Published online September 14, 2017

RESEARCH

Genomic Prediction in a Large African Maize Population

Vahid Edriss, Yanxin Gao, Xuecai Zhang, MacDonald Bright Jumbo, Dan Makumbi, Michael Scott Olsen, José Crossa, Kevin C. Packard, and Jean-Luc Jannink*

ABSTRACT

Genomic prediction (GP) combines genomewide marker data with phenotypic data in a training population to predict the genomic estimated breeding values of untested individuals in a relevant testing population. Our objective was to evaluate the effects of population structure, genotype × trial, tester, and management interactions, and imputation methods on the accuracy of GP for grain yield in the CIMMYT's African maize (*Zea mays* L.) program. The dataset included 2022 diverse breeding lines in 156 Stage 1 yield trials and 66,000 single-nucleotide polymorphism markers. The first two principal components from principal component analysis explained 10.5% of the variance in marker data. Based on marker data, five clusters were detected, but cluster of origin explained only 2% of the phenotypic variation. Prediction accuracy, assessed by cross validation, ranged from 0.20 to 0.36 within clusters and from 0.04 to 0.26 across clusters. Mean GP accuracy within clusters (0.27) outperformed pedigree-based prediction (0.03). Imputation methods did not strongly affect prediction accuracy. Testers and management had large effects. To achieve acceptable GP accuracy within such a diverse population, one can employ (i) a very large training population size, (ii) carefully planned and relevant testers, and (iii) common trial environments and management between the training and validation populations and related genetic materials.

V. Edriss and J.-L. Jannink, Plant Breeding and Genetics Section, School of Integrative Plant Science, 240 Emerson Hall, Cornell Univ., Ithaca, New York 14853; V. Edriss, current address, Nordic Seed, Grindsnabevej 25, Odder, 8300, Denmark; Y. Gao, Genomic Open-Source Breeding Informatics Initiative (GOBII) and Institute of Biotechnology, 608 Frank H. T. Rhodes Hall, Cornell Univ., Ithaca, NY 14853; X. Zhang and J. Crossa, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600, Mexico, DF, Mexico; M.B. Jumbo, D. Makumbi, and M.S. Olsen, International Maize and Wheat Improvement Center (CIMMYT), PO Box 1041-00621, Nairobi, Kenya; K.C. Packard, Cornell Statistical Consulting Unit, B13 Savage Hall, Cornell Univ., Ithaca, NY 14853; J.-L. Jannink, USDA-ARS, R.W. Center for Agriculture and Health, Ithaca, NY 14853. Received 30 Aug. 2016. Accepted 12 June 2017. *Corresponding author (jeanluc.work@gmail.com). Assigned to Associate Editor Manjit Kang.

Abbreviations: AC, across clusters; BLUP, best linear unbiased predictions; COP, coefficient of parentage; CV, cross validation; G × E, genotype × environment; GBS, genotyping by sequencing; GEBV, genomic estimated breeding values; GP, genomic prediction; GS, genomic selection; GY, grain yield; MAF, minor allele frequency; MAS, marker-assisted selection; OPV, open-pollinated variety; PCA, principal component analysis; PEV, prediction error variance; PS, phenotypic selection; QPM, quality protein maize; REML, restricted maximum likelihood; rrBLUP, ridge regression best linear unbiased predictions; SCA, specific combining ability; SNPs, single nucleotide polymorphisms; WC, within cluster.

Genomic prediction (GP) combines genomewide marker data with phenotypic and pedigree data (when available) in a training dataset to predict the genomic estimated breeding values (GEBVs) of untested individuals in a candidate dataset (Meuwissen et al., 2001; Heffner et al., 2009; Wray et al., 2013). Genomic selection (GS) uses whole-genome molecular markers to predict and select individuals with top-ranking GEBVs from a selection

Published in Crop Sci. 57:2361–2371 (2017).
doi: 10.2135/cropsci2016.08.0715

population, where individuals are not phenotyped but only genotyped (Meuwissen et al., 2001; Heffner et al., 2011; Zhang et al., 2015). Factors affecting the accuracy of GS include training population size, relatedness between the training and selection population, marker density, trait heritability, genetic architecture, and other factors that are interrelated with trait architecture (Goddard, 2009; Albrecht et al., 2011; Desta and Ortiz, 2014). In contrast with traditional phenotypic selection (PS), GS offers the advantage of enabling selection prior to phenotyping, saving costs, and has the potential to greatly accelerate genetic gains (Schaeffer, 2006; Heffner et al., 2010, 2011; García-Ruiz et al, 2016). Genomic selection was found to be more accurate than marker-assisted selection (MAS) (Heffner et al., 2011; Arruda et al., 2016). A GP accuracy derived from the literature of 0.53 would correspond to a threefold annual genetic gain in maize (*Zea mays* L.) and a twofold gain in wheat (*Triticum aestivum* L.) relative to MAS (Heffner et al., 2010). The advantage of GS over MAS and PS is particularly true for traits that are expensive to measure, take a long time to measure, or that have low heritability (Desta and Ortiz, 2014). Another advantage of GS over typical MAS is that it does not require identification or validation of association between specific markers and target traits (Arruda et al., 2016), which is a very time-consuming and expensive process in many crop species (Crossa et al., 2011). As marker technology has continuously reduced the cost per data point, the number of available markers has dramatically increased. Genomic selection has been routinely applied in large commercial breeding programs (Endelman et al., 2014) but has lagged in the public sector as limited by resources (Jonas and de Koning, 2016). On the other hand, difficulties that hinder PS accuracy also pose similar challenges to GS accuracy. These include population structure (Lorenz et al., 2012; Riedelsheimer et al., 2013), genotype × environment (G × E) interaction (Comstock and Moll, 1963; Crossa et al., 2014), and optimization of resource allocation in different stages of testing (Lorenz, 2013; Ly et al., 2013). Although overall diversity is desirable in a breeding population, for optimal prediction accuracy, the testing population should be as related to the training population as possible (Habier et al., 2010; Clark et al., 2012; Windhausen et al., 2012). Early stages of testing, such as first-year Stage 1 yield trial at CIMMYT, often involve testcrossing a large number of inbred lines derived from genetically diverse parents and populations onto different testers in different environments (Albrecht et al., 2011; Riedelsheimer et al., 2012; Wu et al., 2015). As part of a routine varietal development pipeline, Stage 1 testing focuses on selecting for high general combining ability and GEBVs, whereas Stage 2 or 3 trials focus on specific combining ability (SCA) and matching hybrids to the environments in which they perform best. To achieve accelerated breeding, resource allocation needs to be optimized between efficiently screening a large number of breeding lines in Stage 1 trials and testing a small number of advanced individuals on multiple testers and in multiple environments for adaptability in Stage 2 and 3 trials (Lorenz, 2013; Jonas and de Koning, 2016).

Another challenge for implementing GP in a large population is marker density and data quality (Crossa et al., 2013; Zhang et al., 2015). Although it is a relatively inexpensive high-density marker technology, genotyping by sequencing (GBS) (Elshire et al., 2011; Poland et al., 2012) is also associated with large quantities of missing data, which arises from no- or low-depth tag sequencing, because of the random sequencing of all tags available in the sequencing library. Different methods can be used to impute missing data, including haplotype hidden Markov models implemented in the software Beagle (Browning and Browning, 2007). Furthermore, multiple imputation methods have been reported to affect (Hickey et al., 2012; Rutkoski et al., 2013) and not affect (Crossa et al., 2013) GP accuracy.

Despite the potential benefits of GP in breeding programs, prediction accuracy studies using extensive empirical multiyear and multilocation, early stage of yield trials involving diverse and complex population structure, testers, and management trial datasets remain quite limited in the public maize breeding programs (Jonas and de Koning, 2016). Therefore, this study was initiated with the following objectives: (i) to investigate GP accuracies in a genetically diverse and structured population for grain yield (GY) in CIMMYT's Africa maize Stage 1 testing program, (ii) to study differences between marker- and pedigree-based prediction on prediction accuracies, (iii) to evaluate effect of marker imputation methods on GP accuracies, and (iv) to determine prediction accuracy as affected by interactions of lines with different trials, testers, and management regimes.

MATERIALS AND METHODS

Phenotypic Data

The experimental data in this study consisted of the first-year yield trial results, obtained between 2007 and 2011, of 2022 maize breeding lines in CIMMYT's African maize breeding program in Kenya. In these early stages of evaluation, breeding lines were testcrossed to one to several testers, and the hybrids derived were placed into 47 trial series, totaling 156 trials, which were set up according to genetic improvement objectives, source population, and pedigree derivation history, maturity, and testers (Supplemental Table 1). The number of lines tested in each trial ranged from 6 to 448, with a median of 40 lines. All maize lines were evaluated in hybrid combinations by testcrossing them with 1 to 15 CIMMYT testers (advanced lines or F_1 crosses between two advanced lines). The trials were planted in 18 different locations in Kenya, Tanzania, Ethiopia, and Uganda under three different management regimes: optimal, drought, and low nitrogen. A minimum of two replications of each entry within each trial were planted under managed drought, low nitrogen, and optimal conditions. Different phenotypic traits were recorded at different locations and trials. Among them, GY was evaluated in all trials and locations, which is summarized and reported here.

Marker Data

DNA was extracted from the leaf tissue of each of the 2022 maize lines and subsequently genotyped using GBS. A GBS protocol commonly used by the maize research community was applied in this study (Elshire et al., 2011). Briefly, GBS libraries were constructed in 96-plex, and genomic DNA was digested with the restriction enzyme ApeK1. Each library was sequenced on a single lane of Illumina flow cell (Cornell Life Science Core Laboratory Center, Ithaca, NY). To increase the genome coverage and read depth for SNP discovery, raw read data from the sequencing samples were analyzed together with an additional 30,000 global maize accessions (Crossa et al., 2013). We used TASSEL 4.0 SNP GBS Discovery Pipeline, with B73 as the reference genome (Glaubitz et al., 2014) to identify SNPs. Initially, 955,690 SNPs were generated for each line; markers with >50% missing scores were discarded from the dataset. After filtering for minor allele frequency (MAF) of >0.01, a subset of 65,995 SNPs remained as the unimputed dataset, with mean missing rate of 36.6% and mean MAF of 0.15.

Imputation Methods

Before GP was applied to the filtered marker datasets, missing markers were imputed with three imputation methods: (i) replace the missing genotypes of each marker with its population expectation (EX-POP), which is simple and computationally fast; (ii) impute using an expectation–maximization algorithm implemented in the ridge regression best linear unbiased predictions (rrBLUP) package in R (Endelman, 2011); and (iii) impute using Beagle 3.3 software (Browning and Browning, 2007) with default parameters, where each chromosome was imputed independently. First, haplotypes were reconstructed with default parameter values. After that, based on the inferred haplotypes, missing genotypes were imputed using a hidden Markov model.

Statistical Analyses

Similar to Ly et al. (2013), a two-step approach to GP was used in this study. First, raw phenotypes were corrected by partitioning genetic and environmental effects to calculate the deregressed best linear unbiased predictions (BLUP). Second, the deregressed BLUP ($y^\star$) was used as the response variable to calculate the GEBV. Prediction accuracy is the correlation between GEBV and deregressed BLUP ($y^\star$) in the validation set and the bias is the regression of GEBV on deregressed BLUP ($y^\star$). We fitted a linear mixed model to estimate BLUP for each line to correct environmental effect on phenotypes, as shown below:

$$y_{ijkl} = \mu + g_i + t_j + r_{k(j)} + s_l + e_{ijkl} \quad [1]$$

where y_{ijkl} is phenotype (i.e., GY), μ is the overall mean and the only fixed effect in the model, g_i is the random genetic effect of the ith maize line, t_j is the random effect of the jth trial, $r_{k(j)}$ is the random effect of the replication within a trial, s_l is the random effect of the lth tester, and e_{ijkl} is the residual effect that includes the genotype × trial interaction. A management effect was not included in the model as a separate variable but is considered part of the trial effect because management and trial effects were confounded with each other. Mixed model analysis was performed using the *lme4* package version 1.1 in R (R Core Team, 2015). Variance components were estimated via restricted maximum likelihood (REML). Plot-based broad-sense heritability (H^2) was calculated from these variance components as the ratio of genotypic variance to the total phenotypic variance. Total phenotypic variance is given by the sum of all variance components ($\sigma_g^2 + \sigma_r^2 + \sigma_t^2 + \sigma_s^2 + \sigma_e^2$, where σ_g^2 is genotypic variance, σ_r^2 is replication variance, σ_t^2 is trial variance, σ_s^2 is tester variance, and σ_e^2 is residual variance). The dataset was considered unbalanced because the 2022 lines were not replicated equally across all the 156 trials. The prediction error variance (PEV) for each line depended on the number of times each line was replicated. According to Rincent et al. (2012), the relation between replication number and PEV is negative; lines with many replications have lower PEVs and are less shrunken toward the mean than those with few replications. To overcome this issue, BLUPs were deregressed on the basis of PEV (Garrick et al., 2009). The deregressed BLUPs were calculated as:

$$y_i^\star = \frac{\hat{g}_i}{1 - \frac{\mathrm{PEV}_i}{\hat{\sigma}_g^2}} \quad [2]$$

where $y_i^\star$ is the deregressed BLUP of the ith line, $\hat{g}_i$ is the BLUP of the ith line, PEV_i is the prediction error variance of the ith line, and $\hat{\sigma}_g^2$ is the total genetic variance. Both PEV_i and $\hat{\sigma}_g^2$ were obtained from Eq. [1].

All marker effects were estimated simultaneously (Meuwissen et al., 2001), as described by Endelman (2011), using the rrBLUP package in R. The following model was used to obtain estimates of the marker effects:

$$\mathbf{y}^\star = \mu + \mathbf{Za} + e \quad [3]$$

where $\mathbf{y}^\star$ is a $n \times 1$ vector of deregressed BLUPs for GY estimated across trials; n is the number of lines; μ is the overall mean; **Z** is an identity matrix; **a** is the vector of additive genetic effects $\mathbf{a} \sim N(0, \mathbf{K}\sigma_a^2)$, in which **K** is the realized additive relationship matrix, estimated using the A.mat function in the rrBLUP package in R (Endelman, 2011); and e is random error [$e \sim N(0, \mathbf{W}\sigma_e^2)$, where **W** is the diagonal weight matrix calculated from the lines' PEV (PEV_i) from Eq. [1]. To calculate the **W** matrix, the reliability of each line was obtained via $r_i^2 = 1 - (\mathrm{PEV}_i/\sigma_g^2)$ so that the weight for the ith line was calculated as $w_i = r_i^2 / \left(1 - r_i^2\right)$.

Population Structure

To visualize the relatedness and potential subpopulation structure of the 2022 maize breeding lines used in this study, the kinship matrix (**K**) used in Eq. [3] was decomposed by principal component analysis (PCA) and the first two principal components were plotted. To determine subpopulation clusters, k-means clustering was applied to the kinship matrix using the kmeans function implemented in R (R Core Team, 2015), which minimizes the distance between lines within clusters and maximizes that across clusters (AC; Saatchi et al., 2011). The number of clusters was determined according to Caliński and Harabasz (1974) using the R package NbClust.

The pedigree of all the 2022 lines was traced back for four generations. Pedigree-based relationships were calculated using the BROWSE software (McLaren et al., 2005; McLaren, 2008). The entries of this numerical relation matrix obtained from the pedigree equal to twice the coefficient of parentage (COP) between pairs of lines. Note that some of the lines were derived from open-pollinated varieties (OPVs) as parents and therefore had no specific pedigree. To account for the relatedness of pairs of lines derived from the same OPV, a COP of 0.05 was assigned, whereas 0 was assigned to pairs of lines derived from different OPVs or from other sources. Prediction of the deregressed BLUPs using the pedigree data was performed by replacing the genomic marker-based matrix $\mathbf{K}$ in Eq. [3] with a pedigree-derived numerical relation matrix.

The impact of population structure on prediction accuracy was measured by two cross-validation (CV) strategies: within cluster (WC), where all the observations within a subpopulation cluster were randomly divided into training (70%) and validation (30%) datasets, and AC, where one of the subpopulation clusters was used as a validation dataset and the remaining four clusters were used as a training dataset (Table 1). The average values of the correlations between the GEBVs and deregressed BLUP values ($y^\star$) in the validation set from 50 runs were calculated and defined as the prediction accuracies for within cluster (rWC) and for across clusters (rAC) (Table 1). Similar to Zhang et al. (2015), correlation between rWC and H was calculated, where rWC is the GP accuracy and H is the square root of the plot-based broad-sense heritability (H^2).

Effect of Line Interaction with Trial, Tester, and Management on Accuracy

When lines in the training and validation populations are evaluated in the same environments, GP accuracy can be inflated by interaction between lines and environments (Ly et al., 2013). This occurs because lines by environments interaction is shared between training and validation observations, even though they will generally not be reproduced in future environments (Lorenz et al., 2011). To estimate the size of this bias, only trials with >60 lines were included in the CV. Similar to Spindel et al. (2015), we altered the training population composition by including or excluding common testing units between the training and validation datasets. We took one trial at a time as a focal trial and used the rest of the trials to predict the focal trial. In the focal trial, 35 lines were randomly picked as validation dataset, whereas observations of the remaining entries from the focal trial were either included or removed from the training dataset, corresponding to two CV schemes (CV1 and CV2). In CV1, observations from the focal trial were excluded from the training dataset, whereas in CV2, those observations were included in the training dataset, but observations of corresponding lines in other trials were removed from the training dataset so that the amount of phenotypic data stayed the same in the training set between CV1 and CV2. For both validation schemes, this procedure was repeated 50 times for the focal trial, and the mean of 50 prediction accuracies was calculated. An estimate of the bias caused by line × trial interaction was obtained from the difference between CV1 and CV2 accuracy (Fig. 1). The same procedure was applied to testers and managements. The number of lines crossed to each tester ranged

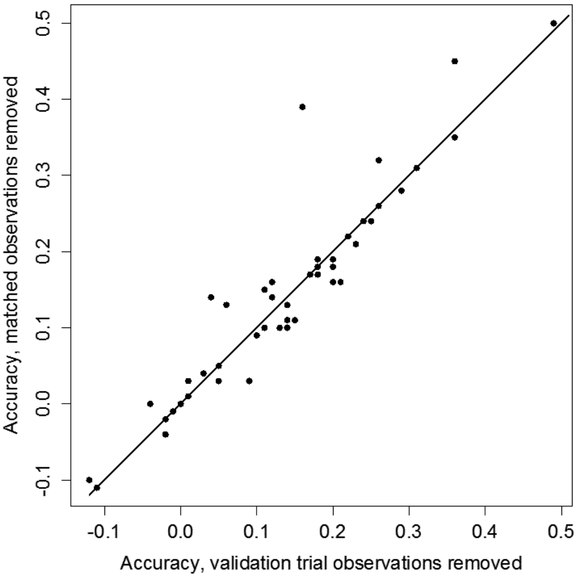

Fig. 1. Prediction accuracy with two cross-validation schemes (CV1 on *x*-axis and CV2 on *y*-axis) for trials. In CV1, observations from the focal trial were excluded from the training dataset, whereas in CV2, observations from the focal trial were included in the training dataset, but matched observation from other trials were removed

Table 1. Genetic variance, heritability (H^2), and the accuracies for within- (rWC) and across-cluster (rAC) predictions.

Cluster	Origins	No. of lines	Mean grain yield	Pedigree-based	rWC†	rAC†	Genetic variance	H^2
			t ha^{-1}					
1	QPM‡	182	3.84	0.08	0.29	0.26	0.03	0.15
2	Latin America	422	4.44	−0.06	0.24	0.20	0.04	0.15
3	Zimbabwe	606	4.80	−0.03	0.33	0.06	0.11	0.21
4	Kenya	403	5.63	0.02	0.36	0.15	0.14	0.22
5	IITA	409	5.30	0.03	0.20	0.04	0.15	0.24
All lines		2022		−0.03	0.27			0.22
Mean			4.80	0.01	0.28	0.14		0.19

† The average values of the correlations between the phenotype and the genomic estimated breeding values from 50 runs were calculated and defined as the prediction accuracies within clusters (rWC) and across clusters (rAC).

‡ QPM, quality protein maize.

between 6 and 1339, with a median of 67. The CV (CV1 and CV2) was applied to 9 of the 15 testers that had >60 lines per tester and to the three managements. Estimate of tester (Table 2) or management (Table 3) effects was obtained from the difference between CV1 and CV2 prediction accuracy.

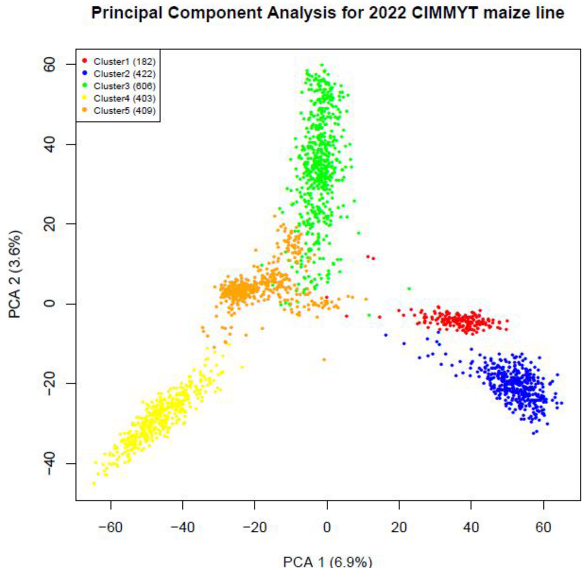

Fig. 2. Population and subpopulation structure as plotted by the first two principal components of principal component analysis (PCA) and *K*-means clustering. Five clusters were determined using the index given by Caliński and Harabasz (1974). Each dot represents one maize line, colors are determined via *K*-means clusters, and cluster sizes are shown in the legend.

RESULTS

Population Structure

The 2022 CIMMYT maize breeding lines evaluated in this study were known for their diverse genetic background (Wu et al., 2015), origins (tropical, subtropical, tropical × temperate, or OPV), type of crosses (three-way, four-way, synthetic, or doubled haploid), and selection history (multiple biotic and abiotic stresses prevalent in Africa). Therefore, some population structure was expected. From the PCA on the realized genomic relationship matrix, the first two principal components explained 6.9 and 3.6% of the total marker variation, respectively (Fig. 2). The optimum number of clusters was determined to be five according to Caliński and Harabasz (1974) index for *k*-means clustering. Cluster sizes ranged from 182 to 606 lines per cluster (Table 1). Further, cluster of line of origin analysis tracked the breeding program that majority of the lines originated from: Cluster 1 contained breeding lines mostly belonging to CIMMYT's quality protein maize (QPM) program, Cluster 2 lines originated mainly from the Latin American maize breeding program, Cluster 3 lines mainly belonged to the Zimbabwean maize breeding program, Cluster 4 lines were mostly derived from the Kenyan maize breeding program, and Cluster 5 lines traced back to the maize breeding program of the International Institute of Tropical Agriculture (IITA) in Nigeria. Clusters 1 (QPM) and 2 (Latin America) were more closely related to each other than to Cluster 3 (Zimbabwe), Cluster 5 (IITA), and Cluster 4 (Kenya). Kenyan lines were more closely related to Zimbabwean and IITA lines than the Zimbabwean and IITA lines were related with each other. Mean GY by cluster ranked as follows: Kenya (5.63 t ha^{-1}) > IITA (5.30 t ha^{-1}) > Zimbabwe (4.80 t ha^{-1})

Table 2. Number of lines and observations tested for each tester, genomic prediction accuracy with two cross-validation schemes (CV1 and CV2). In CV1, observations from the focal tester were excluded from the training dataset, whereas in CV2, those observations were included in the training dataset.

Testers	No. of lines	No. of observations	CV1	CV2	Difference between CV1 and CV2	Classification of testers per accuracy by CV1 and CV2
1. CML440	67	458	0.36	0.36	0.00	Similar, high
2. CML395 × CML444	1,339	10,486	0.26	0.27	0.01	Similar, high
3. CML144	61	244	0.15	0.15	0.00	Similar
4. CML144 × CML159	68	723	0.06	0.08	0.02	Similar, low
5. CML159	67	354	0.04	0.05	0.01	Similar, low
6. CML312 × CML442	916	6,300	0.18	0.31	0.13	Different, high
7. CML202 × CML395	360	1,882	0.09	0.33	0.24	Different
8. CML445	230	1,001	−0.08	0.06	0.14	Different, low
9. ECA-EE-55	154	1,820	−0.03	0.05	0.07	Different, low
Mean			0.12	0.18	0.06	Different

Table 3. Number of lines and observations for each management, genomic prediction accuracy with two cross-validation schemes (CV1 and CV2). In CV1, observations from the focal management were excluded from the training dataset, whereas in CV2, those observations were included in the training dataset.

Management	No. of lines	No. of observations	CV1	CV2	Difference between CV1 and CV2
Optimal	2,022	16,827	0.07	0.24	0.17
Drought	1,911	7,744	−0.02	0.13	0.15
Low nitrogen	357	992	−0.01	0.01	0.02
Mean			0.01	0.13	0.12

> Latin America (4.44 t ha^{-1}) > QPM (3.84 t ha^{-1}), with an overall mean of (5.08 t ha^{-1}). Variance components estimated via REML were as follows: genotypic variance (σ^2_g) = 0.3401, replication variance (σ^2_r) = 0.4189, trial variance (σ^2_t) = 2.8135, tester variance (σ^2_s) = 0.7394, and residual variance (σ^2_e) = 1.2326. Cluster variance was 0.11, meaning only 2% of the total phenotypic variance for GY was contributed by the clusters of origin. Plot-based broad-sense heritability for GY ranged from 0.15 in Clusters 1 (QPM) and 2 (Latin America) to 0.24 in Cluster 5 (IITA), with overall heritability across all lines being 0.22 and a mean heritability being 0.19 across all clusters.

Genomic Prediction and Pedigree-Based Prediction Accuracies

To verify that markers captured information beyond pedigree relatedness in these predictions, we also performed pedigree-based predictions, where the marker-based kinship matrix **K** in Eq. [3] was replaced with a pedigree-based matrix. Pedigree-based prediction accuracies were not much different from zero, ranging from −0.06 in Cluster 2 lines (Latin America) to 0.075 in Cluster 1 lines (QPM), with a mean of all lines of −0.03 (Table 1), which was lower than the WC (0.27) and AC means (0.14). Clearly, GP outperformed pedigree-based prediction.

Attributing its cluster mean of 0.14 to each hybrid would thus generate a cluster mean prediction accuracy of 0.02 (i.e., 0.14 × 0.14 ≈ 2%); however, we observed a prediction accuracy of 0.27. In other words, GP accuracy in our study was almost twice as high as cluster mean prediction accuracy, indicating an advantage of GP over cluster mean prediction, validating that markers captured more than population structure effects. In summary, marker-based GP outperformed cluster mean-based prediction and was greatly improved over pedigree-based prediction.

Within- and Across-Cluster Genomic Prediction

Genomic prediction accuracy within each cluster was calculated using a random CV method (i.e., the mean accuracy of 50 times), splitting 70% of all observations randomly from a cluster to predict GEBVs of the remaining 30% lines within the same cluster. Among the five clusters identified, Cluster 4 (Kenya) had the highest GP accuracy (0.36), whereas Cluster 5 (IITA) had the lowest accuracy (0.20) (Table 1). The accuracy across all lines was 0.27, close to the average of the five clusters' WC accuracy of 0.28. The standard error for 50 repeated CV prediction accuracies was low, ranging from 0.01 to 0.03 (data not shown), with the highest value being for Cluster 4 lines (Kenya) and the lowest for Cluster 2 lines (Latin America), indicating that variation among the 50 repeated, random CVs did not affect the reported accuracies much. The WC prediction accuracy did not appear to correlate with broad-sense heritability, nor with cluster size.

The AC GP accuracies were calculated by leaving one cluster out and using the remaining four clusters as training set. We found that, for all clusters, AC prediction accuracy was lower than WC prediction accuracy. However, Clusters 1 (QPM, 0.26) and 2 (Latin America, 0.20) had accuracy close to WC prediction accuracies. Cluster 3 (Zimbabwe, 0.06) and 5 (IITA, 0.03) accuracies were close to zero (Table 1). The higher across-cluster accuracies for the QPM and Latin American clusters probably came from the similarity of the two clusters (Fig. 2), indicating that they predict each other reasonably well.

Imputation Method Effect on Accuracy

Initially, 955,690 SNPs were generated for each line by GBS; a filtered subset of 65,995 SNPs, with a missing rate of <50% and minimum MAF >0.01, was used for the GP in this study. The mean missing rate was 36.6%, where the missing pattern was random and the mean MAF was 0.15 in the unimputed 65,955 SNP dataset. EX-POP, Beagle (Browning and Browning, 2007), and expectation-maximization (Endelman, 2011) were not only used to impute missing marker data points but also to evaluate the impact of missing data points on GP accuracy. Imputation using Beagle was performed for each chromosome independently, and haplotypes were first reconstructed with default parameter values. Conditional on the inferred haplotypes, missing genotypes were then imputed using a hidden Markov model. In general, little difference was found between imputation methods for WC prediction accuracy or across the total dataset, as shown in Fig. 3. The expectation-maximization method had the highest or second highest accuracy of all clusters. EX-POP and the Beagle methods had the highest difference in Cluster 3 (Beagle was higher by 3.1%). The ranking of computational time was EX-POP < Beagle < expectation-maximization methods.

Differences among imputation methods for AC GP accuracy were small but slightly bigger than those of WC prediction (Fig. 3). Again, the highest difference was between EX-POP and Beagle methods (Beagle was higher by 5.1%) in Cluster 3. Results similar to those for WC prediction were obtained with AC, with the expectation-maximization method tracking very closely the prediction accuracy of EX-POP, whereas Beagle showed more variability than the other two methods. Nevertheless, all imputation methods affected prediction accuracy very similarly, consistent with other studies (Weng et al., 2012). As there was not much difference between imputation methods, the EX-POP method was chosen to impute for the rest of the analysis in the study because of its low computation time.

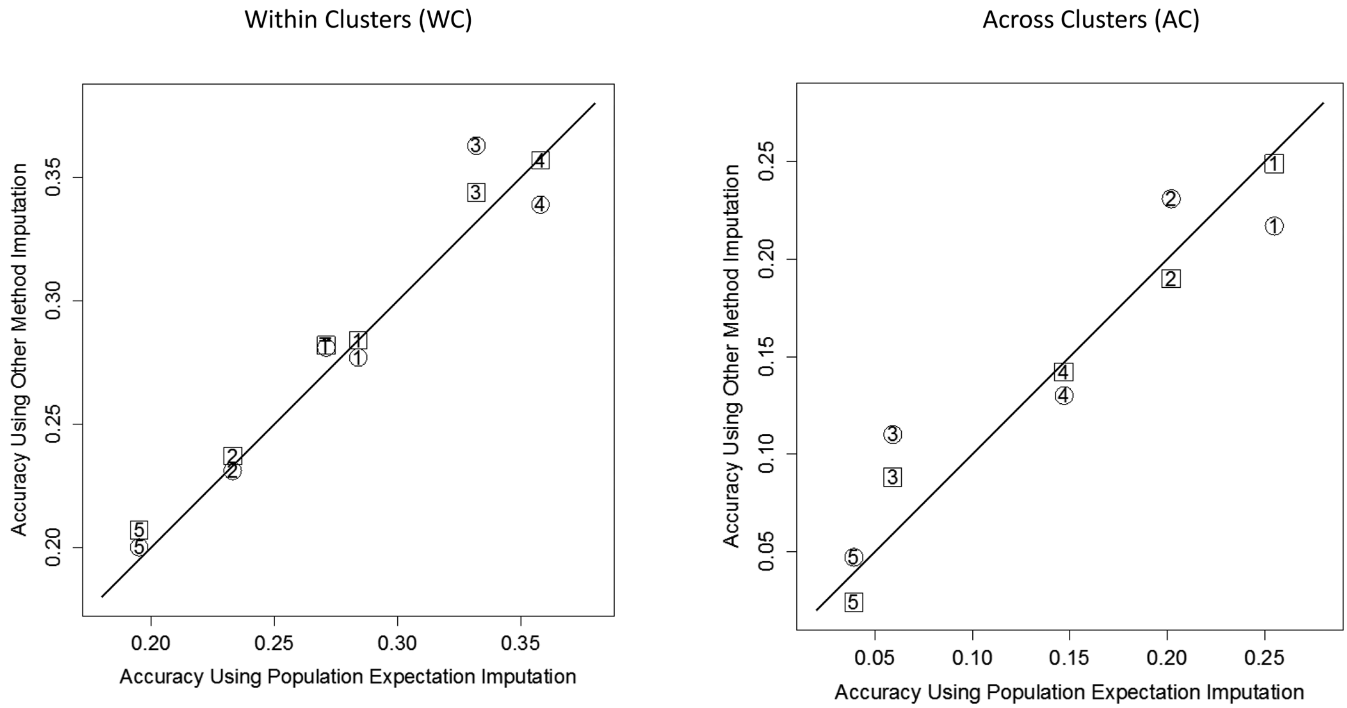

Fig. 3. The accuracy of genomic prediction within clusters (left) and across clusters (right) using Beagle (circles) or expectation-maximization (squares) plotted against accuracies when missing marker data points were set to the cluster mean for their marker. Numbers in the symbol indicate the cluster. "T" indicates the total dataset.

Impact of Trials, Testers, and Managements on Prediction Accuracy

To evaluate the prediction accuracy as affected by different trials, testers, and managements, we adopted a CV scheme (CV1 and CV2) similar to Spindel et al. (2015). The trial sizes ranged from 61 to 233 lines (trials with <60 lines were excluded from these CV analyses). The number of phenotypic observations per trial ranged from 122 to 496, with each line replicated two to three times in each trial. The prediction accuracy for CV1 and CV2 differed very little (Fig. 1), averaging 0.13 and 0.14, respectively, having observations from a particular trial biased the accuracy upward by only 1%. A paired *t* test on the accuracies of CV1 and CV2 gave a *p*-value of 0.034. The assumptions of the *t* test were not met in this analysis because the trials were not independent. Nevertheless, this low *p*-value suggested that, while small, there might be a genuine upward bias between CV1 and CV2 for trials.

The same approach was used to estimate the effect of line × tester interaction on prediction accuracy. In CV1, lines crossed to the focal tester were excluded from the training dataset, whereas in CV2, those observations remained, but matched lines crossed to other testers were removed from the training dataset so that the amount of phenotypic data stayed the same in the training set between CV1 and CV2. The number of lines and observations ranged from 61 to 1339 lines and 244 to 10,486 observations (Table 2). The mean accuracy for CV1 and CV2 was 0.12 and 0.18, respectively. Clearly, the effect of line × tester interaction was larger than the line × trial interaction, suggesting that evaluating lines in hybrid combination with relevant testers was important for GP accuracy.

Among the nine testers evaluated (Table 2), CV2 and CV1 had no bias for five testers, where Testers 1 and 2 had relatively high accuracy (0.36 or 0.26, respectively). Testers 4 and 5 had low accuracy (0.06 and 0.08, respectively), and Tester 3 had intermediate prediction accuracy (i.e., 0.15 for both CV1 and CV2). For the other four testers (Testers 6–9), CV2 had higher prediction than CV1, which indicated the interaction between testers and lines affected prediction accuracy. Tester 6 almost doubled the prediction accuracy when validated by CV2 (0.31) versus CV1 (0.18); for Tester 7, CV2 (0.33) was almost four times larger than CV1 (0.09), and for Testers 8 and 9, CV1s were negative and CV2s were low (0.06 or 0.05). These results indicated that including or excluding common testing units between training and validation datasets made a big difference in prediction accuracy for Testers 6 and 7 but made little difference for Testers 8 and 9.

Our results have demonstrated that testers themselves had a major impact on prediction accuracy, where Testers 1, 2, 3, and 6 had medium to high prediction accuracy and Testers 4, 5, 8, and 9 had very weak prediction accuracy across both CV1 and CV2 validation schemes. Therefore, choosing the right testers and evaluating lines in hybrid combinations with relevant testers are very critical for routine implementation of GS in Stage 1 testing in the CIMMYT East African Maize breeding program.

Finally, we used the CV approach to study the impact of different managements on prediction accuracy. A high number of lines was observed under each management. All lines were tested under optimal conditions, with 16,827 observations. Optimal management received the biggest advantage from CV2, with a difference of 0.17 in the GP accuracy over CV1 (Table 3). Performance under drought or low nitrogen could not be predicted without observations from those management practices, as shown by nonsignificant negative correlations between prediction and observations under CV1. Prediction for low nitrogen was generally unsuccessful. Even when low-nitrogen observations were available in the training dataset (CV2), the accuracy was not different from zero (Table 3).

DISCUSSION

Stage 1 yield trials in the CIMMYT maize breeding program involve testcrossing many diverse breeding lines to multiple testers and evaluating the testcrosses across multiple environments and management regimes. Depending on the traits' genetic architecture and selection intensity, yield trials and PS can be very expensive with low gain from selection per unit of investment. As genotyping costs have continued to decrease, training dataset sizes have tended to increase, and training models have been optimized, which, in turn, have improved prediction accuracies. In this study, we evaluated the feasibility of GS in the CIMMYT East African maize breeding program, using (i) a phenotypic dataset that consisted of GY observations from hybrids generated from 2022 diverse breeding lines obtained from 156 Stage 1 yield trials conducted across multiple testers, environments, and managements; and (ii) a genotypic dataset that consisted of 2022 breeding lines' DNA samples × 65,995 unimputed GBS SNP marker matrix calls. Overall, the prediction accuracies in this study were low: WC accuracies ranged from 0.2 to 0.36, AC accuracies ranged from 0.04 to 0.26, and WC pedigree-based accuracies ranged from −0.06 to 0.08. Despite the prediction accuracies being low, the CV results across different clusters, imputation methods, testers, and managements were consistent with other maize genomic-prediction studies, including breeding populations (Albrecht et al., 2011; Crossa et al., 2011), biparental and multiparental populations (Guo et al., 2013; Riedelsheimer et al., 2013; Schulz-Streeck et al., 2013), and diversity panels (Riedelsheimer et al., 2012; Rincent et al., 2012). For example, using multiple GS models and environments, prediction accuracy for maize flowering time ranged from 0.46 to 0.79, and for maize GY, it ranged from 0.42 to 0.53 (Crossa et al., 2011). The low level of prediction accuracies in this study could be mostly attributed to the highly diverse population structure in the Stage 1 breeding lines evaluated. As Albrecht et al. (2011) and Hickey et al. (2014) reported, a high degree of relatedness between the training and the validation test set corresponded to high accuracies (0.72–0.74), and distantly related families corresponded to low to intermediate accuracies (0.47–0.48). The CIMMYT germplasm constitutes diverse genetic backgrounds (Crossa et al., 2014; Wu et al., 2015). The diversity of the germplasm evaluated in this study is consistent with that assessment. This diversity is very different from temperate maize that has very well-defined heterotic pools and well-kept selection history and pedigree records (Lu et al., 2009).

Zhang et al. (2015) reported that the ascertainment bias and SNP calling error of GBS increased when B73 was used as reference genome in the GBS production pipeline for tropical germplasm, where allele frequency in temperate maize and tropical maize was different and novel alleles in the tropical maize could have been missed. Using the B73 temperate line as the reference genome to call SNP markers for the tropical germplasm could also cause reduced GBS marker quality. Other factors affecting the prediction accuracy in this study could be variable GY measurements across multiyear and multilocation yield trials and variable training population size and composition. As shown in the realized genomic relationship matrix, the total variation for these diverse lines was so large that the first two principal components together explained 10.5% of the total marker variation. Population structure, as tracked to the clusters of line of origins, explained only 2% of the total phenotypic variation. Although the prediction accuracies in this study were low, they were consistent. Clearly, the results indicated that, for highly diverse breeding lines in the African mid-altitude maize, GS (WC mean 0.27 and AC mean 0.22) was more effective than pedigree-based prediction (mean 0.03), which was consistent with Heffner et al. (2011) and Burgueño et al. (2012). Differences in accuracies in the all clusters between GS and PS were not large, indicating that partially replacing Stage 1 yield trial testing with GS could offer a significant cost reduction, more rapid breeding and selection cycles, and thus higher genetic gain along the lines of the analysis provided by (Heffner et al., 2010). Genomic selection offers great opportunity for optimizing breeding schemes within the same resources, which is especially true for closely related lines (Jonas and de Koning, 2016). A prediction accuracy of 0.36 within Kenyan lines still allows a breeding program to discard the worst lines; for example, discarding the lines ranked in the lowest quartile GEBVs before placing them in expensive yield trials. Savings from partially replacing yield trials with GS in the Stage 1 testing would allow the Kenyan breeding program to screen more lines in the subsequent cycles or to add testers, replications, or environments for SCA in Stage 2 and Stage 3 testing.

Windhausen et al. (2012) and Guo et al. (2013) reported predictive accuracies that were highly affected by population structure when the calibration set comprised genetic groups with significantly different mean performance.

Our study showed that WC was more accurate than AC prediction, which is consistent with the above-cited studies. The average relatedness of the individuals from the training population with those from the validation population has been shown to have a strong effect on prediction accuracy (Habier et al., 2010). Windhausen et al. (2012) found that prediction accuracy was not greatly different from the accuracy that would be obtained by predicting an individual's value with the mean phenotype of the cluster to which the individual belongs, which suggested that individuals belonging to different clusters basically contributed no useful information for prediction in a focal cluster. Our study reached a slightly different result. We found that AC prediction accuracy was >0.2 for two clusters and >0.10 for three clusters. Thus, individuals from each cluster contributed information valuable to the prediction of the other clusters. In the case presented here, only 2% of the variation in GY was explained by cluster of line origin. Attributing its cluster mean to each hybrid would thus generate a prediction accuracy of 0.14 (that is, 0.14 × 0.14 ≈ 2%), whereas we observed an accuracy of 0.27. In other words, GP accuracy in our study was almost twice as high as cluster mean prediction accuracy, indicating that markers captured more than population structure effects. To verify that markers captured information beyond pedigree relatedness in these predictions, we also performed predictions where the pedigree-based relationship matrix replaced the marker-based kinship matrix **K** in Eq. [3]. Accuracies for pedigree-based predictions were close to zero (−0.03 across all lines), which was much lower than marker-based prediction accuracy for all lines. These results agree with many other studies in that markers consistently increased prediction ability over the baseline pedigree-derived model (Vazquez et al., 2010; Heffner et al., 2011; Crossa et al., 2014).

The analyses we performed to evaluate the effect of the presence of data from different trials, testers, and managements indirectly addressed the issue of the effect of G × E interaction on prediction accuracy. The G × E interaction generates a common error component between the predictions and the training estimates based on the observations (Lorenz et al., 2011, 2012; Burgueño et al., 2012), which is attributed to a confounding factor that upwardly biases the prediction accuracy (Ly et al., 2013). With respect to the question of what trials and environments to include in training populations, when validation measurements are taken in environments that were also sampled in training population evaluations, there will be a positive bias in the accuracy that depends on the total number of environments used and on the ratio of the G × E interaction variance to genetic variance.

Predicting the performance of newly developed lines that have never been evaluated in the field (CV1) is more challenging than predicting the performance of lines that have been evaluated in different but correlated environments (CV2) (Crossa et al., 2014). Table 3 contains correlations for two CV schemes (CV1 and CV2). The CV1 predicted unobserved phenotypes of untested lines, whereas CV2 predicted unobserved phenotypes of lines that had been evaluated in some environments but not others. Relative to CV1, correlations in CV2 were 240% greater under optimal and 750% greater under drought management, indicating the importance of having information from correlated environments when predicting performance. For correlated management, higher prediction accuracies can be achieved by borrowing information from correlated trials and environments; for example, the mean correlations in CV2 were 120% greater than those in CV1 (Table 3). Consistent with Crossa et al. (2013) using unrelated populations (CV1) as a training population, the prediction accuracy became negligible for drought and low nitrogen management. When GP includes modeling G × E interaction, an increase in prediction accuracy can be achieved by borrowing information from correlated environments (CV2, Table 3). Our results indicated that to achieve high prediction accuracy, the training dataset for GP should represent the full genetic and environmental spectrum of a breeding program. These results are consistent with Albrecht et al. (2014) and Crossa et al. (2014) in that the optimum training data for GP should represent the full genetic and environmental spectrum of the respective breeding program. Albrecht et al. (2014) reported that data heterogeneity can be reduced by experimental designs that maximize the connectivity between data sources by common or highly related test units. Our prediction results across testers showed that borrowing common testing units between training and validation datasets was important for some testers but not for others (Table 2), which demonstrated that testers themselves had a major impact on prediction accuracy. Overall, the results indicated that tester selection was a very important factor in GP accuracy for Stage 1 yield trials. Also, choosing the right testers and evaluating lines in hybrid combinations with relevant testers are critical for routine implementation of GS in Stage 1 testing in the CIMMYT African maize breeding program.

Agreeing with Heffner et al. (2011) and Jonas and de Koning (2016), GP for GY in our study was much more accurate than pedigree-based prediction and was consistent with PS; hence, GP could possibly or partially replace PS in CIMMYT maize breeding programs. Clearly, the CIMMYT African maize program contained a highly diverse subpopulation structure. However, further improvement in GP accuracies could still be achieved by (i) employing a very large training population size, (ii) correctly choosing relevant testers, and (iii) common trial units between the training and validation populations, including similar environment, management, and related genetics.

This report represents the largest empirical GP accuracy CV case study among public maize breeding programs.

Conflict of Interest

The authors declare that there is no conflict of interest.

Supplemental Material Available

Supplemental material for this article is available online.

Acknowledgments

We thank all collaborators in Africa who were involved in conducting the multienvironment trials providing extensive yield trial phenotypic data. We are grateful to Drs. Julie Dawson and Kassa Semagn for their support during genotyping activities for this study. We acknowledge the financial support provided to the Cornell-CIMMYT Genomic Selection project by the Bill and Melinda Gates Foundation (Gates Global Development Grant no. OPP1016384).

References

Albrecht, T., H.J. Auinger, V. Wimmer, J.O. Ogutu, C. Knaak, M. Ouzunova et al. 2014. Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor. Appl. Genet. 127:1375–1386. doi:10.1007/s00122-014-2305-z

Albrecht, T., V. Wimmer, H.J. Auinger, M. Erbe, C. Knaak, M. Ouzunova et al. 2011. Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123:339–350. doi:10.1007/s00122-011-1587-7

Arruda, M.P., A.E. Lipka, P.J. Brown, A.M. Krill, C. Thurber, G. Brown-Guedira et al. 2016. Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum* L.). Mol. Breed. 36:84. doi:10.1007/s11032-016-0508-5

Browning, S.R., and B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097. doi:10.1086/521987

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. 52:707–719. doi:10.2135/cropsci2011.06.0299

Caliński, T., and J. Harabasz. 1974. A dendrite method for cluster analysis. Commun. Stat. Simul. Comput. 3:1–27. doi:10.1080/03610917408548446

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H.J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4. doi:10.1186/1297-9686-44-4

Comstock, R.E., and R.H. Moll. 1963. Genotype-environment interactions. In: W.D. Hanson and H.F. Robinson, editors, Statistical genetics and plant breeding. Natl. Acad. Sci., Natl. Res. Counc., Washington, DC. p. 164–194.

Crossa, J., Y. Beyene, S. Kassa, P. Pérez-Rodríguez, J.M. Hickey, C. Chen et al. 2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 (Bethesda) 3:1903–1926. doi:10.1534/g3.113.008227

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

Crossa, J., P. Pérez-Rodríguez, J. Hickey, J. Burgueño, L. Ornella, J. Ceron-Rojas et al. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112:48–60. doi:10.1038/hdy.2013.16

Desta, Z.A., and R. Ortiz. 2014. Genomic selection: Genome-wide prediction in plant improvement. Trends Plant Sci. 19:592–601. doi:10.1016/j.tplants.2014.05.006

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379. doi:10.1371/journal.pone.0019379

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R Package rrBLUP. Plant Genome 4:250–255. doi:10.3835/plantgenome2011.08.0024

Endelman, J.B., G.N. Atlin, Y. Beyene, K. Semagn, X. Zhang, M.E. Sorrells, and J.L. Jannink. 2014. Optimal design of preliminary yield trials with genome-wide markers. Crop Sci. 54:48–59. doi:10.2135/cropsci2013.03.0154

García-Ruiz, A., J.B. Cole, P.M. VanRaden, G.R. Wiggans, F.J. Ruiz-López, and C.P. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. USA 113:E3995–E4004. doi:10.1073/pnas.1519061113 [erratum: 113:E3995].

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet., Sel., Evol. 41:55. doi:10.1186/1297-9686-41-55

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One 9:e90346. doi:10.1371/journal.pone.0090346

Goddard, M.E. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica (The Hague) 136:245–257.

Guo, Z., D.M. Tucker, D. Wang, C.J. Basten, E. Ersoz, W.H. Briggs et al. 2013. Accuracy of across-environment genome-wide prediction in maize nested association mapping populations. G3 (Bethesda) 3:263–272. doi:10.1534/g3.112.005066

Habier, D., J. Tetens, F.R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5. doi:10.1186/1297-9686-42-5

Heffner, E.L., J.L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4:65–75. doi:10.3835/plantgenome.2010.12.0029

Heffner, E.L., A.J. Lorenz, J.L. Jannink, and M.E. Sorrells. 2010. Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci. 50:1681–1690. doi:10.2135/cropsci2009.11.0662

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12. doi:10.2135/cropsci2008.08.0512

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. 52:654–663. doi:10.2135/cropsci2011.07.0358

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, B.M. Prasanna et al. 2014. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 54:1476–1488. doi:10.2135/cropsci2013.03.0195

Jonas, E., and D.J. de Koning. 2016. Goals and hurdles for a successful implementation of genomic selection in breeding programme for selected annual and perennial crops. Biotechnol. Genet. Eng. Rev. 32:18–42. doi:10.1080/02648725.2016.1177377

Lorenz, A.J. 2013. Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: A simulation experiment. G3 (Bethesda) 3:481–491. doi:10.1534/g3.112.004911

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata et al. 2011. Genomic selection in plant breeding: Knowledge and prospects. In: D.L. Sparks, editor, Advances in agronomy. Vol. 110. Academic Press, Cambridge, MA. doi:10.1016/B978-0-12-385531-2.00002-5 p. 77–123.

Lorenz, A.J., K.P. Smith, and J. Jannink. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci. 52:1609–1621. doi:10.2135/cropsci2011.09.0503

Lu, Y., J. Yan, C.T. Guimarães, S. Taba, Z. Hao, S. Gao et al. 2009. Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theor. Appl. Genet. 120:93–115. doi:10.1007/s00122-009-1162-7

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch et al. 2013. Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Sci. 53:1312–1325. doi:10.2135/cropsci2012.11.0653

McLaren, C.G. 2008. TDM GMS Browse: The GMS BROWSE application. Int. Crop Inf. Syst. https://cropforge.github.io/iciswiki/articles/t/d/m/TDM_GMS_Browse_53fa.html (accessed 19 July 2017).

McLaren, C.G., R. Bruskiewich, A.M. Portugal, and A.B. Cosico. 2005. The international rice information system. A platform for meta-analysis of rice crop data. Plant Physiol. 139:637–642. doi:10.1104/pp.105.063438

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

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253. doi:10.1371/journal.pone.0032253

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Riedelsheimer, C., J.B. Endelman, M. Stange, M.E. Sorrells, J.L. Jannink, and A.E. Melchinger. 2013. Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503. doi:10.1534/genetics.113.150227

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla et al. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics 192:715–728. doi:10.1534/genetics.112.141473

Rutkoski, J.E., J. Poland, J.L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda) 3:427–439. doi:10.1534/g3.112.005363

Saatchi, M., M.C. McClure, S.D. McKay, M.M. Rolf, J. Kim, J.E. Decker et al. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet. Sel. Evol. 43:40. doi:10.1186/1297-9686-43-40

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223. doi:10.1111/j.1439-0388.2006.00595.x

Schulz-Streeck, T., J.O. Ogutu, and H.-P. Piepho. 2013. Comparisons of single-stage and two-stage approaches to genomic selection. Theor. Appl. Genet. 126:69–82. doi:10.1007/s00122-012-1960-1

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redoña et al. 2015. Genomic selection and association mapping in rice (*Oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet. 11:e1004982. doi:10.1371/journal.pgen.1004982 [erratum: 11:e1005350].

Vazquez, A.I., G.J.M. Rosa, K.A. Weigel, G. de los Campos, D. Gianola, and D.B. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93:5942–5949. doi:10.3168/jds.2010-3335

Weng, Z., Z. Zhang, X. Ding, W. Fu, P. Ma, C. Wang, and Q. Zhang. 2012. Application of imputation methods to genomic selection in Chinese Holstein cattle. J. Anim. Sci. Biotechnol. 3:6. doi:10.1186/2049-1891-3-6

Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J.L. Jannink, M.E. Sorrells et al. 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 (Bethesda) 2:1427–1436. doi:10.1534/g3.112.003699

Wray, N.R., J. Yang, B.J. Hayes, A.L. Price, M.E. Goddard, and P.M. Visscher. 2013. Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. 14:507–515. doi:10.1038/nrg3457

Wu, Y., F. San Vicente, K. Huang, T. Dhliwayo, D.E. Costich, K. Semagn et al. 2015. Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. Theor. Appl. Genet. 129:753–765. doi:10.1007/s00122-016-2664-8

Zhang, X., P. Pérez-Rodríguez, K. Semagn, Y. Beyene, R. Babu, M.A. López-Cruz et al. 2015. Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity 114:291–299. doi:10.1038/hdy.2014.99