Published online July 13, 2017

Single-Step Genomic and Pedigree Genotype × Environment Interaction Models for Predicting Wheat Lines in International Environments

Paulino Pérez-Rodríguez, José Crossa,* Jessica Rutkoski, Jesse Poland, Ravi Singh, Andrés Legarra, Enrique Autrique, Gustavo de los Campos, Juan Burgueño, and Susanne Dreisigacker

ABSTRACT

Genomic prediction models have been commonly used in plant breeding but only in reduced datasets comprising a few hundred genotyped individuals. 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 from genotyped individuals to pedigree information from a larger number of phenotyped individuals in order to combine relationship information on all members of the breeding population. Furthermore, genomic prediction models that incorporate genotype × environment interactions (G × E) have produced substantial increases in prediction accuracy compared with single-environment genomic prediction models. Our main objective was to show how to use single-step genomic and pedigree models to assess the prediction accuracy of 58,798 CIMMYT wheat (*Triticum aestivum* L.) lines evaluated in several simulated environments in Ciudad Obregon, Mexico, and to predict the grain yield performance of some of them in several sites in South Asia (India, Pakistan, and Bangladesh) using a reaction norm model that incorporated G × E. Another objective was to describe the statistical and computational challenges encountered when developing the pedigree and single-step models in such large datasets. Results indicate that the genomic prediction accuracy achieved by models using pedigree only, markers only, or both pedigree and markers to predict various 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 do not include the G × E term.

Core Ideas

- Genomic prediction accuracy models have been commonly used in plant breeding but only in reduced datasets comprising a few hundred genotyped individual plants.
- In this study we used pedigree and genomic data from 58,798 wheat lines evaluated in different environments.
- We use pedigree and genomic information in a model that incorporates genotype × environment interactions to predict wheat line performance in environments in South Asia.

GLOBAL WHEAT PRODUCTION is increasing by 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, because of climate change, temperatures could increase well beyond the optimal for growing wheat, which would further reduce

P. Pérez-Rodríguez, Colegio de Postgraduados, CP 56230, Montecillos, Estado de México, México; J. Crossa, J. Rutkoski, R. Singh, E. Autrique, J. Burgueño, and S. Dreisigacker, CIMMYT, Apdo. Postal 6-641, 06600 México City, México; J. Poland, USDA-ARS and Dep. of Agronomy, Kansas State Univ., 4011 Throckmorton Hall, Manhattan KS, 66506; A. Legarra, Institut National de la Recherche Agronomique, UR631 Station d'Amélioration Génétique des Animaux, BP 52627, 32326 Castanet-T, France; G. de los Campos, Dep. of Epidemiology & Biostatistics, Michigan State Univ., 909 Fee Road, Room B601, East Lansing, MI 48824. Received 3 Sept. 2016. Accepted 6 Jan. 2017. *Corresponding author (j.crossa@cgiar.org). Assigned to Associate Editor Shawn Kaeppler.

Abbreviations: G × E, genotype × environment interaction effects; GS, genomic selection; Matrix **A**, relationship matrix containing pedigree information; Matrix **G**, relationship matrix containing genome marker information; Matrix **H**, matrix combining Matrices **A** and **G** (pedigree and marker information); RAM, random access memory; ssGBLUP, single-step genomic best linear unbiased predictor.

Published in Plant Genome
Volume 10. doi: 10.3835/plantgenome2016.09.0089

© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA

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 achieving 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 the training and the testing sets, and 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; González-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, because of the genotypic costs. Nejati-Javaremi et al. (1997) were the first to propose incorporating genotypic information for predicting the breeding values of animals in a similar manner to the way pedigree information is used in the best linear unbiased predictor 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 a pedigree-based relationship matrix (Matrix **A**) with contributions from a genomic relationship matrix (Matrix **G**) of the genotyped individuals. They showed how to modify the original Matrix **A** to obtain Matrix **H**, which 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., CIMMYT's Global Wheat Program (GWP), which generates thousands of new breeding lines that are candidates for field evaluation every cropping cycle]. Applying GS in the GWP is economically feasible (i) when advancing breeding lines in the first preliminary yield trials to predict the performance of the selected lines in multienvironment trials or (ii) for predicting a selected set of lines in different international target environments using the parents evaluated in Mexico and the progeny to be predicted in international environments such as those in South Asia as a training set.

In recent years, the GWP aimed to form a large reference dataset comprising 58,798 breeding lines, including the lines' phenotypic and pedigree data from the last seven cropping cycles in Ciudad 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 that have molecular markers only with those that have pedigree and phenotype must be generated.

The main objectives of this study were (i) to use the large reference set for predicting the performance of wheat lines in several environments in South Asia; and (ii) to perform predictions using phenotypic, pedigree, and genomic information to evaluate the wheat lines genetically using a single-step model that combines pedigree and marker information into a unified **H** matrix. Here, we used information for genotyped and nongenotyped 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 (Matrix **A**), genomic information (Matrix **G**), or both (Matrix **H**) and comparing its prediction accuracy results with those of a genomic model that does not include the G × E interaction. This reaction norm model uses highly random dimensional matrices for the genomic and pedigree

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	Optimal
Delayed planting	Late heat
Bed planting and five irrigations	Optimal
Bed planting and two irrigations	Drought
Zero-till, bed planting and five irrigations	Optimal
Zero-till, bed planting and two irrigations	Drought
Melgas flat planting and five irrigations	Optimal
Melgas flat planting and drip irrigation	Severe drought
Bed planting and drip irrigation	Severe drought
Early heat	Early heat
Late heat	Late heat

matrices. We also describe the statistical and computational challenges encountered when developing the pedigree and single-step models in such large datasets.

MATERIALS AND METHODS

Experimental Data

The dataset 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 to 2016. The original data from each year comprise a large number of trials, each established using an α-lattice design with three replicates. The field management conditions under which each trial was established in each year are described in Table 1. The condition–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 (accessed 5 Apr. 2017). Given the dimensions of **A**, it is difficult to hold it in random access memory (RAM) 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 results and speed up the computations; R was recompiled from the source and linked with OpenBLAS [http://www.openblas.net (accessed 5 Apr. 2017)]. For further details on the computations, see Appendix A. In total, 29,484 individuals were genotyped using genotyping-by-sequencing (e.g., Elshire et al., 2011). We kept all the single nucleotide polymorphism markers and imputed the missing values using observed data. Markers with a minor allele frequency of less than 0.05 were removed; after this process, 9045 markers were available for prediction.

Table 2. Number of lines evaluated in different environments during 2009–2016 by the Global Wheat Program.

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, Ciudad Obregon, Mexico; PUS, Pusa, India; STN, standard management conditions; DLP, delayed planting; B5I, bed planting and five irrigations; B2I, bed planting and two irrigations; MEL, Melgas flat planting and five irrigations; DRB, bed planting and drip irrigation; EHT, early heat; LHT, late heat.

Statistical Models

Recently, Jarquín et al. (2014) and López-Cruz et al. (2015) proposed statistical models for performing genomic predictions taking G × E into account. 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. Jarquín's 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 reparametrized 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 the main effect of environments (E), the main effect of genotypes and the interaction between genotypes and the environment into account. In matrix notation, the model can be written as:

$$\mathbf{y} = \mu 1 + \mathbf{Z}_{\mathrm{E}}\boldsymbol{\beta}_{\mathrm{E}} + \mathbf{Z}_{\mathrm{g}}\boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{e}\,, \qquad [1]$$

where $\mathbf{y} = (\mathbf{y}_1,\ldots,\mathbf{y}_{\mathrm{E}})'$ is the response vector and $\mathbf{y}_j$ represents the observations in the j^{th} environment ($j = 1,\ldots,E$). The general mean is μ; $\mathbf{Z}_{\mathrm{E}}$ is an incidence matrix for environments, which is assumed to be multivariate with $\boldsymbol{\beta}_E \sim MN(\mathbf{0},\sigma_E^2\mathbf{I})$; $\mathbf{Z}_{\mathrm{g}}$ is an incidence matrix that connects genotypes with phenotypes; $\mathbf{u}_1$ represents the random effect of genotypes, which is assumed multivariate with $\mathbf{u}_1 \sim MN(\mathbf{0},\sigma_{\mathrm{u}}^2\mathbf{A})$; and $\mathbf{u}_2$ represents the effect of G × E interaction. We assume $\mathbf{u}_2 \sim MN(\mathbf{0},\sigma_{\mathrm{ge}}^2(\mathbf{Z}_{\mathrm{g}}\mathbf{G}\mathbf{Z}_{\mathrm{g}}')\#(\mathbf{Z}_{\mathrm{E}}\mathbf{Z}_{\mathrm{E}}'))$, 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: $\mathbf{e} \sim MN(\mathbf{0},\sigma_e^2\mathbf{I})$, where **e** is the residual error; *MN* is the multivariate normal, and **I** is the identity matrix.

Since **A** is positive definite and symmetric, it can be factored as $\mathbf{A} = \mathbf{L}\mathbf{L}'$ by using Cholesky decomposition where Matrix **L** is a lower triangular matrix with positive diagonal entries and is usually named the Cholesky factor. Therefore, from Eq. [1]:

$$\mathbf{Z}_{\mathrm{g}}\mathbf{u}_1 \overset{d}{=} \mathbf{Z}_{\mathrm{g}}\mathbf{L}\mathbf{u}_1^*\,, \qquad [2]$$

where $\mathbf{u}_1^* \sim MN(\mathbf{0},\sigma_{\mathrm{u}}^2\mathbf{I})$. Furthermore, it is not necessary to calculate the $\mathbf{Z}_{\mathrm{g}}\mathbf{L}$ product because for each row of the resulting matrix, we just need to copy the k^{th} row of **L**, where *k* is the column in the i^{th} row of $\mathbf{Z}_{\mathrm{g}}$ that is different from zero (i.e., $\mathbf{Z}_{\mathrm{g}}$ (*i*, *k*) = 1). The matrix $\mathbf{Z}_{\mathrm{E}}\mathbf{Z}_{\mathrm{E}}'$ is a block diagonal; blocks different from zero correspond to matrices with ones:

$$\mathbf{Z}_{\mathrm{E}}\mathbf{Z}_{\mathrm{E}}' = \begin{pmatrix} \mathbf{J}_1 & & & \\ & \mathbf{J}_2 & & \\ & & \ddots & \\ & & & \mathbf{J}_{\mathrm{E}} \end{pmatrix}, \qquad [3]$$

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

$$\tilde{\mathbf{A}} = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}' = \begin{pmatrix} \tilde{\mathbf{Z}}_{11} & \tilde{\mathbf{Z}}_{12} & \cdots & \tilde{\mathbf{Z}}_{1\mathrm{E}} \\ \tilde{\mathbf{Z}}_{21} & \tilde{\mathbf{Z}}_{22} & \cdots & \tilde{\mathbf{Z}}_{2\mathrm{E}} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{Z}}_{\mathrm{E}1} & \tilde{\mathbf{Z}}_{\mathrm{E}2} & \cdots & \tilde{\mathbf{Z}}_{\mathrm{EE}} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{Z}}_{11} & \tilde{\mathbf{Z}}_{12} & \cdots & \tilde{\mathbf{Z}}_{1\mathrm{E}} \\ \tilde{\mathbf{Z}}_{21} & \tilde{\mathbf{Z}}_{22} & \cdots & \tilde{\mathbf{Z}}_{2\mathrm{E}} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{Z}}_{\mathrm{E}1} & \tilde{\mathbf{Z}}_{\mathrm{E}2} & \cdots & \tilde{\mathbf{Z}}_{\mathrm{EE}} \end{pmatrix}', \qquad [4]$$

The block diagonal elements of $\tilde{\mathbf{A}}$ can be computed as follows:

$$\begin{aligned} \tilde{\mathbf{A}}_{11} &= \sum_{\text{Environments}} \tilde{\mathbf{Z}}_{1\mathrm{j}}\tilde{\mathbf{Z}}_{1\mathrm{j}}' = \mathbf{A}_{11}\,, \\ &\vdots \\ \tilde{\mathbf{A}}_{\mathrm{EE}} &= \sum_{\text{Environments}} \tilde{\mathbf{Z}}_{\mathrm{Ej}}\tilde{\mathbf{Z}}_{\mathrm{Ej}}' = \mathbf{A}_{\mathrm{EE}}\,, \end{aligned} \qquad [5]$$

where $\mathbf{A}_{jj}$, corresponds to the relationship matrix for individuals evaluated in environment *j*. From Eq. [3] and Eq. [5] and by using Cholesky decomposition, the term $\mathbf{Z}_{\mathrm{g}}\mathbf{A}\mathbf{Z}_{\mathrm{g}}'\#\mathbf{Z}_{\mathrm{E}}\mathbf{Z}_{\mathrm{E}}'$ can be obtained as follows:

$$\begin{aligned} \mathbf{Z}_{\mathrm{g}}\mathbf{A}\mathbf{Z}_{\mathrm{g}}'\#\mathbf{Z}_{\mathrm{E}}\mathbf{Z}_{\mathrm{E}}' &= BDiag\left(\mathbf{A}_{11},\ldots,\mathbf{A}_{\mathrm{EE}}\right) \\ &= BDiag(\mathbf{L}_1\mathbf{L}_1',\ldots,\mathbf{L}_{\mathrm{E}}\mathbf{L}_{\mathrm{E}}') = \mathbf{L}_{\mathrm{ge}}\mathbf{L}_{\mathrm{ge}}' \end{aligned}, \qquad [6]$$

where $\mathbf{L}_{\mathrm{ge}} = BDiag(\mathbf{L}_1,\ldots,\mathbf{L}_{\mathrm{E}})$. Therefore, from Eq. [6], we obtain:

$$\mathbf{u}_2 \overset{d}{=} \mathbf{L}_{\mathrm{ge}}\mathbf{u}_2^*\,, \qquad [7]$$

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

Therefore, using the results from Eq. [2] and Eq. [7], Model 1 can be written as:

$$\mathbf{y} = \mu 1 + \mathbf{Z}_{\mathrm{E}}\boldsymbol{\beta}_{\mathrm{E}} + \mathbf{Z}_{\mathrm{g}}\mathbf{L}\mathbf{u}_1^* + \mathbf{L}_{\mathrm{ge}}\mathbf{u}_2^* + \mathbf{e} \qquad [8]$$

Equation [1] and Eq. [8] are equivalent, but Eq. [8] has at least two advantages over Eq. [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 **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 $\mathbf{G} = \dfrac{\mathbf{W}\mathbf{W}'}{p}$ be the genomic relationship matrix (López-Cruz et al., 2015). A model similar to Eq. [8] can be obtained by replacing **A** with **G**.

Model 3: G × E Interaction Using Molecular Markers and Pedigree (Single-Step Approach)

In this model, the information for genotyped and nongenotyped 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 as:

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{\mathrm{nn}} + \mathbf{A}_{\mathrm{gn}}'\mathbf{A}_{\mathrm{gg}}^{-1}\left(\mathbf{G}_{\mathrm{a}} - \mathbf{A}_{\mathrm{gg}}\right)\mathbf{A}_{\mathrm{gg}}^{-1}\mathbf{A}_{\mathrm{gn}} & \mathbf{A}_{\mathrm{gn}}'\mathbf{A}_{\mathrm{gg}}^{-1}\mathbf{G}_{\mathrm{a}} \\ \mathbf{G}_{\mathrm{a}}\mathbf{A}_{\mathrm{gg}}^{-1}\mathbf{A}_{\mathrm{gn}} & \mathbf{G}_{\mathrm{a}} \end{bmatrix}, \qquad [9]$$

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

$$\mathbf{G}_a = \beta\mathbf{G} + \alpha\,, \qquad [10]$$

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

$$\text{Avg}\left(\text{diag}\left(\mathbf{G}\right)\right)\beta + \alpha = \text{Avg}\left(\text{diag}\left(\mathbf{A}_{gg}\right)\right), \qquad [11]$$

$$\text{Avg}\left(\mathbf{G}\right)\beta + \alpha = \text{Avg}(\mathbf{A}_{gg})\,, \qquad [12]$$

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

The Appendix shows the R code that allowed us to build Matrix **H**. A parametric G × E interaction model takes the effect of the environments, the main effect of genotypes and the G × E interaction into account. A model that uses information obtained from markers and pedigree can be obtained by replacing the **A** matrix in Model 1 with the Matrix **H** described above (Eq. [8]).

Model without G × E Interactions

Note that models that do not include the G × E term can be derived from Model 1 to Model 3 just by removing the corresponding random G × E term. For example, by removing the term $\mathbf{u}_2$ representing the effect of G × E from Model 1 (Eq. [8]), it becomes $\mathbf{y} = \mu\mathbf{1} + \mathbf{Z}_E\boldsymbol{\beta}_E + \mathbf{Z}_g\mathbf{u}_1 + \mathbf{e}$.

In this case, the resulting models are equivalent to the cross-environment genomic best linear unbiased predictor model of López-Cruz et al. (2015). We include models without the G × E term to compare the prediction accuracy of models with and without G × E interactions. 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, and thus developing Matrix **H** for the single-step model did not make sense.

Assessing the Models' Predictive Ability

The main interest of breeders is to predict the performance of nonevaluated lines in South Asian sites (Jalbapur, Ludhiana, and 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 all available records under drought, late heat, optimal, and severe drought conditions obtained in Ciudad Obregon (Mexico), and 20% of available records in each of the South Asian sites assigned at random as the training set. 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 abilities were compared by using Pearson's correlation coefficient. The models that used the **A** and **H** matrices included the phenotypic information of the 58,798 wheat lines, whereas the model that was based on markers only included information for 29,484 wheat lines that correspond to the individuals that were genotyped. The genotyped individuals were a subset of the individuals with pedigree information; therefore, lines in the testing set had pedigree and marker information. The numbers of individuals in the testing sets in South Asian sites were 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.

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

† FAS, Faisalabad, Pakistan; JAM, Jamalpur, Bangladesh; JBL, Jabalpur, India; LDH, Ludhiana, India; STN, standard management conditions; DLP, delayed planting.

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 to accept big.matrix objects created using the bigmemory package as input (Kane et al., 2013). The bigmemory package was used to handle the huge matrices that had 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/ (accessed 5 Apr. 2017).

RESULTS

Descriptive Statistics

Figure 1 shows a boxplot of grain yield per location and median yield per location. From the plot, it can be seen that the optimal conditions had the highest grain yield, whereas 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

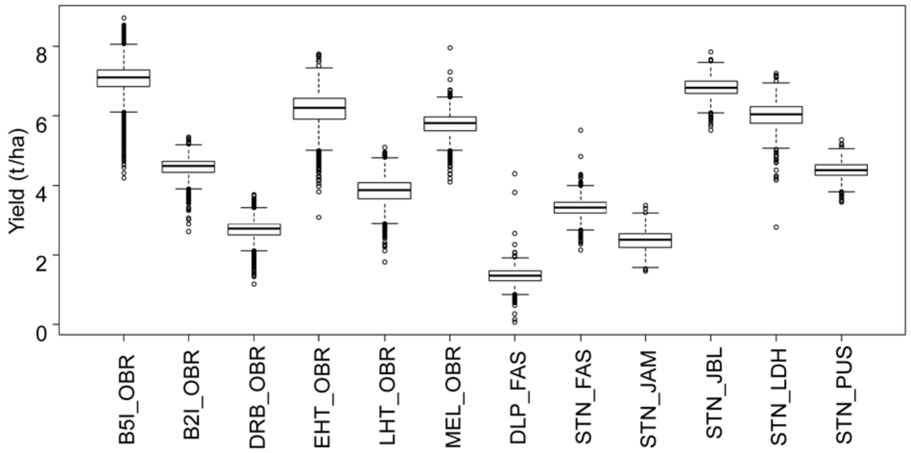

Fig. 1. Boxplot of wheat grain yield (t ha^{-1}) per environment (condition–location combination). OBR, Obregon, Mexico; FAS, Faisalabad, Pakistan; JAM, Jamalpur, Bangladesh; JBL, Jabalpur, India; LDH, Ludhiana, India; PUS, Pusa, India; STN, standard management conditions; DLP, delayed planting; B5I, bed planting and five irrigations; B2I, bed planting and two irrigations; Z5I, zero-till, bed planting, and five irrigations; Z2I, zero-till, bed planting, and two irrigations; MEL Melgas flat planting and five irrigations; DRM, Melgas flat planting and drip irrigation to impose drought; DRB, bed planting and drip irrigation; EHT, early heat; LHT, late heat.

than those in Mexican environments. Table 4 shows the number of lines evaluated in each environment and the number of 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 4735. The phenotypic sample correlation ranged from –0.05 to 0.53, which suggests large G × E effects.

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 Matrix **G** are around 1.0, reflecting the fact that the markers were centered and standardized. The diagonal entries of Matrix **H** are around 1.5, which stems from standarding **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 wheat lines in common between the environment with the bed planting with five irrigations at Obregon and the bed planting with two irrigations at Obregon is 0.156, whereas the phenotypic correlation of the 1537 wheat lines in common between the bed planting with five irrigations at Obregon and standard management conditions at Pusa, India, is 0.210. In general, the phenotypic correlations were not high, ranging from −0.051 to 0.481.

Table 4. Number of genotypes (diagonal, in bold), number of genotypes in common in a pair of environments (upper triangular), and sample phenotypic correlations (lower triangular) per environment (Env.).

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	**56,964**	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	0.120	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; STN, standard management conditions; DLP, delayed planting; B5I, bed planting and five irrigations; B2I, bed planting and two irrigations; MEL, Melgas flat planting; DRB, bed planting and drip irrigation; EHT, early heat; LHT, late heat.

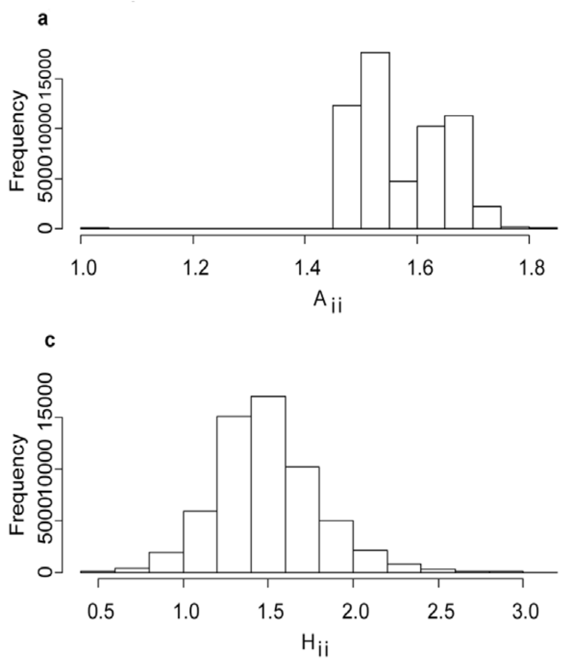

Fig. 2. Distribution of the diagonal entries of (a) the additive relationship matrix derived from pedigree (**A**), (b) the genomic relationship matrix (**G**), and (c) the combined matrix (**H**).

Table 5 shows the average Pearson's correlations between observed and predicted phenotypes and their corresponding SD for the model without 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 for 80% of the entries included in the six South Asian environments. The prediction accuracies are relatively low, with those based on pedigree being slightly higher than those based on markers or on both pedigree and markers.

Table 5. Correlations (plus SD in parentheses) between predicted and observed values obtained by using the cross-validation where all the wheat lines from Ciudad 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 for models without genotype × environment effects (G × E).

	Models 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)

† The highest correlations in each environment are in bold typeface.

‡ FAS, Faisalabad, Pakistan; JAM, Jamalpur, Bangladesh; JBL, Jabalpur, India; LDH, Ludhiana, India; PUS, Pusa, India; STN, standard management conditions; DLP, delayed planting conditions.

Prediction Accuracy of G × E Models

Table 6 shows the average Pearson's correlations between the observed and predicted phenotypes and the corresponding standard obtained using the same cross-validation scheme described above but including the G × E

Table 6. Correlations (plus SD in parentheses) between predicted and observed values obtained using the cross-validation where all the wheat lines from Ciudad 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 for models with genotype × environment effects (G × E).

	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)

† The highest correlations in each environment are in bold typeface.

‡ FAS, Faisalabad, Pakistan; JAM, Jamalpur, Bangladesh; JBL, Jabalpur, India; LDH, Ludhiana, India; PUS, Pusa, India; STN, standard management conditions; DLP, delayed planting conditions.

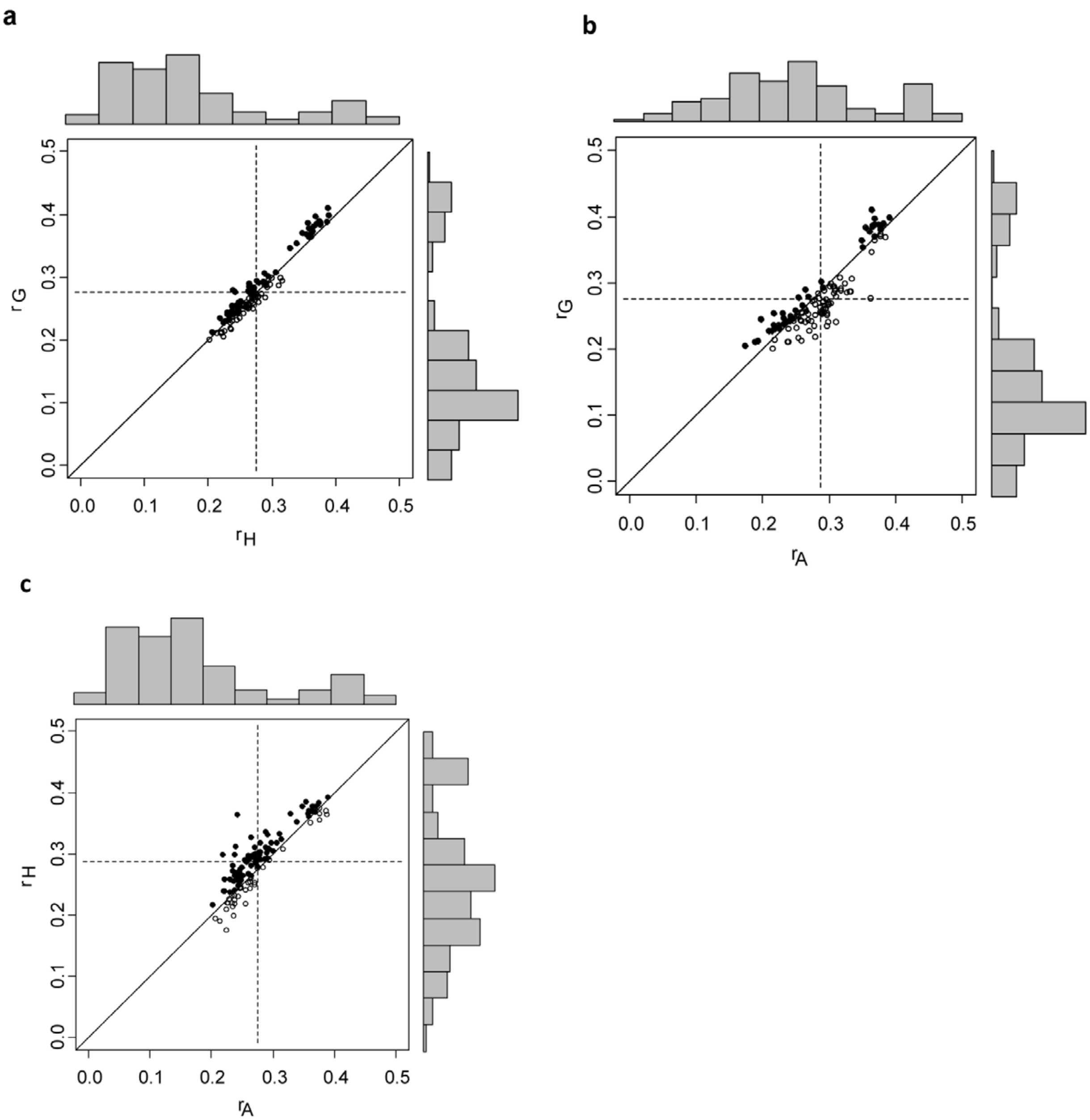

Fig. 3. Plots of the predictive correlations for each of 20 cross-validations and six environments in South Asia for wheatr grain yield. (a) When the best linear model is based on Matrix **G**, this is represented by black squares; when the best model is based on Matrix **H**, this is represented by white squares. (b) When the best model is based on Matrix **G**, this is represented by black squares; when the best linear model is based on Matrix **A**, this is represented by white squares. (c) When the best model is based on Matrix **H**, this is represented by black squares; when the best linear model is based on Matrix **A**, this is represented by white squares. 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).

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 (delayed planting at Faisalabad, standard management at Faisalabad, standard management at Jamalpur, and standard conditions at Pusa). 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 3a–c shows scatterplots of the predictive correlations for each of the 20 cross-validations across the six environments in South Asia. Figure 3a depicts the correlations between predicted values based on markers (Matrix **G**) versus those based on Matrix **H** and shows that the prediction accuracy based on Matrix **G** was superior to that obtained based on **H**. Figure 3b displays

Table 7. Comparison of the predictive ability of models with and without genotype × environment effects (G × E).

Environment‡	Pedigree (**A**)	Markers (**G**)	Pedigree + markers (**H**)
	% change†		
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 was calculated via Eq. [13].

‡ FAS, Faisalabad, Pakistan; JAM, Jamalpur, Bangladesh; JBL, Jabalpur, India; LDH, Ludhiana, India; PUS, Pusa, India; STN, standard management conditions; DLP, delayed planting conditions.

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

Table 7 shows the percentage of change in the prediction accuracy of models with and without G × E. The percentage of change was calculated as:

$$\frac{r_{\mathrm{G\times E}} - r_{\mathrm{noG\times E}}}{r_{\mathrm{noG\times E}}} \times 100 \ , \qquad [13]$$

where $r_{\mathrm{G\times E}}$ is the Pearson's correlation for a model with the G × E term and $r_{\mathrm{no\ G\times E}}$ is the Pearson's correlation for a model without the G × E term. From the 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 with 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. Black bars 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:

$$r_{\mathrm{j}} = \sum_{k\neq j} \frac{n_{\mathrm{jk}}}{n_{\mathrm{j}}} r_{\mathrm{jk}} \ , \qquad [14]$$

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*, $n_{\mathrm{j}} = \sum_{k} n_{\mathrm{jk}}$, 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 the environment with delayed planting at Faisalabad; 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 by using the approach described above. The gray bars represent the means of the correlations between the observed and predicted values obtained from cross-validations. Note that in general, the G × E models gave good predictions, usually better than those from the phenotypic correlations. Although we predicted 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.

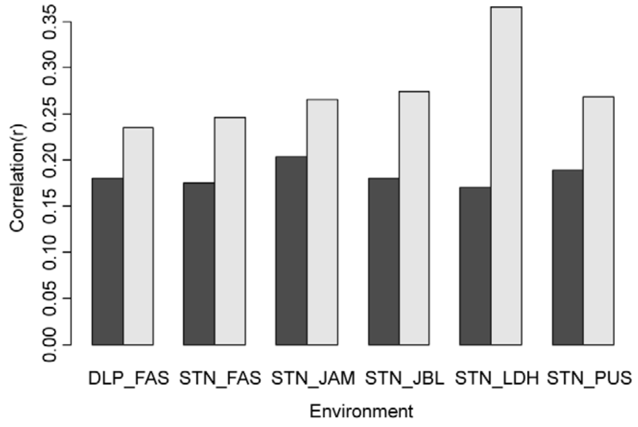

Fig. 4. Barplot of correlations for each predicted environment in South Asia. Gray bars represent the means of the correlations between observed and predicted values obtained from the cross-validation in Table 6 using Matrix **H**. Black bars represent the weighted mean of the phenotypic correlation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

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 problem would be to augment the numerical relationship Matrix (**A**) of all individuals with the genomic relationship matrix (**G**) of the genotyped individuals and to perform predictions based on the resulting complete Matrix **H**, which would allow us to predict nongenotyped 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 one to use all phenotypic data and not merely data from phenotypes that have pedigree and marker information; this extra phenotypic information should also enhance prediction

Table 8. Phenotypic correlations and numbers of lines in common between delayed planting conditions at Faisalabad, Pakistan (DLP_FAS) and the rest of the environments in Mexico and South Asia.

j	Environment in South Asia	k†	Other environments	r_{jk}	n_{jk}	$\frac{n_{jk}}{n_j}$	$\frac{n_{jk}}{n_j} r_{jk}$
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 =	15,448		r_1 = 0.18

† k = 1,…,11 represents the environments, r_1 represents the weighted phenotypic correlation for Environment 1 in South Asia (i.e. Faisalabad), and n_1 is the total of the column labeled as n_{jk}.

‡ FAS, Faisalabad, Pakistan; JAM, Jamalpur, Bangladesh; JBL, Jabalpur, India; LDH, Ludhiana, India; PUS, Pusa, India; STN, standard management conditions; DLP, delayed planting; B5I, bed planting and five irrigations; B2I, bed planting and two irrigations; MEL, Melgas flat planting; DRB, bed planting and drip irrigation; EHT, early heat; LHT, late heat

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 nongenotyped plants.

In a recent article, Fernando et al. (2014) proposed a single-step Bayesian regression strategy that allows the use of all genotyped and nongenotyped individuals by means of imputed marker covariates for nongenotyped individuals. The advantage of the Bayesian approach over the single-step best linear unbiased predictor is that it does not require one to compute 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 1176 genotyped CIMMYT wheat lines and 11,131 nongenotyped wheat lines tested in five environments in Ciudad Obregon, Mexico, during the 2012–2013 cycle. We developed optimized weighting factors for Matrix **H** and applied a multivariate method for assessing G × E and 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 Ciudad Obregon and predicted 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 Table 5 to Table 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 the 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 phenotypic data only.

Genomic Prediction Accuracy Versus the 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 Ciudad Obregon, Mexico, and predicted lines in several South Asian environments. For Ludhiana, Pusa, and Jabalpur, about 1227 wheat lines were predicted on the basis of the performance of these lines in six environments in Ciudad Obregon plus the performance of about 57,000 wheat lines related to those to be predicted (1227) and evaluated in previous years in Ciudad Obregon, Mexico.

Prediction accuracy was the correlation between the predicted values of the lines in Ciudad Obregon plus a low proportion of them (20%) in six environments in South Asia using three G × E models (those using Matrices **A**, **G**, and **H**) with the observed values of 80% of the lines in the six environments in South Asia (which were not phenotyped). The correlations for all the environments were around 0.25 to 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 Ciudad Obregon, Mexico, is lower than the accuracy they could obtain by performing genomic selection and prediction. Although wheat breeders expect that lines selected in Ciudad Obregon will perform well in South Asian environments, the results of this study should prompt them to increase research on genomic selection in Ciudad 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 Matrices **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**, one to use 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 lines with pedigree data and another subset of phenotype data from lines with marker data only. The single-step method for computing Matrix **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 the Mendelian sample component (or with family variability) accounted for the 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 GS literature. This is the first study to show 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 has been 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 dataset in conjunction with pedigree, markers, or both, in GS and prediction. We compared models including and excluding G × E. 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 **H** allows the use of all the available information. Also, the reaction norm G × E model allows us to swap information among positively correlated environments, although the predictive power of the model was similar to that of the model that included markers only. Ashraf et al. (2016) used the single-step **H** approach on a set of 11,131 nongenotyped and 1176 genotyped wheat lines.

Animal breeders make extensive use of the fact that the relationship Matrix **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 $\mathbf{H}^{-1}$ structure as well (Aguilar et al., 2010; Christensen and Lund, 2010), with a storage cost that is quadratic in the number of genotyped individuals but is only linear in the number of nongenotyped 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 as used in Eq. [8].

CONCLUSIONS

This study shows how to solve statistical and computational challenges when incorporating and combining high-dimensional pedigree and genomic matrices into a single-step model for predicting unobserved individuals in other environments. We found that the genomic prediction of genotyped and nongenotyped wheat lines produces higher prediction accuracy than that of lines predicted from 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.

APPENDIX

R Script Used 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 three columns separated by tabs, ID (the identification number of the individual), Sire (male parent), and Dam (female parent).

(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 with OpenBLAS (http://www.openblas.net, accessed 5 Apr. 2017). At the end of the process, the relationship matrix was also stored as a partitioned matrix on 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(“Mparent”,“FParent”,“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$MParent,dam=a$FParent,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:

$$\mathbf{A}_{\text{full}} = \mathbf{X}_{\text{full}}{}'\mathbf{X}_{\text{full}}\ , \qquad [A1]$$

where $\mathbf{X}_{\text{full}}$ is an upper triangular, sparse (right) Cholesky factor of the relationship matrix. In this case, $\mathbf{X}_{\text{full}}$ 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 $\mathbf{A}_{\text{full}}$; we just need a subset of this matrix with the 58,798 individuals so we can take a subset of 58,798 columns from $\mathbf{X}_{\text{full}}$. The columns correspond to the individuals that we are interested in. Let X be the resulting matrix; we then have $\mathbf{A} = \mathbf{X'X}$, where $\mathbf{X}$ 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 $\mathbf{X}$ is a huge matrix, it is very difficult to obtain $\mathbf{A}$ directly; furthermore, since $\mathbf{X}$ is sparse, the product cannot be parallelized easily. We then partitioned $\mathbf{X}$ into several submatrices and saved the submatrices as binary files that can later be retrieved in order to obtain the product. For example:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \\ \mathbf{X}_{31} & \mathbf{X}_{32} \\ \mathbf{X}_{41} & \mathbf{X}_{42} \\ \mathbf{X}_{51} & \mathbf{X}_{52} \end{pmatrix}; \mathbf{X}' = \begin{pmatrix} \mathbf{X}'_{11} & \mathbf{X}'_{21} & \mathbf{X}'_{31} & \mathbf{X}'_{41} & \mathbf{X}'_{51} \\ \mathbf{X}'_{12} & \mathbf{X}'_{22} & \mathbf{X}'_{32} & \mathbf{X}'_{42} & \mathbf{X}'_{52} \end{pmatrix}, \qquad [A2]$$

where $\mathbf{X}_{ij}$ is a submatrix obtained from $\mathbf{X}$.
The R code below was used to partition matrix $\mathbf{X}$ into five submatrices 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 Matrix $\mathbf{A}$ as follows:

$$\mathbf{A} = \mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}. \qquad [A3]$$

where:

$$\begin{aligned}
\mathbf{A}_{11} &= \mathbf{X}'_{11}\mathbf{X}_{11} + \mathbf{X}'_{21}\mathbf{X}_{21} + \mathbf{X}'_{31}\mathbf{X}_{31} + \mathbf{X}'_{41}\mathbf{X}_{41} + \mathbf{X}'_{51}\mathbf{X}_{51} \\
\mathbf{A}_{22} &= \mathbf{X}'_{12}\mathbf{X}_{12} + \mathbf{X}'_{22}\mathbf{X}_{22} + \mathbf{X}'_{32}\mathbf{X}_{32} + \mathbf{X}'_{42}\mathbf{X}_{42} + \mathbf{X}'_{52}\mathbf{X}_{52} \\
\mathbf{A}_{12} &= \mathbf{X}'_{11}\mathbf{X}_{12} + \mathbf{X}'_{21}\mathbf{X}_{22} + \mathbf{X}'_{31}\mathbf{X}_{32} + \mathbf{X}'_{41}\mathbf{X}_{42} + \mathbf{X}'_{51}\mathbf{X}_{52} \\
\mathbf{A}_{21} &= \mathbf{X}'_{12}\mathbf{X}_{11} + \mathbf{X}'_{22}\mathbf{X}_{21} + \mathbf{X}'_{32}\mathbf{X}_{31} + \mathbf{X}'_{42}\mathbf{X}_{41} + \mathbf{X}'_{52}\mathbf{X}_{51}\ ,
\end{aligned} \qquad [A4]$$

Note that now we need to calculate several products of matrices. There are optimized libraries that can be used for this task. For example, in R, we can recompile the program so that we can use OpenBLAS. Details are given at http://www.openblas.net/ (accessed 5 Apr. 2017) and http://www.rochester.edu/college/psc/thestarlab/help/moreclus/BLAS.pdf (accessed 5 Apr. 2017).

We recompiled R version 3.2.3 (R Core Team, 2016) in order to use OpenBLAS so it can perform matrix operations in parallel. The next fragment of code obtains Matrix $\mathbf{A}_{11}$ using the partitioned matrices.

```
rm(list=ls())
n_submatrix=5
A11=matrix(0,nrow=25000,ncol=25000)
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 obtained similarly. With this approach and by using eight cores for the matrix product, we obtained the 58,798 × 58,798 Matrix ***A*** in less than 3 hr in the CIMMYT-BSU server, which has 12 Intel Xeon Cores (Intel, Santa Clara, CA) @ 3.47 GHz and ~ 48 Gb of RAM.

R Script to Obtain Matrix H

The script presented below computes the relationship Matrix **H** that includes 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 the relationship Matrix **A** share the same scale (Christensen et al., 2012).

Inputs:

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

Output:

1) Matrix **H**.

```
#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(genotypednotin
  pedigree),“\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). Singlestep methods
# for genomic evaluation in pigs. animal,6, pp 15651571
#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)
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")
```

Conflict of Interest Disclosure

The authors declare no conflicts of interest.

Acknowledgments

We are grateful to the field and lab workers of the CIMMYT's GWP and to national program researchers who collected the data used in this study. This work was funded by the US Agency for International Development's 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. The authors would also like to thank the Bill and Melinda Gates Foundation for providing financial support through the grant OPPGD1389.

References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752. doi:10.3168/jds.2009-2730

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computations of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128(6):422–428. doi:10.1111/j.1439-0388.2010.00912.x

Ashraf, B., V. Edriss, D. Akdemir, E. Autrique, D. Bonnett, J. Crossa, et al. 2016. Genomic prediction using phenotypes from pedigree lines with no markers. Crop Sci. 56:957–964. doi:10.2135/cropsci2015.02.0111

Bates, D., and A. Vazquez. 2009. pedigreemm: Pedigree-based mixed-effects models. R Foundation for Statistical Computing. http://CRAN.R-project.org/package=pedigreemm (accessed 5 Apr. 2016).

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 Sci. 52(2):707–719. doi:10.2135/cropsci2011.06.0299

Christensen, O.F., and M.S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2. doi:10.1186/1297-9686-42-2

Christensen, O.F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal 6:1565–1571. doi:10.1017/S1751731112000742

Crossa, J., G. de los Campos, P. Pérez-Rodríguez, D. Gianola, J. Burgueño, J.L. Araus, et al. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi:10.1534/genetics.110.118521

Crossa, J., P. Pérez-Rodríguez, G. de los Campos, G. Mahuku, S. Dreisigacker, and C. Magorokosho. 2011. Genomic selection and prediction in plant breeding. J. Crop Improv. 25(3):239–261. doi:10.1080/15427528.2011.558767

de los Campos, G., and P. Pérez-Rodríguez. 2015. BGLR: Bayesian generalized linear regression. R package version 1.0.4. R Foundation for Statistical Computing. http://CRAN.R-project.org/package=BGLR (accessed 5 Apr. 2017).

de los Campos, G., D. Gianola, G.J.M. Rosa, K.A. Weigel, and J. Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92:295–308. doi:10.1017/S0016672310000285

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, et al. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385. doi:10.1534/genetics.109.101501

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E. Buckler, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:E19379. doi:10.1371/journal.pone.0019379

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

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6(6):721–741. doi:10.1109/TPAMI.1984.4767596

González-Camacho, J.M., G. de los Campos, P. Pérez-Rodríguez, D. Gianola, 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 T.P. Callanan. 1989. Computational aspects of likelihood-based inference for variance components. In: D. Gianola and K. Hammond, editors, Advances in statistical methods for genetic improvement of livestock. Springer-Verlag, Berlin. p. 136–176.

Henderson, C.R. 1975. Rapid method for computing the inverse of a relationship matrix. J. Dairy Sci. 58(11):1727–1730. doi:10.3168/jds.S0022-0302(75)84776-X

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. doi:10.2307/2529339

Henderson, C.R. 1977. Best linear unbiased prediction of breeding values not in the model for records. J. Dairy Sci. 60(5):783–787. doi:10.3168/jds.S0022-0302(77)83935-0

Heslot, N., H.P. Yang, M.E. Sorrells, and J.L. Jannink. 2012. Genomic selection in plant breeding: A comparison of models. Crop Sci. 52:146–160. doi:10.2135/cropsci2011.06.0297

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 Sci. 52(2):654–663. doi:10.2135/cropsci2011.07.0358

Jarquín, D., J. Crossa, X. Lacaze, P. Cheyron, J. Daucourt, J. Lorgeou, et al. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127(3):595–607. doi:10.1007/s00122-013-2243-1

Kane, M.J., J. Emerson, and S. Weston. 2013. Scalable strategies for computing with massive data. J. Stat. Softw. 55(14):1–19. doi:10.18637/jss.v055.i14

Kerr, R.J., L. Li, B. Tier, G.W. Dutkowski, and T.A. McRae. 2012. Use of the numerator relationship matrix in genetic analysis of autopolyploid species. Theor. Appl. Genet. 124(7):1271–1282. doi:10.1007/s00122-012-1785-y

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–4663. doi:10.3168/jds.2009-2061

López-Cruz, M., J. Crossa, D. Bonnett, S. Dreisigacker, J. Poland, J.-L. Jannink, et al. 2015. Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3 (Bethesda) 5:569–582. doi:10.1534/g3.114.016097

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

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648–4655. doi:10.3168/jds.2009-2064

Nejati-Javaremi, A., C. Smith, and J.P. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. 75:1738–1745. doi:10.2527/1997.7571738x

Pérez Rodríguez, P., J. Crossa, K. Bondalapati, G. De Meyer, F. Pita, and G. de los Campos. 2015. A pedigree reaction norm model for prediction of cotton (*Gossypium* sp.) yield in multi-environment trials. Crop Sci. 55:1143–1151. doi:10.2135/cropsci2014.08.0577

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

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/ (accessed 5 Apr. 2017).

Ray, D.K., N. Ramankutty, N.D. Mueller, P.C. West, and J.A. Foley. 2012. Recent patterns of crop yield growth and stagnation. Nat. Commun. 3:1293. doi:10.1038/ncomms2296

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

Weber, V.S., A.E. Melchinger, C. Magorokosho, D. Makumbi, M. Bänziger, and G.N. Atlin. 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. doi:10.2135/cropsci2011.09.0486