2361** (0.0143)	0.1989 (0.0112)	0.2117 (0.0107)

FAS=Faisalabad, Pakistan; JAM=Jamalpur, Bangladesh; JBL=Jabalpur, India; LDH=Ludhiana, India, OBR=Obregon, Mexico; PUS=Pusa, India. Standard management conditions (STN), and delayed planting conditions (DLP).

Table 6: Correlations between predicted and observed values obtained using the cross-validation where all the wheat lines from Cd. Obregon, Mexico, plus 20% of the wheat lines in sites in India, Pakistan, and Bangladesh were used in the training set to predict 80% of the lines in the corresponding sites in India, Pakistan, and Bangladesh. The highest correlations in each environment are in boldface.

	G×E model		
Environment	Pedigree (***A***)	Markers (***G***)	Pedigree + Markers (***H***)
DLP_FAS	**0.2462** (0.0294)	0.2327 (0.0132)	0.2345 (0.0123)
STN_FAS	0.2360 (0.0227)	0.2414 (0.0180)	**0.2455** (0.0175)
STN_JAM	**0.2942** (0.0414)	0.2681 (0.0293)	0.2656 (0.0309)
STN_JBL	**0.2921** (0.0183)	0.2741 (0.0163)	0.2739 (0.0165)
STN_LDH	0.3699 (0.0109)	**0.3785** (0.0157)	0.3651 (0.0155)
STN_PUS	**0.2842** (0.0175)	0.2622 (0.0191)	0.2684 (0.0185)

FAS=Faisalabad, Pakistan; JAM=Jamalpur, Bangladesh; JBL=Jabalpur, India; LDH=Ludhiana, India, OBR=Obregon, Mexico; PUS=Pusa, India. Standard management conditions (STN), and delayed planting conditions (DLP).

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

Table 7. Comparing the predictive ability of models with and without G×E.

Environment	% Change		
	Pedigree (***A***)	Markers (***G***)	Pedigree + Markers (***H***)
DLP_FAS	16.52	35.61	26.88
STN_FAS	46.49	95.47	65.91
STN_JAM	20.18	44.06	34.59
STN_JBL	17.78	42.17	32.10
STN_LDH	44.83	53.11	52.81
STN_PUS	20.37	31.83	23.85

% Change = $(r_{G\times E} - r_{no\ G\times E,C})/r_{no\ G\times E} \times 100$, where $r_{G\times E}$ is the Pearson's correlation for a model with the G×E term and $r_{no\ G\times E}$ is the Pearson's correlation for a model without the G×E term.

FAS=Faisalabad, Pakistan; JAM=Jamalpur, Bangladesh; JBL=Jabalpur, India; LDH=Ludhiana, India, OBR=Obregon, Mexico; PUS=Pusa, India. Standard management conditions (STN), and delayed planting conditions (DLP).

Table 8: Phenotypic correlations and numbers of lines in common between DLP_FAS and the rest of the environments in Mexico and South Asia.

j	Env. in South Asia	k	Other environments	r_{jk}	n_{jk}	n_{jk}/n_j	$n_{jk}/n_j\, r_{jk}$
1	DLP_FAS	1	B5I_OBR	0.154	1537	0.099	0.015
1	DLP_FAS	2	B2I_OBR	0.094	1515	0.098	0.009
1	DLP_FAS	3	DRB_OBR	0.111	1535	0.099	0.011
1	DLP_FAS	4	EHT_OBR	0.131	1062	0.069	0.009
1	DLP_FAS	5	LHT_OBR	0.067	1537	0.099	0.006
1	DLP_FAS	6	MEL_OBR	0.174	1537	0.099	0.017
1	DLP_FAS	7	STN_FAS	0.338	1547	0.100	0.033
1	DLP_FAS	8	STN_JAM	0.170	537	0.035	0.005
1	DLP_FAS	9	STN_JBL	0.235	1547	0.100	0.023
1	DLP_FAS	10	STN_LDH	0.206	1547	0.100	0.020
1	DLP_FAS	11	STN_PUS	0.280	1547	0.100	0.028
				$n_1 =$	15448		$r_1 = 0.18$

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

APPENDIX

R script to obtain and store relationship matrix *A*

This script computes relationship matrix ***A***.

Inputs:

1) A text file with pedigree information for the individuals that we are interested in. The file should have 3 columns separated by tabs, ID (Id of the individual), Sire and Dam.
2) A text file with the individuals that we are interested in.

Output: The relationship matrix.

To speed up the computations, we used dense partitioned matrixes and linked R against OpenBLAS (http://www.openblas.net). At the end of the process, the relationship matrix was also stored as a partitioned matrix on the hard disk in binary R format (RData). Below we detail the steps used to build the matrix.

Step 1: Read the data and compute the relationship matrix from the pedigree information

```
#Clean workspace
rm(list=ls())


#Load
library(pedigreemm)


#Read the pedigree file
a=read.csv("pedigree/RAVI_58K_GIDS_PROGEN.csv",header=TRUE)
a=a[,c(1:3)]
a=a[a[,1]!=0 & a[,2]!=0,]

colnames(a)=c("SIRE","DAM","ID")
a=a[!duplicated(a),]

cat("nrow=",nrow(a),"\n")
cat("selfing=",sum(a[,1]==a[,2]),"\n")

#Read the ids of individuals with phenotypic records
ids=scan("GIDsForUSAIDprediction_20160406.csv")
```

```
ids=as.character(ids)

pede=editPed(sire=a$SIRE,dam=a$DAM,label=a$ID,verbose=TRUE)
ped=with(pede, pedigree(label=label, sire=sire, dam=dam))
```

Now use the **relfactor** function for the pedigree, that is

$$\boldsymbol{A_{full} = X_{full}' X_{full}}$$

where $\boldsymbol{X_{full}}$ is an upper triangular, sparse (right) Cholesky factor of the relationship matrix. In this case, $\boldsymbol{X_{full}}$ is a matrix with n=177,376 rows and the same number of columns. The code for obtaining the relfactor is given below.

```
Xfull=relfactor(ped)
```

We do not need $\boldsymbol{A_{full}}$**;** we just need a subset of this matrix with the 58,798 individuals so we can take a subset of 58,798 columns from $\boldsymbol{X_{full}}$. The columns correspond to the individuals that we are interested in. Let $\boldsymbol{X}$ be the resulting matrix; then

$$\boldsymbol{A = X'X}$$

where **X** has *n*=177,376 rows and *p*=58,798 columns. The R code for obtaining this matrix is shown below.

```
index=ped@label%in%ids
X=Xfull[,index]
```

Step 2: Partition the relationship factor

Since $\boldsymbol{X}$ is a huge matrix, it is very difficult to obtain $\boldsymbol{A}$ directly; furthermore, since X is sparse, the product cannot be parallelized easily. So we partitioned $\boldsymbol{X}$ into several sub-matrices and saved the sub-matrices as binary files that can later be retrieved in order to perform the product.

For example:

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{11} & \boldsymbol{X}_{12} \\ \boldsymbol{X}_{21} & \boldsymbol{X}_{22} \\ \boldsymbol{X}_{31} & \boldsymbol{X}_{32} \\ \boldsymbol{X}_{41} & \boldsymbol{X}_{42} \\ \boldsymbol{X}_{51} & \boldsymbol{X}_{52} \end{pmatrix}; \; \boldsymbol{X}' = \begin{pmatrix} \boldsymbol{X}'_{11} & \boldsymbol{X}'_{21} & \boldsymbol{X}'_{31} & \boldsymbol{X}'_{41} & \boldsymbol{X}'_{51} \\ \boldsymbol{X}'_{12} & \boldsymbol{X}'_{22} & \boldsymbol{X}'_{32} & \boldsymbol{X}'_{42} & \boldsymbol{X}'_{52} \end{pmatrix}$$

where $\boldsymbol{X}_{ij}$ is a sub-matrix obtained from $\boldsymbol{X}$.

The R code below was used to partition matrix $\boldsymbol{X}$ into 5 sub-matrices and save the results to binary files.

```
n_submatrix=5
n=nrow(X)
p=ncol(X)

to_row=0;
delta=as.integer(n/n_submatrix);

for(i in 1:n_submatrix)
{
    from_row=to_row+1;
    to_row=delta*i;
    if(i==n_submatrix) to_row=n;

    #Another slice for X
    for(j in 1:2)
    {
      if(j==1)
      {
        from_column=1
        to_column=29401
      }else{
        from_column=29402
        to_column=p
      }
     cat("**********************\n")
     cat("Submatrix: ",i," ",j,"\n");
     cat("from_row: ",from_row,"\n");
```

```
      cat("to_row: ",to_row,"\n");
      cat("from_column: ",from_column,"\n");

      #Conventional matrix object so that we can use
      #optimized dense matrix products
      Xij=as.matrix(X[from_row:to_row,from_column:to_column])
      save(Xij,file=paste("X_",i,j,".RData",sep=""))

    }
}
```

Step 3: Compute the relationship matrix using the partitioned matrices from step 2

Given the partition of the relationship factor, we can compute the A matrix as follows:

$$\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$$

where

$$\boldsymbol{A}_{11} = \boldsymbol{X}'_{11}\boldsymbol{X}_{11} + \boldsymbol{X}'_{21}\boldsymbol{X}_{21} + \boldsymbol{X}'_{31}\boldsymbol{X}_{31} + \boldsymbol{X}'_{41}\boldsymbol{X}_{41} + \boldsymbol{X}'_{51}\boldsymbol{X}_{51}$$
$$\boldsymbol{A}_{22} = \boldsymbol{X}'_{12}\boldsymbol{X}_{12} + \boldsymbol{X}'_{22}\boldsymbol{X}_{22} + \boldsymbol{X}'_{32}\boldsymbol{X}_{32} + \boldsymbol{X}'_{42}\boldsymbol{X}_{42} + \boldsymbol{X}'_{52}\boldsymbol{X}_{52}$$
$$\boldsymbol{A}_{12} = \boldsymbol{X}'_{11}\boldsymbol{X}_{12} + \boldsymbol{X}'_{21}\boldsymbol{X}_{22} + \boldsymbol{X}'_{31}\boldsymbol{X}_{32} + \boldsymbol{X}'_{41}\boldsymbol{X}_{42} + \boldsymbol{X}'_{51}\boldsymbol{X}_{52}$$
$$\boldsymbol{A}_{21} = \boldsymbol{X}'_{12}\boldsymbol{X}_{11} + \boldsymbol{X}'_{22}\boldsymbol{X}_{21} + \boldsymbol{X}'_{32}\boldsymbol{X}_{31} + \boldsymbol{X}'_{42}\boldsymbol{X}_{41} + \boldsymbol{X}'_{52}\boldsymbol{X}_{51}$$

Note that now we need to perform several products of matrices. There are optimized libraries for that. For example, in R we can recompile the program so that we can use OpenBLAS. Details are given at the following links:

http://www.openblas.net/

http://www.rochester.edu/college/psc/thestarlab/help/moreclus/BLAS.pdf

We recompiled R-3.2.3 (http://r-project.org) in order to use OpenBLAS so it can perform matrix operations in parallel. The next fragment of code obtains the matrix $\boldsymbol{A}_{11}$ using the partitioned matrices.

```
rm(list=ls())
n_submatrix=5
A11=matrix(0,nrow=25000,ncol=25000)
```

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

```
for(i in 1:n_submatrix)
{
  cat("i=",i,"\n")
  load(paste("X_",i,"1.RData",sep=""))
  A11=A11+crossprod(Xij);
}
save(A11,file="A11.RData")
```

The rest of the matrices can be similarly obtained. With this approach and using 8 cores for the matrix product, we obtained the 58, 798 × 58, 798 matrix ***A*** in less than 3 hours in the CIMMYT-BSU server (which has 12 Intel (R) Xeon Cores @ 3.47 GHz and ~ 48 Gb of RAM).

R script to obtain the *H* matrix

The script presented below computes a relationship matrix ***H*** including full pedigree and genomic information (see equation 4 in Legarra et al., 2009). It adjusts the elements of genomic relationship matrix ***G***, so that the entries of relationship matrix ***A*** share the same scale (Christensen et al., 2012).

Inputs:

1) ***A*** and ***G*** matrices. The row and column names of both matrices include the IDs of the individuals.

Output:

1) ***H*** matrix.

```
#Clean workspace
rm(list=ls())

#Load A
load("../output/A11.RData")
load("../output/A12.RData")
load("../output/A21.RData")
load("../output/A22.RData")
```

```
A=rbind(cbind(A11,A12),
      cbind(A21,A22))

rm(A11,A12,A21,A22)

# read G and construct matrix of pedigree relationships of
# genotyped individuals, Agg (called A22 in Legarra et al., 2009 and A11 in OF Christensen notation)

#Read the genotypes (markers)
load("G80_42706_29489_correctedgid.RData")

#Center and scale the markers
X=scale(X,center=TRUE,scale=TRUE)

#Compute the genomic relationship matrix (López-Cruz et al., 2015)
G=tcrossprod(X)/ncol(X)

#Ids of genotyped individuals
genotyped=colnames(G)
cat("genotyped: ",length(genotyped),"\n")

#Ids of individuals with pedigree
inpedigree=colnames(A)
cat("inpedigree: ",length(inpedigree),"\n")

#Ids of individuals not genotyped
nongenotyped=setdiff(inpedigree,genotyped)
cat("in pedigree nongenotyped: ",length(nongenotyped),"\n")

genotypednotinpedigree=setdiff(genotyped,inpedigree)
cat("genotyped not in pedigree",length(genotypednotinpedigree),"\n")

genotypedinpedigree=intersect(genotyped,inpedigree)
cat("genotyped in pedigree",length(genotypedinpedigree),"\n")

# we have individuals with genotype that are NOT in matrix A
# we get rid of these individuals
G=G[genotypedinpedigree,genotypedinpedigree]
genotyped=genotypedinpedigree

#extract submatrix of A concerning genotyped individuals
Agg=matrix(NA,ncol(G),nrow(G))
Agg=A[genotyped,genotyped]

# now we need to make both matrices compatible. Use here Christensen et al. 2012 to make
# average inbreeding and average relationships compatible
# so that G <- a+bG
# O. F. Christensen, P. Madsen, B. Nielsen, T. Ostersen and G. Su (2012). Single-step methods
# for genomic evaluation in pigs. animal,6, pp 1565-1571 doi:10.1017/S1751731112000742

meanG=mean(G)
meandiagG=mean(diag(G))
meanAgg=mean(Agg)
meandiagAgg=mean(diag(Agg))
cat(meanG,meandiagG,meanAgg,meandiagAgg,"\n")
b=(meandiagAgg-meanAgg)/(meandiagG-meanG)
```

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

```
a=meandiagAgg-meandiagG*b
cat(a,b,"\n")

# a should be positive !!!
# modification to make G compatible
G=a+b*G

# invert Agg as we need it
Aggi=solve(Agg)

# a problem here is to divide A neatly between genotyped and not genotyped individuals.
# Usually we use sparse operators and this is easier.
# here I use the colnames and should be efficient

# ---------------------------------------- #
# option 1 to construct H not its inverse
# ---------------------------------------- #
# use expression (4) in Legarra-Aguilar-Misztal 2009
H=matrix(NA,ncol(A),nrow(A))
colnames(H)=colnames(A)
rownames(H)=rownames(A)
H[genotyped,genotyped]=G
H[nongenotyped,genotyped]=A[nongenotyped,genotyped]%*%(Aggi%*%G)
#H[genotyped,nongenotyped]=G%*%Aggi%*%A[genotyped,nongenotyped]
H[genotyped,nongenotyped]=t(H[nongenotyped,genotyped])
H[nongenotyped,nongenotyped]=A[nongenotyped,nongenotyped] +
                    A[nongenotyped,genotyped]%*%(Aggi%*%(G-
 Agg)%*%Aggi)%*%A[genotyped,nongenotyped]
cat(mean(diag(H)),mean(H),"\n")

# in principle H is (SEMI-)positive definite but can be quite bad conditioned,
# e.g. if there are pedigree errors or label switching
save(H,file="H.Rdata")
```

Figure 1

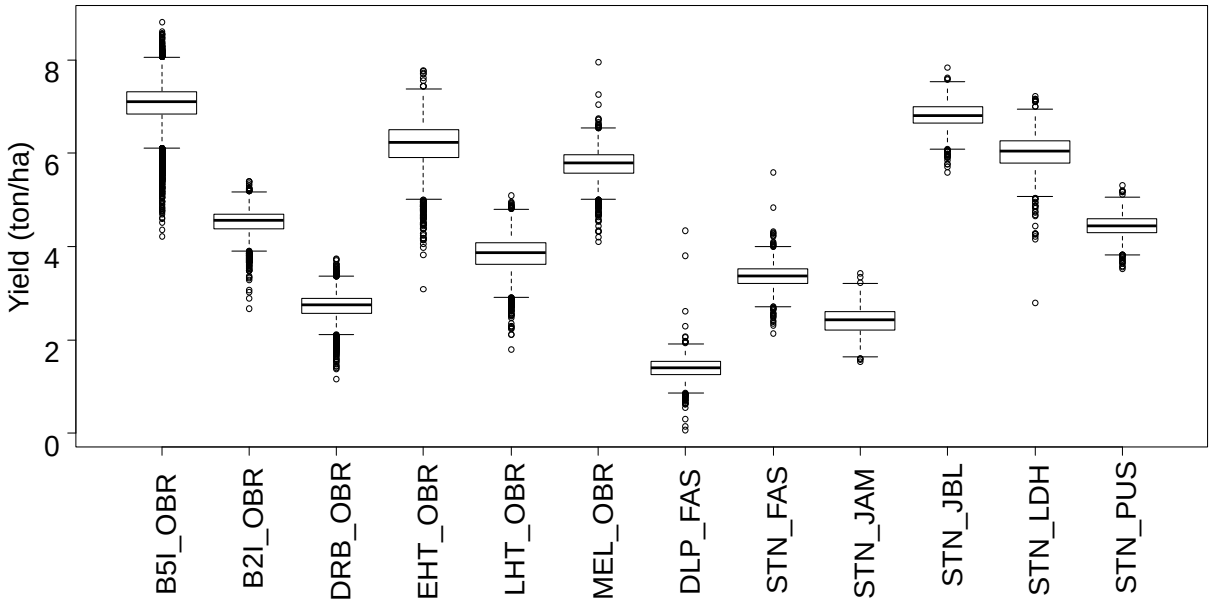


Figure 1: Boxplot of grain yield (t/ha) by environment (condition-location combination). OBR=Obregon, Mexico, FAS=Faisalabad, Pakistan; JAM=Jamalpur, Bangladesh; JBL=Jabalpur, India; LDH=Ludhiana, India; PUS=Pusa, India. Standard management conditions (STN); Delayed planting (DLP); Bed planting and 5 irrigations (B5I); Bed planting and 2 irrigations (B2I); Zero till, bed planting and 5 irrigations (Z5I); Zero till, bed planting and 2 irrigations (Z2I); Melgas flat planting (MEL); Drip irrigation to impose drought in flat (DRM); drip irrigation to impose drought, beds (DRB); Early heat (EHT); Late heat (LHT).

Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

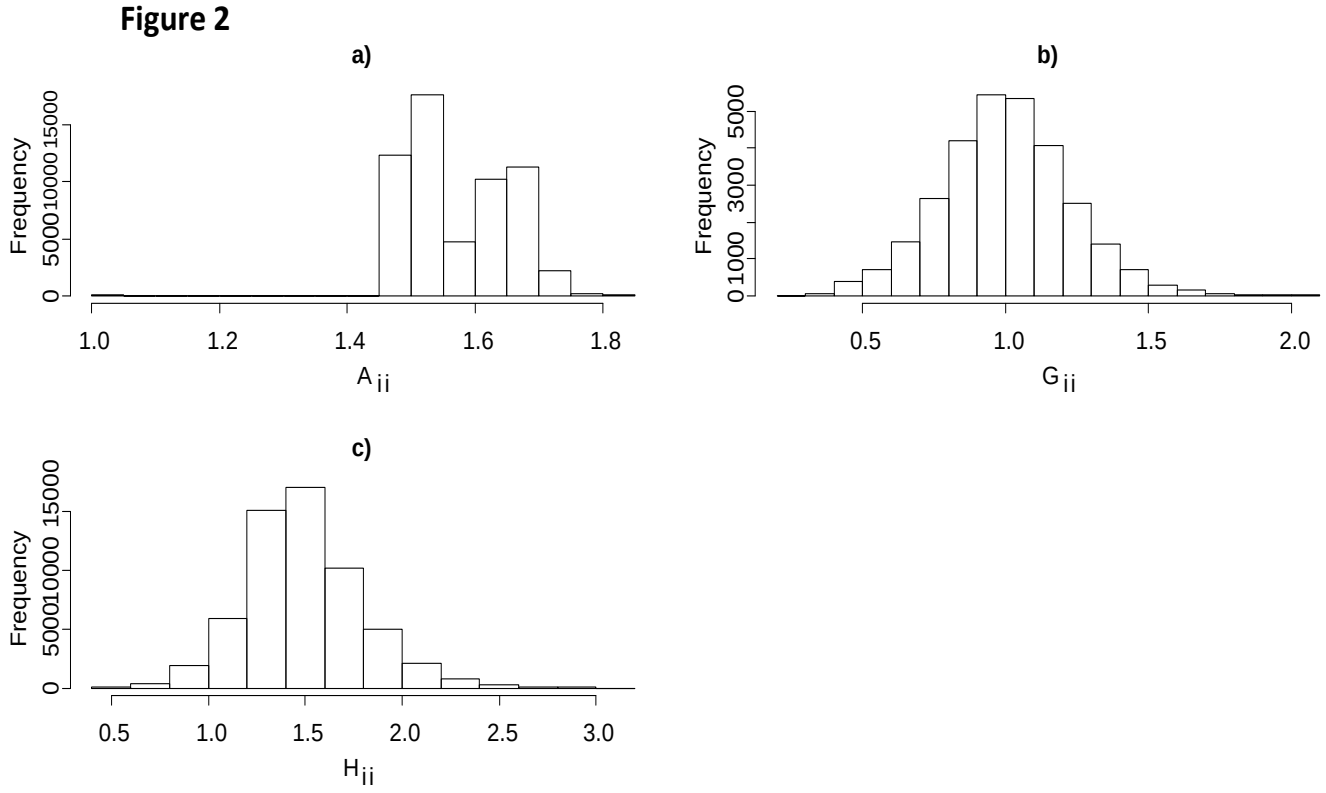


Figure 2: Distribution of the diagonal entries of a) the additive relationship matrix derived from pedigree (***A***); b) the genomic relationship matrix (***G***); and c) the ***H*** matrix.

Figure 3a

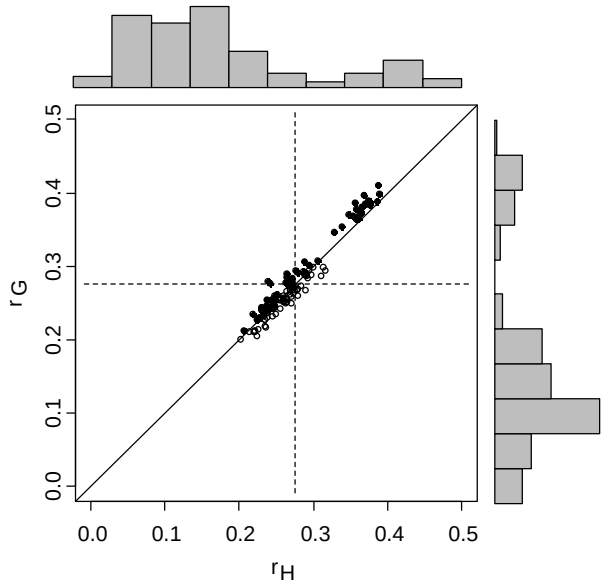
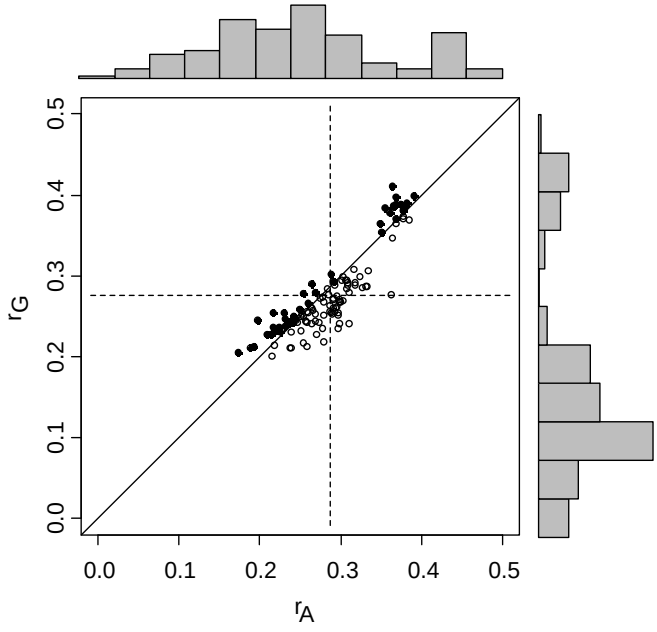


Figure 3b



Plant Gen. Accepted Paper, posted 01/25/2017. doi:10.3835/plantgenome2016.09.0089

Figure 3c

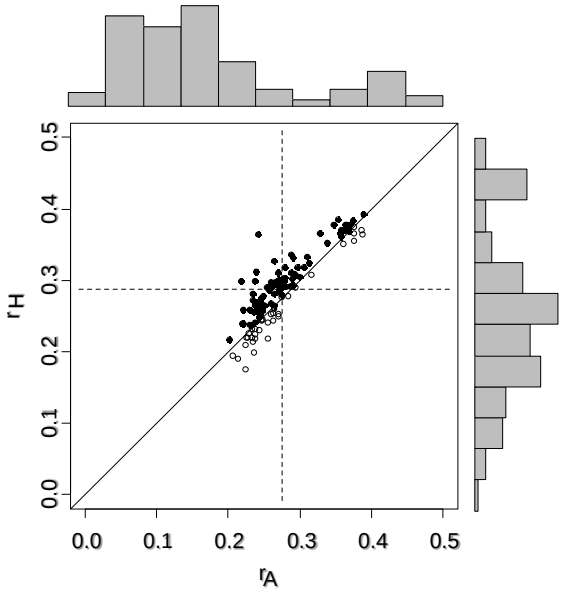


Figure 3: Plots of the predictive correlations for each of 20 cross-validations and 6 environments in South Asia for grain yield. (a) When the best linear model is based on ***G*** matrix, this is represented by black squared; when the best model is based on the ***H*** matrix, this is represented by a white squared; (b) When the best model is based on the ***G*** matrix this is represented by a black squared; when the best linear model is based on the ***A*** matrix, this is represented by a white squared, (c) When the best model is based on ***H***, this is represented by a black squared; when the best linear model is based on the ***A*** matrix, this is represented by an white squared. The histograms depict the distribution of the correlations in the testing set obtained from the partitions for different models. The horizontal (vertical) dashed line represents the average of the correlations for the testing set in the partitions for the model shown on the Y (X) axis. The solid line represents Y = X; i.e., both models have the same prediction ability.

Figure 4

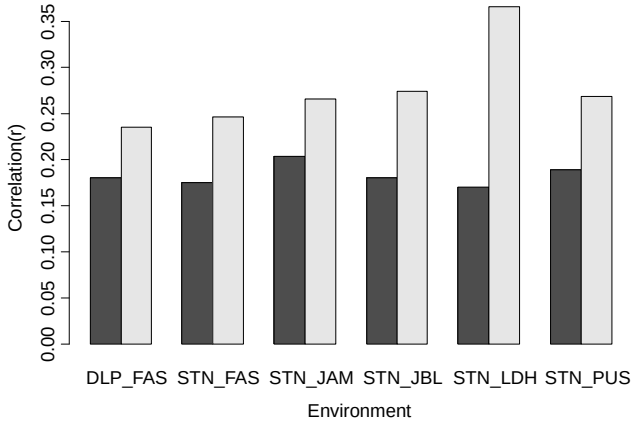


Figure 4: Barplot of correlations for each predicted environment in South Asia. Gray bars represent the means of the correlations between observed and predicted values obtained from cross-validation in Table 6 using the ***H*** matrix. Black bars represent a weighted mean of the phenotypic correlation of a given environment and the rest of environments in Table 4; for example, for DPL_FAS, the weighted correlation can be obtained using data shown in Table 8. FAS=Faisalabad, Pakistán; JAM=Jamalpur, Bangladesh; JBL=Jabalpur, India; LDH=Ludhiana, India, OBR=Obregon, Mexico; PUS=Pusa, India. Standard management conditions (STN), and delayed planting conditions (DLP).