Genomic Prediction with Genotype by Environment Interaction Analysis for Kernel Zinc Concentration in Tropical Maize Germplasm

Edna K. Mageto,† Jose Crossa,† Paulino Pérez-Rodríguez,‡ Thanda Dhliwayo,‡ Natalia Palacios-Rojas,‡ Michael Lee,*,1 Rui Guo,‡ Félix San Vicente,‡ Xuecai Zhang,‡ and Vemuri Hindu**

*Department of Agronomy, Iowa State University, Ames, IA 50011, †International Maize and Wheat Improvement Center (CIMMYT), El Batan, Texcoco CP 56237, Mexico, ‡Colegio de Postgraduados, Department of Statistics and Computer Sciences, Montecillos, Edo. De México 56230, México, **Asia Regional Maize Program, International Maize and Wheat Improvement Center (CIMMYT), ICRISAT Campus, Patancheru, Hyderabad, Telangana 502324, India

ABSTRACT Zinc (Zn) deficiency is a major risk factor for human health, affecting about 30% of the world’s population. To study the potential of genomic selection (GS) for maize with increased Zn concentration, an association panel and two doubled haploid (DH) populations were evaluated in three environments. Three genomic prediction models, M (M1: Environment + Line, M2: Environment + Line + Genomic, and M3: Environment + Line + Genomic + Genomic x Environment) incorporating main effects (lines and genomic) and the interaction between genomic and environment (G x E) were assessed to estimate the prediction ability ($r_{MP}$) for each model. Two distinct cross-validation (CV) schemes simulating two genomic prediction breeding scenarios were used. CV1 predicts the performance of newly developed lines, whereas CV2 predicts the performance of lines tested in sparse multi-location trials. Predictions for Zn in CV1 ranged from -0.01 to 0.56 for DH1, 0.04 to 0.50 for DH2 and -0.001 to 0.47 for the association panel. For CV2, $r_{MP}$ values ranged from 0.67 to 0.71 for DH1, 0.40 to 0.56 for DH2 and 0.64 to 0.72 for the association panel. The genomic prediction model which included G x E had the highest average $r_{MP}$ for both CV1 (0.39 and 0.44) and CV2 (0.71 and 0.51) for the association panel and DH2 population, respectively. These results suggest that GS has potential to accelerate breeding for enhanced kernel Zn concentration by facilitating selection of superior genotypes.

KEYWORDS Zea mays L. genetics breeding zinc prediction GenPred Shared data resources Genomic Prediction

Malnutrition arising from zinc (Zn) deficiency is a major risk factor for human health affecting nearly 30% of the world’s population (Bouis and Saltzman 2017; Gannon et al. 2017). The problem is more prevalent in low-and middle income countries (LMICs), and is highly attributed to lack of access to a balanced diet, reliance on cereal-based diets and ignorance of good nutritional practices (Welch and Graham 2004). Several approaches, such as food fortification, diversification and supplementation have been tried to reduce Zn deficiency. However, in LMICs, these methods have not been entirely successful (Misra et al. 2004; Stein 2010).

Breeding maize for increased Zn concentration may offer some relief. The Zn-enriched varieties can be widely accessible, will not require continued investment once developed, and they remain after the initial successful investment and research (Govindan 2011). Recently, maize varieties with 15–36% more Zn were released in Guatemala and Colombia (Listman 2019). Nevertheless, increased breeding efforts are required to develop more Zn-enriched varieties for a diverse range of environments and management practices. Progress toward developing those varieties has mainly relied upon conventional plant breeding approach that is labor-intensive and
time-consuming. However, with the recent advances in genomics, new methods for plant breeding such as genomic selection (GS) can be used to identify genotypes with enhanced Zn concentration more efficiently and rapidly.

Kernel Zn concentration is determined at the end of a plant’s life cycle, so GS can enable selection of promising genotypes earlier in the life cycle. This reduces the time and cost of phenotypic evaluation and may increase the genetic gain per unit time and cost (Heslot et al. 2015; Manickavelu et al. 2017). The utility and effectiveness of GS has been examined for many different crop species, marker densities, traits and statistical models and varying levels of prediction accuracy have been achieved (de los Campos et al. 2009, 2013; Crossa et al. 2010, 2013, 2014; Jarquin et al. 2014; Pérez-Rodríguez et al. 2015; Zhang et al. 2015; Velu et al. 2016). Although the number of markers needed for accurate prediction of genotypic values depends on the extent of linkage disequilibrium between markers and QTL (Meuwissen et al. 2001), a higher marker density can improve the proportion of genetic variation explained by markers and thus result in higher prediction accuracy (Albrecht et al. 2011; Zhao et al. 2012; Combs and Bernardo 2013; Liu et al. 2018). Importantly, higher prediction accuracies have been obtained when genotypes of a population are closely related than when genetically unrelated (Pszczola et al. 2012; Combs and Bernardo 2013; Spindel and McCouch 2016).

Initially, GS models and methods were developed for single-environment analyses and they did not consider correlated environmental structures due to genotype by environment (G x E) interactions (Crossa et al. 2014). The differential response of genotypes in different environments is a major challenge for breeders and can affect heritability and genotype ranking over environments (Monteverde et al. 2018). Multi-environment analysis can model G x E using genetic and residual covariance functions (Burgueño et al. 2012), markers and environmental covariates (Jarquin et al. 2014), or marker by environment (M x E) interactions (Lopez-Cruz et al. 2015). This approach to GS can successfully be used for biofortification breeding of maize because multi-environment testing is routinely used in the development and release of varieties.

Modeling covariance matrices to account for G x E allows the use of information from correlated environments (Burgueño et al. 2012). Mixed models that allow the incorporation of a genetic covariance matrix calculated from marker data, rather than assuming independence among genotypes improves the estimation of genetic effects (VanRaden 2008). The benefit of using genetic covariance matrices in G x E mixed models is that the model relates genotypes across locations even when the lines are not present in all locations (Monteverde et al. 2018). GS models capable of accounting for multi-environment data have extensively been studied in different crops (Zhang et al. 2015; Cuevas et al. 2016, 2017; Velu et al. 2016; Jarquin et al. 2017; Sukumaran et al. 2017a; Monteverde et al. 2018; Roorkee et al. 2018). In those studies, incorporating G x E demonstrated a substantial increase in prediction accuracy relative to single-environment analyses.

Kernel Zn has been investigated in several quantitative trait loci (QTL) analyses in maize and each study has reported that Zn concentration is under the control of several loci. The phenotypic variation explained by those loci ranges from 5.9 to 48.8% (Zhou et al. 2010; Qin et al. 2012; Simić et al. 2012; Baxter et al. 2013; Jin et al. 2013; Zhang et al. 2017a; Hindu et al. 2018). A Meta-QTL analysis across several of those studies identified regions on chromosome 2 that might be important for kernel Zn concentration (Jin et al. 2013). Additionally, genomic regions associated with Zn concentration were recently reported in a genome-wide association study of maize inbreds adapted to the tropics (Hindu et al. 2018). Whereas some of the regions were novel, four of the twenty identified were located in previously reported QTL intervals.

A wide array of maize genetic studies has reported considerable effects of G x E interactions for kernel Zn concentration (Oikeh et al. 2003, 2004; Long et al. 2004; Chakraborti et al. 2009; Prasanna et al. 2011; Agrawal et al. 2012; Guleria et al. 2013). However, genotypes with high-Zn concentration have been identified in both tropical and temperate germplasm (Ahmadi et al. 1993; Bänziger and Long 2000; Brícik et al. 2004; Menkir 2008; Chakraborti et al. 2011; Prasanna et al. 2011; Hindu et al. 2018). Additionally, evaluation procedures for kernel Zn are labor-intensive, expensive and time-consuming (Palacios-Rojas 2018). To the best of our knowledge, no study has examined the predictive ability of GS methods that incorporate G x E for Zn concentration in maize. Within the framework of the reaction norm model (Jarquin et al. 2014), the potential of GS for Zn using maize inbreds adapted to tropical environments were assessed. The objectives of this study were: (i) to evaluate the prediction ability for Zn using an association mapping panel and two bi-parental populations evaluated in three tropical environments, (ii) to assess and compare the predictive ability of different GS models, and (iii) to examine the effects of incorporating G x E on prediction accuracy for Zn.

### MATERIALS AND METHODS

#### Zinc association mapping (ZAM) panel

The ZAM panel consists of 923 inbreds from maize breeding programs of the International Maize and Wheat Improvement Center (CIMMYT). The panel represents wide genetic diversity for kernel Zn concentration (Hindu et al. 2018).

#### Bi-parental DH populations

From the ZAM panel, four inbreds with contrasting Zn concentrations were selected and used to form two bi-parental (doubled haploid [DH]) populations. DH1 was derived from the F1 generation of a mating between CML503, a high-Zn inbred (31.21 µg/g) with CLWN201, a low-Zn inbred (22.62 µg/g). DH2 was derived from the F1 generation of a mating between CML465, another high-Zn inbred (31.55 µg/g) with CML451, a moderate-Zn inbred (27.88 µg/g). DH1 and DH2 were comprised of 112 and 143 inbreds, respectively.

#### Experimental design and phenotypic evaluation

**Zinc association mapping (ZAM) panel:** The ZAM panel was grown at CIMMYT research stations in Mexico, during the months of June through September and November through March at Agua Fria in 2012 and 2013, and Celaya in 2012. Plot sizes and the experimental designs (Hindu et al. 2018).

**Bi-parental DH populations:** The DH populations were grown at CIMMYT research stations in Mexico; Celaya in 2014 and Tlaltizapán (18°41’N, 99° 07’ W; 962.5 m asl) in 2015 and 2017. In 2014 and 2015, both populations were evaluated in single-replication trials (Hindu et al. 2018). In 2017, a randomized complete block design (RCBD) with two replications was used. The rows were 2.5 m long and 75 cm apart and each genotype was grown in a single row plot. All plots were managed according to the recommended agronomic practices for each environment.

From the ZAM panel and each DH population, four to six plants in each plot were self-pollinated, hand-harvested at physiological...
maturity, hand-shelled and dried to a moisture content of 12.5%. The bulked kernels from each plot are considered a representative sample and were used in subsequent Zn analyses as described (Hindu et al. 2018).

Genotypic data

Genomic DNA was extracted from leaf tissues of all inbred lines (ZAM panel and DH populations) using the standard CIMMYT laboratory protocol (CIMMYT, 2005). The samples were genotyped using the genotyping by sequencing (GBS) method at the Institute for Genomic Diversity, Cornell University, USA (Elshire et al. 2011; Crossa et al. 2013). The restriction enzyme ApelK1 was used to digest DNA, GBS libraries were constructed in 96-plex and sequenced on a single lane of Illumina HiSeq2000 flow cell (Elshire et al. 2011). 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 collections (Zhang et al. 2015).

SNP identification was performed using TASSEL 5.0 GBS Discovery Pipeline with B73 (RefGen_v2) as the reference genome (Elshire et al. 2011; Glauheit et al. 2014). The source code and the TASSEL GBS discovery pipeline are available at https://www.maizegenetics.net and the SourceForge Tassel project https://sourceforge.net/projects/tassel. For each inbred, the pipeline yielded 955, 690 SNPs which were distributed on the 10 maize chromosomes. After filtering using a minor allele frequency of 0.05 and removing SNPs with more than 10% missing data, 181,889 (ZAM panel) and 170, 798 (bi-parental) SNPs were used for genomic prediction.

Phenotypic data analysis

For the ZAM panel, broad-sense heritability ($H^2$) across environments was estimated as:

$$H^2 = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_{GE} + \frac{l^2}{l + \sigma^2_e/1}}$$

where $\sigma^2_G$ is the variance due to genotype, $\sigma^2_{GE}$ is variance due to genotype x environment, $\sigma^2_e$ is the error variance, $l$ is the number of environments and $r$ is the number of replicates using multi-environment trial analysis with R (META-R) (Alvarado et al. 2016). For the DH populations, variance components based on the genomic relationship matrix were computed using BGLR package as implemented in GBLUP (Pérez and de los Campos 2014). An estimate of narrow-sense heritability ($h^2$) for each DH population was calculated as:

$$h^2 = \frac{\sigma^2_g}{\sigma^2_G + \frac{l^2}{l + \sigma^2_e/1}}$$

where $\sigma^2_g$ is an estimate of the additive genetic variance and $\sigma^2_e$ is an estimate of the residual variance.

Correlation coefficients between Zn and environments, descriptive statistics and phenotypic data distribution using boxplots were generated in R (core Team 2018). Line means (genotypic values) for the ZAM panel were estimated as Best Linear Unbiased Estimators (BLUEs) with a random effect for replications nested within each environment. Raw data (values) were used for the DH populations.

Statistical models

Genomic models used in this study were based on the reaction norm model which models the markers (genomic) by environment interaction (Jarquin et al. 2014). This model is an extension of the Genomic Best Linear Unbiased Predictor (GBLUP) random effect model, where the main effects of lines (genotypes), genomic, environments and their interactions are modeled using covariance structures that are functions of marker genotypes and environmental covariates.

In this study, environment is the combination of site and year (site-by-year) and the adjusted means (BLUEs) to be used in the genomic prediction models are obtained by fitting the phenotypes $y_{ij}$ as:

$$y_{ij} = \mu + E_i + L_j + LE_{ij} + e_{ij},$$

this linear model represents the response of the $j^h$ ($j = 1, \ldots, J$) genotype/line tested in the $i^h$ ($i = 1, \ldots, I$) environment and $\{y_{ij}\}$ as the sum of an overall mean $\mu$ plus random environmental main effect $[E_i \sim N(0, \sigma^2_E)]$, the random genotype effect $[L_j \sim N(0, \sigma^2_L)]$, the random interaction between the $j^h$ genotype and the $i^h$ environment $[LE_{ij} \sim N(0, \sigma^2_{LE})]$ and a random error term $[e_{ij} \sim N(0, \sigma^2_e)]$. From this linear model, $N(\cdot)$ denotes a normal random variable, iid stands for independent and identically distributed responses and $\sigma^2_E, \sigma^2_L, \sigma^2_{LE}, \sigma^2_e$ are the variances for environment, genotype, genotype by environment and residual error, respectively. The model above does not allow borrowing of information among genotypes because the genotypes were treated as independent outcomes. Thus, models used in this study were derived from the baseline model above by subtracting terms or modifying assumptions and/or incorporating genomics/marker information. A brief description of the genomic models used in this study are given below.

M1. Environment + Line

This model is obtained by retaining the first three components from the baseline model (overall mean, random environment main effect and random line main effect) while their underlying assumptions remain unchanged.

$$y_{ij} = \mu + E_i + L_j + e_{ij}. \quad [1]$$

Here environments were considered as site-by-year combinations.

M2. Environment + Line + Genomic

Another representation of the random main effect of line $L_i$ in the previous model is considering a linear combination between markers and their correspondent marker effects, $g_i = \sum_{m=1}^p x_{jm} b_m$, such that

$$y_{ij} = \mu + E_i + L_j + g_i + e_{ij} \quad [2]$$

where $b_m \sim N(0, \sigma^2_g)$ represents the random effect of the $m^{th}$ ($m = 1, \ldots, p$) marker, $x_{jm}$ is the genotype of the $j^h$ line at the $m^{th}$ marker and $\sigma^2_g$ its correspondent variance component.

Therefore, $g_i = (g_{i1}, \ldots, g_{ip})$, is the vector of genetic effects, and follows a normal density with mean zero, and a co-variance matrix $\text{Cov}(g) = \sigma^2_g \mathbf{I}$ being the genomic relationship matrix (Lopez-Cruz et al. 2015) that describes genetic similarities among pairs of individuals. In this model, the line effect $L_i$ is retained to account for imperfect information and model mis-specification because of potential imperfect linkage disequilibrium between markers and quantitative trait loci (QTL).

M3. Environment + Line + Genomic + Genomic $\times$ Environment

This model accounts for the effects of lines $L_i$, of markers (genomic) $g_i$, of environments (E) and the interaction between markers
(genomic) and the environment \((gE)\). The model includes the interaction between markers (genomics) and the environment via co-variance structure (Jarquín et al. 2014). The model is as follows:

\[
y_{ij} = \mu + E_i + I_j + g_j + gE_{ji} + e_{ij} \quad [3]
\]

Where \(gE_{ji}\) is the interaction between the genetic value of the \(j^{th}\) genotype in the \(i^{th}\) environment and \(gE = \{gE_i\} \sim N(0, (Z_gGZ_g)^{-1} \otimes (Z_gZ_g)\sigma_{gE}^2)\), where \(Z_g\) and \(Z_e\) are the correspondent incidence matrices for the effects of genetic values of genotypes and environments, respectively.

\(\sigma_{gE}^2\) is the variance component of \(gE\) and \(\otimes\) denotes the Hadamard product (element-to-element product) between two matrices.

**Model assessment**

Models were first fitted to the entire data set to estimate variance components using the R-package BGLR (de los Campos et al. 2010; de los Campos and Perez-Rodriguez 2016). The information generated from the full data analyses was not used as prior information for the cross-validation schemes (CV1 and CV2) used for assessing the prediction accuracy of the different models.

**Prediction accuracy assessment using cross-validation**

Two distinct cross-validation schemes that mimic prediction problems that breeders may face when performing genomic prediction were used (Burgueño et al. 2012). One random cross-validation (CV1) evaluates the prediction ability of models when a set of lines have not been evaluated in any environment (prediction of newly developed lines). In CV1, predictions are entirely based on phenotypic records of genetically related lines. The second cross-validation (CV2) is related to incomplete field trials also known as sparse testing, in which some lines are observed in some environments but not in others. In CV2, the goal is to predict the performance of lines in environments where they have not yet been observed. In this study, CV2 mimics a situation where lines are evaluated in two environments but missing in the third environment. Thus, information from related lines and the correlated environments is used, and prediction assessment can benefit from borrowing information between lines within an environment, between lines across environments and among correlated environments.

In CV1 and CV2, a fivefold cross-validation scheme was used to generate the training and validation sets to assess the prediction ability for Zn within the ZAM panel and each DH population. The data were randomly divided into five subsets, with 80% of the lines assigned to the training set and 20% assigned to the validation set. Four subsets were combined to form the training set, and the remaining subset was used as the validation set. Permutation of five subsets taken one at a time led to five training and validation data sets. The procedure was repeated 20 times and a total of 100 runs were performed in each population. The average value of the correlations between the phenotype and the genomic estimated breeding values (GEBVs) from 100 runs was calculated for the ZAM panel, and each DH population for Zn in each environment and was defined as the prediction ability \((r_{MP})\).

**Data availability**

All models were fitted in R (core Team 2018) using the BGLR package (Pérez and de los Campos 2014). All phenotypic and genomic data can be downloaded from the link: [http://hdl.handle.net/11529/10548331](http://hdl.handle.net/11529/10548331)

**RESULTS**

**Descriptive statistics**

Mean values of kernel Zn concentration were estimated for each environment and across environments (Tables 1 and 2). For the ZAM panel, kernel Zn ranged from 14.76 to 39.80 \(\mu g/g\) in Celaya 2012, 15.16 to 42.52 \(\mu g/g\) and 17.05 to 46.52 \(\mu g/g\) in Agua Fria 2012 and 2013, respectively (Figure 1). The highest mean (29.53 \(\mu g/g\)) for Zn was observed in Agua Fria 2013. DH1 had Zn values ranging from 16.00 to 48.00 \(\mu g/g\) in Celaya 2012, 16.00 to 35.00 \(\mu g/g\) in Tlaltizapan 2015 and 15.50 to 39.00 \(\mu g/g\) in Tlaltizapan 2017, while the respective values for DH2 were 17.70 to 43.14 \(\mu g/g\), 15.60 to 37.80 \(\mu g/g\) and 14.70 to 37.60 \(\mu g/g\) (Figures 2A and 2B). The highest means for Zn were observed in Celaya 2014 (25.38 \(\mu g/g\)) and 2017 (27.96 \(\mu g/g\)) for DH1 and DH2, respectively (Table 2). Across environments, \(H^2\) for

<p>| Table 1 Descriptive statistics for kernel Zn concentration for the ZAM panel grown in three environments |</p>
<table>
<thead>
<tr>
<th>Population</th>
<th>Population size</th>
<th>Location</th>
<th>Mean ± SE ((\mu g/g))</th>
<th>(\sigma_{g}^2)</th>
<th>(\sigma_{gE}^2)</th>
<th>(H^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZAM panel</td>
<td>923</td>
<td>Agua Fria 2012</td>
<td>26.15 ± 0.15</td>
<td>12.04</td>
<td>2.42</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Celaya 2012</td>
<td>25.06 ± 0.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Agua Fria 2013</td>
<td>29.53 ± 0.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Across</td>
<td>26.94 ± 0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(H^2 = \) Broad-sense heritability for Zn across environments. \(\sigma_{g}^2 = \) variance due to genotypes \(\sigma_{gE}^2 = \) the interaction between genotypes and the environment significant at \(P < 0.001\).

<p>| Table 2 Descriptive statistics for kernel Zn concentration for DH populations grown in three environments |</p>
<table>
<thead>
<tr>
<th>Population</th>
<th>Population size</th>
<th>Location</th>
<th>Mean ± SE ((\mu g/g))</th>
<th>(j^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH1</td>
<td>112</td>
<td>Celaya 2014</td>
<td>25.38 ± 0.48</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tlaltizapan 2015</td>
<td>24.01 ± 0.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tlaltizapan 2017</td>
<td>24.53 ± 0.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Across</td>
<td>24.65 ± 0.26</td>
<td>0.76</td>
</tr>
<tr>
<td>DH2</td>
<td>143</td>
<td>Celaya 2014</td>
<td>27.96 ± 0.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tlaltizapan 2015</td>
<td>24.08 ± 0.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tlaltizapan 2017</td>
<td>24.64 ± 0.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Across</td>
<td>25.59 ± 0.22</td>
<td></td>
</tr>
</tbody>
</table>

\(j^2 = \) Narrow-sense heritability for Zn across environments.
the ZAM panel was 0.85 (Table 1) and the $h^2$ for DH1 and DH2 were 0.83 and 0.76, respectively (Table 2). There were significant positive correlations between environments for Zn (Table 3), accounting for the moderate to high heritability estimates.

Principal component analysis for the ZAM panel suggested presence of a relatively diverse set of lines, and 452 principal components (PCs) were needed to explain 80% of the genotypes’ variance (Figures 3A and 3B). The first two principal components explained 3.85% of the total variance. For the DH populations first two eigenvectors separated the two groups (DH1 and DH2) and 56 principal components were needed to explain 80% of the genotypes’ variance (Figures 3C and 3D). The first two principal components explained 27.50% of the total variation for the DH populations.

Estimates of variance components

Variance component estimates for all models were derived from the full-data analysis (Table 4). In the ZAM panel, variance components for M1 ranged from 16.18 to 7.01 with the main effect of environments explaining the largest proportion (46%). When marker information was incorporated (i.e., M2 and M3), the estimated variance due to environments was reduced to 9.46 for M2 and 8.11 for M3. Inclusion of the interaction term (genomic x environment) reduced the residual variance component by ~30%, from an estimated residual variance of 7.01 in M1 to 4.92 in M3 suggesting that some components of differences among genotypes cannot be fully captured by the main effects of markers and environments. Similar trends were observed in DH1 and DH2 except for M1 of DH1 where the main effect of lines accounted for a slightly higher proportion of the total variance (38%) than the main effects of environments (35%).

The residual variance component values for all models of DH populations were slightly higher than those for the ZAM panel (Table 4), possibly because the populations were evaluated in single-replicated experiments at Celaya and Tlaltizapan (2014 and 2015, respectively). However, estimates from M3 of the DH populations suggest that ≥ 60 of the within-environment variability can be explained by main effects of lines, genomic (markers) and their interaction term. The proportions of within-environment variation explained by the interaction term were ≥ 10% for the ZAM panel and DH populations suggesting the importance of considering such interactions in genomic prediction models.

Prediction ability in different populations

Cross-validated $r_{MP}$ values for kernel Zn were estimated for the ZAM panel and DH populations (Tables 5, 6 and 7). The average $r_{MP}$ values in CV1 were consistently lower than those in CV2, suggesting the importance of using information from correlated environments when predicting performance of inbred lines. The mean $r_{MP}$ values in CV1 and CV2 for the ZAM panel were 0.39 and 0.71, respectively (Table 5). For the DH populations, average $r_{MP}$ values were 0.53 for DH1-CV1, 0.44 for DH2-CV1 (Table 6), 0.70 for DH1-CV2 and 0.51 for DH2-CV2 (Table 7).

In the ZAM panel, the highest values in CV1 (0.47) and CV2 (0.72) were obtained in Celaya and Agua Fria 2012 (Table 5). For the bi-parental populations, both under CV1 and CV2, higher $r_{MP}$ values were observed for DH1 compared to DH2. The highest values in CV1 (0.56) and CV2 (0.71) were observed in Tlaltizapan 2017 and 2015, all for DH1 (Tables 6 and 7). The consistently higher $r_{MP}$ values in CV1 and CV2 of DH1 could be attributed to the higher (0.58 to 0.62) correlation values between environments (Table 3).

Prediction ability of different models

Comparing the $r_{MP}$ values obtained from each model, M1 had the lowest (-0.001, -0.03 and 0.04) accuracies in CV1 for the ZAM panel and DH populations (Tables 5 and 6). Those values were improved in CV2 because the predictions benefited from previous records.
(collected from other environments) of lines whose Zn values were being predicted. When M1 was expanded to M2 by adding the main effects of markers, the $r_{MP}$ values at each environment and across environments were increased. For example, in CV1, M2, 100-fold increase in $r_{MP}$ values were observed for the ZAM panel and DH populations, and in CV2, M2, average $r_{MP}$ values increased by 2.98%, 2.94% and 11.11% for the ZAM panel, DH1 and DH2, respectively (Tables 5, 6 and 7).

The multi-environment model (M3), which includes the interaction between markers (genomic) and the environment (gEji) gave higher prediction accuracy than single-environment models (M1 and M2). In CV1, mean $r_{MP}$ values increased from 0.37 (M2) to 0.39 (M3) for the ZAM panel and from 0.43 (M2) to 0.44 for DH2 (Tables 5 and 6). Similar trends were observed in CV2 for the ZAM panel and DH2 (Tables 5 and 7). However, in both CV1 and CV2 of DH1, incorporating gEji did not improve $r_{MP}$ values for Zn (Tables 6 and 7). For CV1, M3, $r_{MP}$ values for Zn in individual environments ranged from 0.34 to 0.47 for the ZAM panel (Table 5), 0.51 to 0.55 for DH1 and 0.35 to 0.50 for DH2 (Table 6). For CV2, M3, those values ranged from 0.69 to 0.72 for the ZAM panel, 0.68 to 0.70 for DH1 and 0.43 to 0.56 for DH2 (Tables 5, 6 and 7).

**DISCUSSION**

Overall, moderate to high prediction ability values for kernel Zn were observed for the ZAM panel and DH populations. This could be attributed to the heritabilities observed for kernel Zn (Tables 1 and 2). Similar observations were reported for Zn concentration in wheat (Velu et al. 2016; Manickavelu et al. 2017). Higher predicted values with high accuracy for GS programs are expected for traits with moderate to higher heritability estimates (Combs and Bernardo 2013; Lian et al. 2014; Muranty et al. 2015; Saint Pierre et al. 2016; Manickavelu et al. 2017; Zhang et al. 2017b 2019; Arojju et al. 2019). Consistent with a study on Zn and iron (Fe) concentration in spring wheat, the prediction accuracies in this study are sufficient to discard at least 50% of the inbreds with low-Zn concentration (Velu et al. 2016).

Additionally, the moderate to high prediction accuracies reported in this study shows that GS can be used in maize breeding to improve kernel Zn concentration. Assuming two possible seasons of Zn evaluation per year, the predicted genetic gains can be estimated from prediction accuracies and genetic variances of the training populations. The genetic variances for the ZAM panel, DH1 and DH2 were 12.38, 12.20 and 14.88, and prediction accuracies were 0.71, 0.70 and 0.51, respectively. If the inbreds in each predicted population are ranked based on their predicted Zn values and the top 10% selected, then their expected average Zn values can be estimated from the proportion of inbreds selected, their respective training population genetic variances, prediction accuracies and the time interval for evaluating the lines. With reference to this, the expected average values of Zn are approximately 31 μg/g for the ZAM panel, 30 μg/g for DH1 and 27 μg/g for DH2. These averages are higher |

| Table 3 Phenotypic correlation between environments for kernel Zn |
|-------------------|------------------|------------------|
|                   | DH1  | DH 2  | ZAM Panel |
| Env1 vs. Env2     | 0.62 | 0.46  | 0.63       |
| Env1 vs. Env3     | 0.58 | 0.29  | 0.66       |
| Env2 vs. Env3     | 0.62 | 0.45  | 0.61       |

Phenotypic correlation coefficients were significant at $\alpha = 0.001$.

**Figure 3** Scree plots (A and C) and loadings of the first two eigenvectors (B and D) of the covariance matrices derived from markers for the ZAM panel (A and B) and for the DH populations (C and D).
than the averages of the respective training populations (~27 μg/g for the ZAM panel, ~25 μg/g for DH1 and ~26 μg/g for DH2) suggesting that the prediction accuracies achieved are sufficient to select at least 10% of the predicted inbreds with higher Zn concentration.

Data from both bi-parental populations and diverse collection of inbreds have been used for GS and based on cross-validation (CV), it has been established that prediction accuracies could also be affected by the relatedness between training and prediction sets (Habier et al. 2007; de Roos et al. 2009; Asoro et al. 2011; Daetwyler et al. 2013; Cericola et al. 2017; Crossa et al. 2017). In this study, average predicted accuracies were higher for CV1 of the bi-parental populations (0.53 for DH1 and 0.44 for DH2) compared to the ZAM panel (0.39). Higher predicted values in CV1 of the DH populations could be attributed to the closer relationship among DH lines in the training and prediction sets, maximum linkage disequilibrium (LD) between a marker and a QTL, and controlled population structure (Bernardo and Yu 2007; Albrecht et al. 2011; Zhang et al. 2015). In collections of diverse inbreds, prediction accuracy may depend on the ancestral relationships between the lines. So, in experiments using such collections of lines, prediction accuracies have been more variable than accuracies achieved using bi-parental populations (Spindel and McCouch 2016).

Cross-validation (CV) schemes are used in genomic prediction to estimate the accuracy with which predictions for different traits and environments can be made (Burgueño et al. 2012; Zhang et al. 2015; Saint Pierre et al. 2016; Velu et al. 2016; Sukumaran et al. 2017a, 2017b; Monteverde et al. 2018; Roorkiwal et al. 2018). In this study, two CV schemes (CV1- predicting the performance of newly developed lines, and CV2- predicting the performance of lines that have been evaluated in some environments, but not in others) were used. The utility of these schemes indicated that prediction values for newly developed lines (CV1) were generally lower (0.39 for the ZAM panel, 0.53 for DH1 and 0.44 for DH2) than the values for lines which have been evaluated in different but correlated environments (CV2; 0.71, 0.70 and 0.51 for the ZAM panel, DH1 and DH2, respectively). Such observations indicate the importance of using information from correlated environments when predicting the performance of inbred lines. However, selection of new lines without direct field testing, as simulated in CV1, may enhance the breeding process by replacing the time and labor-intensive field testing for Zn with genomic-estimated breeding values. But, the prediction accuracy values obtained may be

<table>
<thead>
<tr>
<th>Table 4 Estimated variance components (estimate ± SD) and percentage of within-environment variance accounted for by each random effect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>ZAM panel</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>L</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>G x E</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>DH1</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>L</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>G x E</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>DH2</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>L</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>G x E</td>
</tr>
<tr>
<td>Residual</td>
</tr>
</tbody>
</table>

**E** = Environment, **L** = Line, **G** = genomic (marker information), **G x E** = genomic x environment.

**a**Relative to the total variance minus the variance due to main effect of the environment. The percentages of within-environment variance were computed without taking into account the variance of the environment.

<table>
<thead>
<tr>
<th>Table 5 Correlations (mean ± SD) between observed and genomic estimated breeding values for kernel Zn in the three environments for three GBLUP models for cross-validations CV1 and CV2 of the ZAM panel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>ZAM panel (923)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Population</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>ZAM panel (923)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Models: M1= Environment + Line, M2 = Environment + Line + Genomic, M3 = Environment + Line + Genomic + Genomic x Environment.**
lower such that the annual rate of genetic progress in a GS program is compromised (Burgueño et al. 2012). So, the ultimate decision of how a breeding scheme should be structured could depend on the compromise between the desired prediction accuracy and the generation interval (Burgueño et al. 2012).

Genotype by environment interaction is an important factor affecting kernel Zn concentration in maize and genomic prediction models that incorporate G x E may enhance the potential of GS for biofortification breeding. For different crop species and traits, genomic prediction models which incorporated G x E achieved higher prediction accuracies in both CV1 and CV2 schemes relative to models which did not include G x E (Burgueño et al. 2012; Guo et al. 2013; Jarquin et al. 2014; Lopez-Cruz et al. 2015; Zhang et al. 2015; Monteverde et al. 2018). In this study, the impact of modeling G x E variance structures for multi-environment trials was investigated and results indicated that the average predicted values from M3 (G x E model) were higher (0.39 and 0.44 for CV1 and 0.71 and 0.51 for CV2) than the values from M2 (non-G x E; 0.37 and 0.43 for CV1-M2, 0.69 and 0.50 for CV2-M2) for the ZAM panel and DH2. These findings agree with those reported on Zn concentration in wheat (Velu et al. 2016), providing evidence that incorporating G x E in GS models can enhance their power and suitability for improving maize for kernel Zn concentration. Conversely, the average predicted values for CV1 and CV2 of DH1 were higher in M2 (0.53 and 0.70) than in M3 (0.53 and 0.69). Except for differences in population size (112 lines vs. 143 lines), this was unexpected since DH1 and DH2 were grown in the same environments.

The gains in prediction accuracies for the GS model that accounted for G x E were dependent on the correlation between environments and CV method used. In this study, the phenotypic correlations between environments were all positive (ranging from 0.58 to 0.62 for DH1, 0.29 to 0.46 for DH2 and 0.61 to 0.66 for the ZAM panel). Such correlations can be exploited using multi-environment models to derive predictions that use information from across both the lines and environments (Burgueño et al. 2012). For instance, although the phenotypic correlations between environments for DH2 were positive (0.29 to 0.46), the lowest average prediction value (0.51) for CV2 was observed for this population. This was expected because CV2 uses phenotypic information from genotypes which have already been tested; hence, effectively exploiting the correlations between environments (Burgueño et al. 2012; Jarquin et al. 2014; Crossa et al. 2015; Pérez-Rodríguez et al. 2015; Saint Pierre et al. 2016; Monteverde et al. 2018). However, for CV1, the information between environments could only be accounted for through the genomic relationship matrix (Monteverde et al. 2018). Hence, the gains in CV1 may likely attribute to more accurate estimate of environment-specific marker effects (Guo et al. 2013). In contrast, when multiple environments are weakly correlated, prediction accuracies from across environment analyses can be negatively affected relative to prediction accuracies within environments (Bentley et al. 2014; Wang et al. 2014; Spindel and McCouch 2016). Thus, before designing a GS experiment, identifying correlated environments where environments can differ in terms of site, year or season in which data were collected is of great interest (Spindel and McCouch 2016).

However, the prediction accuracy values were of lower quality when genomic predictions were conducted across populations. For instance, when the ZAM panel was used as the training population, prediction accuracies for DH1, DH2 and DH1+DH2 were 0.15, -0.10 and 0.09, respectively. When DH1 and DH2 were used as a training and prediction set for each other, prediction accuracies were 0.08 and 0.16 (Unpublished data). These prediction accuracies are

---

### Table 6 Correlations (mean ± SD) between observed and genomic estimated breeding values for Zn in the three environments for three GBLUP models for cross-validation CV1 of DH populations

<table>
<thead>
<tr>
<th>Population</th>
<th>Environment</th>
<th>Prediction accuracy in CV1</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH1</td>
<td>Celaya, 2014</td>
<td>-0.05 ± 0.10</td>
<td>0.52 ± 0.04</td>
<td>0.51 ± 0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tlaltizapan, 2015</td>
<td>-0.02 ± 0.12</td>
<td>0.52 ± 0.05</td>
<td>0.51 ± 0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tlaltizapan, 2017</td>
<td>-0.01 ± 0.10</td>
<td>0.56 ± 0.05</td>
<td>0.55 ± 0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td><strong>-0.03 ± 0.10</strong></td>
<td><strong>0.53 ± 0.04</strong></td>
<td><strong>0.52 ± 0.04</strong></td>
<td></td>
</tr>
<tr>
<td>DH2</td>
<td>Celaya, 2014</td>
<td>0.05 ± 0.08</td>
<td>0.47 ± 0.03</td>
<td>0.50 ± 0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tlaltizapan, 2015</td>
<td>0.03 ± 0.08</td>
<td>0.45 ± 0.03</td>
<td>0.45 ± 0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tlaltizapan, 2017</td>
<td>0.04 ± 0.08</td>
<td>0.35 ± 0.03</td>
<td>0.35 ± 0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td><strong>0.04 ± 0.06</strong></td>
<td><strong>0.43 ± 0.03</strong></td>
<td><strong>0.44 ± 0.02</strong></td>
<td></td>
</tr>
</tbody>
</table>

*Models: M1 = Environment + Line; M2 = Environment + Line + Genomic; M3 = Environment + Line + Genomic + Genomic x Environment.

### Table 7 Correlations (mean ± SD) between observed and genomic estimated breeding values for Zn in the three environments for three GBLUP models for cross-validation CV2 of DH populations

<table>
<thead>
<tr>
<th>Population</th>
<th>Environment</th>
<th>Prediction accuracy in CV2</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH1</td>
<td>Celaya, 2014</td>
<td>0.67 ± 0.02</td>
<td>0.68 ± 0.02</td>
<td>0.68 ± 0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tlaltizapan, 2015</td>
<td>0.70 ± 0.02</td>
<td>0.71 ± 0.02</td>
<td>0.70 ± 0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tlaltizapan, 2017</td>
<td>0.67 ± 0.02</td>
<td>0.70 ± 0.02</td>
<td>0.69 ± 0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td><strong>0.68 ± 0.01</strong></td>
<td><strong>0.70 ± 0.01</strong></td>
<td><strong>0.69 ± 0.01</strong></td>
<td></td>
</tr>
<tr>
<td>DH2</td>
<td>Celaya, 2014</td>
<td>0.46 ± 0.016</td>
<td>0.53 ± 0.02</td>
<td>0.56 ± 0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tlaltizapan, 2015</td>
<td>0.50 ± 0.020</td>
<td>0.55 ± 0.02</td>
<td>0.55 ± 0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tlaltizapan, 2017</td>
<td>0.40 ± 0.023</td>
<td>0.43 ± 0.02</td>
<td>0.43 ± 0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td><strong>0.45 ± 0.02</strong></td>
<td><strong>0.50 ± 0.01</strong></td>
<td><strong>0.51 ± 0.01</strong></td>
<td></td>
</tr>
</tbody>
</table>

*Models: M1 = Environment + Line; M2 = Environment + Line + Genomic; M3 = Environment + Line + Genomic + Genomic x Environment.
considerably lower than those reported in this study and the differences may be attributed to: (i) weak genetic relationships between the training and prediction population sets and (ii) different methods of analysis because the prediction accuracies reported in this study were partly achieved by modeling the random-effects environment structure to account for G x E while for the unpublished data, the random-effects environment structure of G x E was not included.

The ability to predict kernel Zn concentration using high-throughput SNP markers including G x E interactions creates an opportunity for efficiently enhancing Zn concentration in maize breeding programs. For instance, during early generations of a breeding program, GS can be utilized to identify genotypes with favorable alleles when numbers of progenies and families are large. This could potentially reduce the resource-intensive evaluation process and advancement of false-positive progenies (Velu et al. 2016). Coupled with advances in technologies for assessing Zn, plant scientists can more rapidly measure Zn concentration in maize kernels using the energy dispersive x-ray fluorescence (XRF) assays (Guld et al. 2017). Thus, with more validations and model refinements, GS can potentially accelerate the breeding process to enhance Zn concentration in maize for a wider range of environments.

ACKNOWLEDGMENTS

We thank HarvestPlus through CIMMYT and the R.F. Baker Center for Plant Breeding, Department of Agronomy at Iowa State University for making this research possible. We also would like to thank, Andrea Cruz-Morales, Mayolo Leyva, Balfre Tellez-Noriega and Adolfo Basilio for managing the trials and collecting the unpublished data. The ability to predict kernel Zn concentration using high-throughput SNP markers including G x E interactions creates an opportunity for efficiently enhancing Zn concentration in maize breeding programs. For instance, during early generations of a breeding program, GS can be utilized to identify genotypes with favorable alleles when numbers of progenies and families are large. This could potentially reduce the resource-intensive evaluation process and advancement of false-positive progenies (Velu et al. 2016). Coupled with advances in technologies for assessing Zn, plant scientists can more rapidly measure Zn concentration in maize kernels using the energy dispersive x-ray fluorescence (XRF) assays (Guld et al. 2017). Thus, with more validations and model refinements, GS can potentially accelerate the breeding process to enhance Zn concentration in maize for a wider range of environments.

LITERATURE CITED


CIMMYT, 2005 Laboratory Protocols, CIMMYT Applied Molecular Genetics Laboratory. Ed. 2. CIMMYT, Mexico, D.F.


