OXFORD
G3 Genes | Genomes | Genetics

G3, 2022, jkab440
https://doi.org/10.1093/g3journal/jkab440
Advance Access Publication Date: 28 December 2021
Genomic Prediction

Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data

Anna R. Rogers,[1] James B. Holland[1,2,3,*]

[1]Program in Genetics, North Carolina State University, Raleigh, NC 27695, USA,
[2]USDA-ARS Plant Science Research Unit, North Carolina State University, Raleigh, NC 27695, USA,
[3]Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC 27695, USA

*Corresponding author: Department of Agriculture—Agriculture Research Service, Box 7620 North Carolina State University, Raleigh, NC 27695-7620, USA.
Email: jim.holland@usda.gov

Abstract

Technology advances have made possible the collection of a wealth of genomic, environmental, and phenotypic data for use in plant breeding. Incorporation of environmental data into environment-specific genomic prediction is hindered in part because of inherently high data dimensionality. Computationally efficient approaches to combining genomic and environmental information may facilitate extension of genomic prediction models to new environments and germplasm, and better understanding of genotype-by-environment (G × E) interactions. Using genomic, yield trial, and environmental data on 1,918 unique hybrids evaluated in 59 environments from the maize Genomes to Fields project, we determined that a set of 10,153 SNP dominance coefficients and a 5-day temporal window size for summarizing environmental variables were optimal for genomic prediction using only genetic and environmental main effects. Adding marker-by-environment variable interactions required dimension reduction, and we found that reducing dimensionality of the genetic data while keeping the full set of environmental covariates was best for environment-specific genomic prediction of grain yield, leading to an increase in prediction ability of 2.7% to achieve a prediction ability of 80% across environments when data were masked at random. We then measured how prediction ability within environments was affected under stratified training-testing sets to approximate scenarios commonly encountered by plant breeders, finding that incorporation of marker-by-environment effects improved prediction ability in cases where training and test sets shared environments, but did not improve prediction in new untested environments. The environmental similarity between training and testing sets had a greater impact on the efficacy of prediction than genetic similarity between training and test sets.

Keywords: genotype-by-environment interactions; multienvironment; genomic prediction; environmental covariates; dominance genetic variance; shared data resource

Introduction

Genotype-by-environment (G × E) interactions occur when environmental factors do not have the same effect on all genotypes, such that the relative phenotypic differences among genotypes vary across environments. G × E interactions are commonly observed in complex phenotypes studied across all subfields of genetics (Manolio et al. 2009; Lasky et al. 2015; Edwards 2016; Krishnamurthy et al. 2017; Yang et al. 2017) and complicate the understanding of genetic and genotypic effects. In plant breeding, G × E represents a portion of phenotypic variance that hinders broad-scale adaptation, but may be leveraged for development of varieties with relatively better performance in the presence of specific environmental pressures, for example, drought (Adee et al. 2016) or salinity tolerance (Krishnamurthy et al. 2017). Plant breeders have traditionally treated G × E as noise that hinders stability across a wide target population of environments (Comstock and Moll 1963), in part because of the challenge it poses during the selection process. Identifying superior genotypes in plant breeding programs is difficult because G × E interactions can lead to the selection of genotypes based on performance in 1 or a small set of test environments in the early stages of a breeding program that might have relatively poor performance in later stage multienvironment trials (Crossa et al. 2017; Hickey et al. 2017). Without a good understanding of G × E interactions and their impacts on performance in early-stage selections, G × E interactions become a nuisance at later stages when promising lines meet new stressors. This also poses a problem when breeding for future environments, where climate change threatens to alter the growing environments in ways that may not favor current genetic backgrounds. To breed for future environments where extreme weather events are likely to be more frequent, leveraging G × E will may help to develop more resilient cultivars.

Over the last decade, genomic selection (GS) has become widely used in both plant and animal breeding programs, in part

Received: November 9, 2021. **Accepted:** December 6, 2021

due to the decreased cost of genomic data and increased computational power available for statistical modeling (Cooper et al. 2014; Hickey et al. 2017; Hammer et al. 2019; Voss-Fels et al. 2019). As technologies for capturing environmental data developed and became more cost effective for use in breeding programs, models have been proposed to integrate environmental covariates into GS (Jarquín et al. 2014; Saint Pierre et al. 2016; Millet et al. 2019; Monteverde et al. 2019). These range across a spectrum from extensively researched crop models that rely on biological understanding of the developmental response of specific genotypes to changing environmental conditions (Pauli et al. 2016; Bustos-Korts et al. 2019), to reaction-norm models identifying a few covariates that are posed to have important effects on end point phenotypes (Burgueño et al. 2012; Heslot et al. 2014; Jarquín et al. 2014; Crossa et al. 2016; Millet et al. 2019; Li et al. 2021), and to machine learning models that "learn" what covariates are important to a given trait from a large amount of environmental data based on assumptions of effect distributions (Bandeira E Sousa et al. 2017; Cuevas et al. 2017, 2018; Montesinos-López et al. 2018; Costa-Neto et al. 2021; Li et al. 2021). Inferences about the causes of G × E interactions may be possible with crop physiology models but are difficult to make from machine learning model results.

Phyisologically grounded crop models rely on extensive and detailed experimentation to estimate model parameters (Bustos-Korts et al. 2019; Hammer et al. 2019), which can provide high predictive ability of genotypes used to train models in new environments, but much lower prediction ability for new genotypes not used in model training (Millet et al. 2019). These results suggest that transferring information on crop growth models between populations remains challenging (Millet et al. 2019), limiting their practicality to large-scale breeding programs, where thousands of genotypes and environments may be present and generating the data necessary to create crop growth models may not be feasible (Hammer et al. 2019). Even when such models can be used, computational power presents a challenge and parsimony is often preferred to speed up modeling for real-time breeding decisions (Hammer et al. 2019).

Reaction-norm and machine learning models provide simpler alternatives for integrating both genomic and environmental data for genomic prediction (GP), in part because these methodologies are relatively straightforward extensions of GS models that handle environmental data and G × E interactions in ways analogous to use of genomic marker data in GS (Crossa et al. 2016). One caveat of many GP studies incorporating environmental data into GxE modeling is that they tend to focus on relatively small numbers of environments and genetic backgrounds, which may not be representative of breeding programs that incorporate more heterogeneity of genotypes or target population of environments. Predictive studies using both environmental and genomic data sources have shown that use of G × E terms increase predictive ability, but that the increase in predictive ability depends greatly on phenotypic correlations between testing and training environments (Bandeira E Sousa et al. 2017; Cuevas et al. 2018; Monteverde et al. 2018). When covariances between training and testing environments were close to 0 or negative, linear G × E models had difficulty predicting phenotypic outcomes, but in some cases more flexible kernel models were better able to approximate these environmental covariances and recover prediction ability (Crossa et al. 2017; Cuevas et al. 2017, 2018). These kernel models tend to be costly in memory requirements and computational speeds, creating a significant barrier to entry for small breeding programs with limited resources and large programs requiring speedy selections (Isik et al. 2017; Cuevas et al. 2018; Granato et al. 2018). Because of this, more computationally efficient methods that may make prediction models more useful for making breeding decisions.

Rogers et al. (2021) curated phenotypic, environmental, and genomic data involving 1,918 maize hybrids tested in up to 65 environments from the Genome to Fields (G2F) project in years 2014–2016. Environmental covariates were related to covariances between yield performances in different environments in this data set (Rogers et al. 2021), thus we hypothesize that environment-specific GP could be aided by the inclusion of environmental covariates and G × E interactions models. Because some environmental data are collected many times per environment, integrating these data into GP models require a choice of appropriate resolution of windows over which environmental data are summarized. Higher resolution may help prediction ability, but it increases computational demand. Similarly, very high-resolution genetic marker data may improve prediction ability but there is a tradeoff between prediction ability and computational resource requirements when increasing marker density. Combining genetic marker coefficients and environmental variable coefficients to create G × E interaction coefficients increase the number of model parameters by the product of the 2 components; therefore, identifying an appropriate balance between the density of genetic and environmental factors and the memory demands and computational speed of prediction models is important. One proposed solution to this problem is the use of environmental indexing over an optimal window, which has been successful in the case of maize flowering time where a small set of environmental covariates have large effects on phenotypic outcomes (Li et al. 2018). This approach has used for traits in several species where environmental factors with large influence on phenotypes are well understood (Li et al. 2021). Such approaches are not easily generalizable to prediction of grain yield, which is influenced by a large number of environmental factors and exhibits substantial GxE interaction variance.

Jarquín et al. (2020) utilized the first 2 years of G2F hybrid yield trial data to test reaction-norm models involving 1-h windows of environmental data and genomic relationship matrices to model general and specific combining ability genetic components. These models involved 25,152 environmental covariables summarized into an environmental covariance matrix, which assumes that all environmental covariates contribute equally to the relationships between the environments (Jarquín et al. 2020). This approach for including G×E interactions did not achieve consistent improvement in predictive ability for grain yield (Jarquín et al. 2020). Here, we approach environment-specific prediction in a way that allows marker-environmental variable interactions to have different weights in prediction models, as learned from the data, for greater flexibility in modeling the covariances between environments.

The objectives of this study were to (1) optimize genetic marker and environmental covariable data sets to balance GP ability against model dimensionality (as a proxy for computational resource requirements), and (2) measure the effect on prediction ability of different prediction models when training and testing data sets were stratified by genetic or environmental relationships, to reflect real world prediction scenarios encountered by plant breeders. We compared the prediction ability of genetic marker data sets coded to represent additive vs dominance coefficients and randomly sampled to different marker densities. We also compared temporal window sizes for summarizing environmental variables, and dimension reduction of either the genetic or environmental data for use in incorporation G × E effects in models for environment-specific GP. Finally, we compared the relative utility of different ways to represent G × E effects in prediction models under scenarios with varying levels of genetic and environmental separation between training and testing data sets.

Materials and methods

Environmental, phenotypic, and genetic marker data from years 2014 to 2016 of the Genomes to Fields maize project described in Rogers et al. (2021) were used for predictive modeling. In this study, we included additional soil parameters for each field-testing environment in the US locations from the USDA-NCRS Soil Survey Geographic Database (Soil Survey Staff 2021) obtained using the package soilDB (Beaudette et al. 2021) in R (R Core Team 2020) (Supplementary File 1). Parameters measuring soil particle size, water holding capacity, slope, and erosion factors were obtained for all soil horizons up to 2 m of depth (Supplementary Table 1). Soil horizons are layers defined by physical, chemical, and biological properties. The proportion of each horizon in the first 2 m of soil was computed and then used as a weight in computation of weighted values across horizons for each soil parameter. Missing values for coarse and fine silt representative value (RV, siltco_r, and siltfine_r) correspond to 0 values and were imputed as 0 where missing. Values for exchangeable cations (cec7_r), NH4OAc extracable bases (sumbases_r), and exchangeable hydrogen ions (extracid_r) were observed to be missing in very few cases and imputed using the R package mice (van Buuren and Groothuis-Oushoorn 2011). The USDA Soil Survey does not include data outside of the United States, therefore, to include these parameters in predictive modeling only the 59 US environments were used for this study, dropping 6 yield trial environments in Ontario, Canada, and leaving 16,106 hybrid-environment yield BLUEs for analysis, all of which had corresponding genetic marker data, weather data, and soil data.

Marker matrices

The imputed and filtered set of 20,373 SNP marker calls described in Rogers et al. (2021) was used as the starting genotypic data set for this analysis. Additive allele calls were recorded as counts of the minor allele (0, 1, 2). Dominance genotype calls were derived from the same matrix and called each homozygote as a 0, and heterozygous genotypes as 1 (Vitezica et al. 2013; Muñoz et al. 2014). Each set of marker scores was centered and scaled within loci such that marker i followed a distribution where $m_i \sim N(0,\ 1)$.

Models

Models were fit using R 4.0 (R Core Team 2020) using the North Carolina State University High Performance Computing Cluster (NCSU HPC) Henry2 cluster with the package BGLR (Pérez-Rodríguez and de los Campos 2010), using a custom shell script for batch submission of jobs (Supplementary File 2) and a custom R script for job execution (Supplementary File 3). The Henry2 cluster is a heterogeneous Intel Xeon based Linux cluster, and compute nodes include a mix of several generations of Intel Xeon processors in dual-socket blade servers. Nodes containing the same generation of processor may have varying amounts of memory. Core counts for nodes range from 8 to 32, and memory ranges from 16 to 512 GB. For the purposes of this high throughput calculation, jobs were placed on “first available” nodes having at least 70 GB of RAM. Jobs were submitted using a shell script (Supplementary File 2), which takes arguments for the number of folds to execute, what type of model to run, where to direct output, and what type of cross-validation to run. The jobs submitted using this script then execute the R script (Supplementary File 3) for a given fold, loading a workspace custom for the model type run. All data used for prediction can be found in the folder R_data_objects (Supplementary File 4). Post hoc analyses were done in R (Supplementary File 5).

Models were built strategically to answer questions about what components are useful for building complex G × E models for prediction. Data sets available to use for predictive modeling include both additive and dominance marker matrices of 20,373 markers, windowed environmental data sets covering 5-, 10-, 15-, and 30-day windows along with soil data derived from the USDA soil survey, and trait BLUEs from stage 1 analysis of each environment that accounted for differences in experimental design (Rogers et al. 2021). With these components available, we were able to ask several questions regarding what components would be most useful for building a G × E model—including the necessary number of markers, resolution of environmental data, and what type of G × E term was the most useful for modeling. A model using the complete data would involve 20,373 additive marker coefficients, 20,373 dominance marker coefficients, 377 environmental covariates at the higher resolution of 5-day windows, and 2 G×E terms with 7,680,621 interaction covariates each, for A×E and D×E effects, respectively. This model would require more RAM and computing time than reasonable for a plant breeding program, thus we investigated subsets of this full model to reduce memory and time requirements.

To reduce the complexity of the parameter space, we broke this problem into several steps. We first compared the ability of marker matrices of size 5,093 markers, 10,153 markers, 15,280 markers, and 20,373 markers to predict genetic main effects. Next, models utilized the chosen markers and added environmental data summarized into 1 of 4 different temporal window sizes to determine what resolution of environmental data provided the highest prediction ability for environment-specific hybrid values. Finally, the selected genetic and environmental components were used to create G × E models for environment-specific prediction. This allowed us to test a reasonable subset of all possible models that could be created for their performance in environment-specific prediction while addressing relevant questions to what components were necessary for creation of a useful G × E model.

Genetic main effects

To determine if genotype effects could be accurately modeled with a smaller number of markers than the complete set of 20,373, we generated subsets with 5,093, 10,153, or 15,280 markers by randomly sampling 1 time each from the original data set. Then models for genetic main effects were fit on hybrid yield BLUEs averaged across environments from Rogers et al. (2021). Models of the following form were compared for their ability to predict genetic main effects:

$$\mathbf{y} = \mu + \mathbf{M}b + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is the vector of yield BLUEs averaged across environments, μ is the grand mean, $\boldsymbol{M}$ is a marker matrix containing centered and scaled additive ($\boldsymbol{A}$) marker or dominance ($\boldsymbol{D}$) marker calls with row dimension equal to the number of unique hybrids (1,916) and column dimension equal to the number of markers

(5,093, 10,153, 15,280, or 20,373), b is the vector of marker effects, and $\boldsymbol{\varepsilon}$ is the vector of residual effects.

In addition, models with both additive and dominance effects were tested, using equal numbers of markers for each effect:

$$\mathbf{y} = \mu + \mathbf{A}b + \mathbf{D}c + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$, μ, and $\boldsymbol{\varepsilon}$ are the same as described previously, but both additive ($\boldsymbol{A}$) marker and dominance ($\boldsymbol{D}$) marker matrices of identical dimensions were fit, and b and c are the vectors of marker additive and dominance effects, respectively. All marker effects were fit under Bayesian Ridge Regression using the BGLR package (Pérez-Rodríguez and de los Campos 2010).

Within each class of model complexity (defined by the total number of marker effects modeled), we chose the model with higher average prediction ability using a single replication of 10-fold cross-validation in which a random set of 10% of the hybrids were placed in the test set for a given fold. Prediction ability within each test set was computed as $\mathrm{Cor}(\hat{y}_i,\ \hat{y}_{i,\mathrm{BLUE}})$, the correlation between the predicted value for hybrid i and the BLUE for hybrid i. Prediction abilities were averaged over folds. Then, using the simplest class of models, which involved a single marker matrix of 5,000 markers as a baseline, we selected models with higher complexity if they increased prediction ability by at least 1% compared with simpler models. By this process, we selected the matrix of 10,153 marker dominance effects for use in subsequent models.

Environment main effects

Next, we compared the ability of models incorporating environment effects to predict environment-specific hybrid BLUEs. Models had the form:

$$y = \mu + \mathbf{M}b + \mathbf{E}c + \varepsilon,$$

where y is the vector of hybrid-environment combination yield BLUEs, $\mathbf{M}$ is the matrix of 10,153 marker dominance coefficients, b is the matrix of marker effects, $\mathbf{E}$ is either a matrix of environment label dummy variables or the matrix of environmental variables summarized within 5-, 10-, 15-, or 30-day windows plus additional soil parameters, c is the vector of environment main effects or environmental variable effects, and ε is the vector of residual effects. Marker effects were modeled with Bayesian ridge regression; environment label main effects were modeled as fixed effects, or environmental covariables were modeled with a distribution of effects under LASSO (Tibshirani 1996; Park and Casella 2008; de los Campos et al. 2013).

Models were compared based on the mean ability to predict hybrid-environment BLUEs from a single year held out from the training set, averaged over the 3 possible training-test set combinations, chosen as a challenging scenario for prediction in untested environments. The environmental window with the highest mean predictive ability across the 3 years was utilized as the environmental component in subsequent G × E modeling.

G × E effects and dimension reduction methods

To make inclusion of G × E effects more computationally tractable, dimension reduction was used to model the interaction between the genetic and environmental parts of the model. Without dimension reduction, the G × E term would involve 3,828,058 predictors (derived from all combinations of 10,153 markers and 377 environmental covariates from the selected environment variable matrix), which would require impractically long computation times. Dimension reduction can be applied to either the marker or environment matrices prior to creation of the matrix of G × E effects. To compare dimension reduction of the genetic effects to that of environmental effects, we kept 1 of the 2 effect matrices intact and applied principal components (PCs) analysis to the other matrix, retaining a number of PCs such that the product of genetic and environmental effects was always maintained at approximately 100,000 effects. Specifically, this was achieved by retaining all 10,153 marker dominance effects along with the first 10 PCs of the environmental variable matrix and computing a matrix $\mathbf{GE_{PC}}$ with 101,530 columns by multiplying each column of the marker matrix by each column of the environmental PC matrix element-wise. Alternatively, we multiplied each column of 377 environmental variables by each of the first 265 PCs of the marker matrix to form a matrix $\mathbf{G_{PC}E}$ with 99,905 columns. The first 10 PCs of the environmental variable matrix accounted for 60% of the total variance of that matrix, and the first 265 PCs of the marker matrix accounted for 70% of the total marker variation. We accounted for computation time for dimension reduction when comparing computation time for different models.

Models incorporating G × E effects had the form:

$$y = \mu + \ \mathbf{M}b + \mathbf{E}c + \mathbf{GE_{PC}}d + \varepsilon$$

or

$$y = \mu + \ \mathbf{M}b + \mathbf{E}c + \mathbf{G_{PC}E}d + \varepsilon,$$

where terms are the same as defined in the previous section, with the addition of the matrix of marker-by-environment predictors ($\mathbf{GE_{PC}}$ or $\mathbf{G_{PC}E}$) and the vector of marker-by-environment effects (d). Marker effects were estimated under Bayesian Ridge Regression, whereas environmental covariate main effects and G × E effects were each estimated with LASSO with BGLR (Pérez-Rodríguez and de los Campos 2010).

These 2 models were then compared under different cross-validation scenarios, as described below.

Computation of prediction ability and statistical bias

Our goal was to evaluate different models for prediction ability under different scenarios defined by the scheme to separate training from test data sets. In each case, we held out the specified proportion of hybrid-environment BLUEs from the training set, and then measured prediction ability as the mean correlation over folds between predicted values and observed BLUEs within the held-out test set. The prediction ability correlation within each fold was measured in 2 ways: (1) correlation across environments measured as a single correlation value involving all held-out observations and (2) within-environments where correlation was measured for each environment on all held-out observations in said environment.

Within-environment bias was estimated within each test set to determine if models had bias toward over- or underpredicting held out observations within each environment:

$$\mathrm{Bias} = \bar{x}_{\mathrm{Predicted}} - \bar{x}_{\mathrm{Observed}},$$

where $\bar{x}_{\mathrm{Predicted}}$ is the mean of predicted values for the test set in a particular environment and $\bar{x}_{\mathrm{Observed}}$ is the mean of observed values for the test set in that environment.

Negative bias indicates that the model systematically predicts yield values lower than observed in the given environment, and positive bias indicates prediction of yield higher than observed on average. Within-environment bias relates to how accurately environment main effects are captured by the model. We also estimated the slope of the regression of observed values on predicted values within each site, which indicates the relative shrinkage or expansion of predictions compared with observed values, and which has also previously been referred to as a measure of bias in prediction (Daetwyler et al. 2013). Here, we refer to this statistic as slope to distinguish it from bias in predicting the overall environment mean.

Cross-validation and sampling schemes

Prediction ability of the different models was measured under a variety of cross-validation and sampling schemes that mimic prediction scenarios encountered in plant breeding programs (Fig. 1a). The first 2 methods (CV1 and CV2) were proposed by Burgueño et al. (2012) and are designed to mimic 2 situations often encountered by plant breeders. We also propose additional sampling scenarios that approximate other scenarios that may often be encountered by plant breeders. Code for all sampling schemes is in Supplementary File 3.

Baseline sampling

The different sampling schemes resulted in different sizes of training data sets, so we first measured the effect of reducing the number of observations within training sets on prediction ability by using random sampling to place 10%, 20%, 30%, 40%, 50%, or 60% of the observations in the test set, such that the model was trained on the remaining data. Baseline sampling was done such that a specified percentage of observations were placed in the test set using random sampling. This was repeated 10 times for each hold-out percentage to compute a mean prediction ability.

Cross-validation 1

Cross-validation 1 (CV1) is designed to evaluate prediction of genotypes that have not yet undergone field evaluation, such as in the case of newly developed lines (Burgueño et al. 2012). In this scenario, all observations of a random set of 10% of the hybrids were held out from model training, and the trained model was used to predict the values of this 10% of hybrids at all environments in which they were actually observed (Fig. 1a).

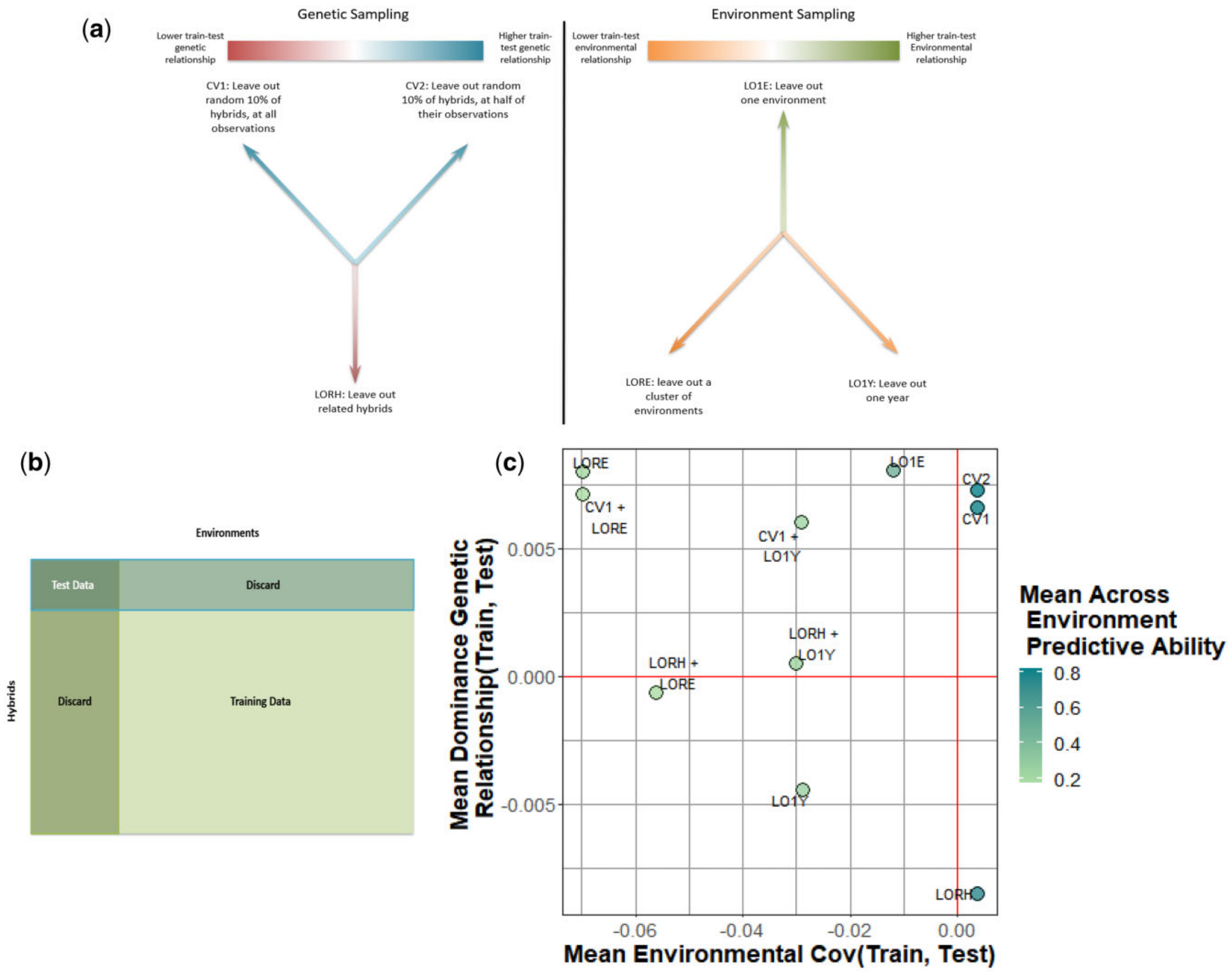

Fig. 1. Cross-validation schemes involved either random or stratified sampling of hybrids or environments. a) Schematic of bidirectional cross-validation structure designed to leave out sets of hybrids and sets of environments at the same time. b) Diagram of training and test data for bidirectional sampling schemes. A set of hybrids and environments were held out from training, and prediction accuracy was computed in the data set representing the intersection of hybrids and environments held out from the training data. c) Average environmental and genomic similarities between training and test sets for each sampling scheme and mean prediction ability from model using interactions of PCs of the marker matrix and all environment variables [PCA(Markers)*Env].

Cross-validation 2

Cross-validation 2 (CV2) is used to approximate the scenario where genotypes are evaluated in a subset of environments (Burgueño et al. 2012). In this scenario, half of the existing observations of 10% of the hybrids from CV1 are held out as part of the test set, while the other half are used in training (Fig. 1a). The half of observations used in training will be part of the test set in a different, nonoverlapping fold. This scheme allows for the correlation of performance between environments to aid in prediction of performance at another related environment.

LO1Y: leave out 1 year

Leave out 1 year (LO1Y) cross-validation was designed to ask the question: How well can environment-specific performance be predicted from training data collected in different years than the test environments? In the practical breeding situation, historical data are available to predict future performance. In our case, we used data from 2 years as training data and the held out third year was the test set (Supplementary Table 2; Fig. 1a).

LORE: leave out related environments

To test the impact of reducing the similarity between testing and training environments on the ability to predict environment-specific performance, we used the 7 environment clusters defined by weather variables described in Rogers et al. (2021) to define training and test sets (Supplementary Table 2). Each environment cluster was used once as the test set, with model training performed on the other 6 environment clusters. If environmental similarity between training and test sets is important for prediction ability, we expect prediction ability to decrease more in this case than when holding out random sets of environments (Fig. 1a).

LORH: leave out related hybrids

Practical breeding programs often introduce new breeding families, prompting the question of how well environment-specific performance of new germplasm subpopulations can be predicted from training data on distinct germplasm. To measure the change in prediction ability due to genetic differentiation between training and test sets, we used the 10 marker-defined hybrid clusters identified by Rogers et al. (2021) to define training and test sets (Fig. 1a). In this scenario, we used environment-hybrid BLUEs from 9 of the 10 hybrid cluster as training data and environment-hybrid BLUEs from the 10th cluster as test data, repeating this process for all 10 clusters as the test set (Supplementary Table 2).

LO1E: leave out 1 environment

A possible practical implementation of environment-specific prediction is to predict performance within a location that is in the same year and geographically similar to a relatively large training set. By careful stratification of testing sites, a breeding program could potentially leverage training data to predict in untested locations within the range of the training sets (Fig. 1a). To evaluate prediction ability in this scenario, we used leave 1 environment out cross-validation, wherein environment-hybrid BLUEs from 58 of the environments are used for model training and prediction ability is measured within the single held-out environment (Supplementary Table 2). This process is repeated by using each individual environment as the test set in turn.

Bidirectional cross-validation methods

The following CV methods combine 2 directions of data censoring to create training and test sets, they are designed to examine problems commonly encountered in breeding that can result in loss of prediction accuracies. Each method censors a set of hybrids, either leaving out 10% of hybrids at random using the folds from CV1 or in a stratified manner by leaving out genetically defined hybrid clusters, and also censors a set of environments by either leaving out 1 entire year of data or an environmental cluster (Fig. 1a). The training data in these cases are comprised of data not in the set of held out hybrids or environments, i.e. the complement of the union of the held-out hybrids and environments (Fig. 1b). The paired test set is the intersection of the held-out hybrids and environments (Fig. 1b). This results in differing levels of similarity between training and testing environments, and between training and test set genetic composition (Fig. 1c). In a few cases where individual test environments contained less than 8 hybrid observations, we did not estimate the within-environment prediction ability.

CV1 + LO1Y: leave out 10% of hybrids and 1 year of data

To simulate prediction of similar genetic materials in future years, we leave out any observations in the union of a random hybrid fold and the observations from a single year (expected and actual number of folds = 30). For example, when the test set of hybrids is the first CV1 fold and the test year is 2014, 470 observations of hybrids from fold 1 in 2014 are censored, all 4,193 additional observations from 2014 are also held out to fully mask year 2014 from the training set, and all observations of the held out 470 hybrids from years 2015 ($n = 505$) and 2016 ($n = 989$) were also censored to fully mask the test set genotypes from the training set. In this case, a total of 5,652 observations are held out of training, and of these, 470 compose the test set while the remainder are discarded (Fig. 1b; Supplementary Table 2).

LORE + LO1Y: leave out related hybrids and 1 year of data

To test the problem of predicting performance within a new genetic group in a different year than the training set, we created training sets that leave out any observations in the union of a related hybrid cluster and a single year of data, paired with test sets that include the intersection of the hybrid cluster and year. This uses the LORH folds along with the year fold. For this set, the actual number of folds (28) is lower than the potential number of folds (30, from all combinations of 3 years and 10 hybrid clusters), because none of the hybrids from cluster 3 were planted in 2014, and none of the hybrids from cluster 10 were planted in 2016 (Supplementary Table 2).

CV1 + LORE: leave out 10% of hybrids and related environments

This method is designed to test environment-specific prediction ability of germplasm related to the training set but evaluated in distinct environments training sets for this scenario hold out a random 10% of hybrids (from a CV1-fold) and 1 environmental cluster, paired with a test set composed of the intersection of the test hybrids and environment cluster (Supplementary Table 2).

LORH + LORE: leave out related hybrids and environments

The most challenging scenario for prediction is the idea of predicting performance of new germplasm in untested environments. For this scenario, training sets leave out the union of a hybrid cluster and an environment clusters, and their paired test sets include the intersection of these hybrid and environment clusters (Supplementary Table 2).

Influence of genetic and environmental similarity between training and test sets on prediction ability

We characterized the genetic similarity between training and test sets for each prediction scenario using the mean dominance genomic relationship coefficient (estimated in Rogers et al. 2021). We also computed the covariances between all pairs of environments using the complete set of scaled environmental variables and used these covariance values as measures of environmental similarities (Fig. 1c). We then performed ANOVA on the 30 prediction ability mean values from the 30 combinations of prediction models and cross-validation schemes using following model:

$$\mathrm{PA}_{ij} = m_i + \bar{g}_j + \bar{e}_j + \varepsilon$$

where m_i is the model (G + E, PC(Markers)*Env, or PC(Env)*Markers), $\bar{g}_j$ is the mean dominance genetic relationship between the train and test sets for sampling strategy j, $\bar{e}_j$ is the mean environmental covariance between train and test sets for sampling strategy j, ε is the residual error, and PA_{ij} is the mean across-environment prediction ability for model i, with mean genetic and environmental covariances $\bar{g}_j$ and $\bar{e}_j$, respectively.

Results and discussion

Genetic main effect prediction

Models for genetic main effects fit using additive marker matrices, dominance marker matrices, and combinations of the 2 demonstrated that using dominance marker matrices resulted in a mean prediction ability increase of 8.2% ($P < 2 \times 10^{-16}$) compared with using an additive marker matrix of with the same marker number (Fig. 2). These results are consistent with Rogers et al. (2021), who demonstrated that although the additive genetic variance component was greater than the dominance variance component for yield in the complete data set used here, GBLUP models utilizing a dominance relationship matrix had better fit to the data than their additive counterparts, and also predicted marginal genotype mean values better.

Fitting both **A** and **D** together did not improve prediction ability compared with fitting **D** of the same marker number alone ($P = 0.423$) (Fig. 2). Rogers et al. (2021) previously demonstrated a moderately high correlation between the additive and dominance relationship matrices ($R_{\mathrm{off-diag}} = 0.83$, $R_{\mathrm{diag}} = 0.54$). Increasing marker number did not increase predictive ability for models using **A** alone, but resulted in predictive ability increases with diminishing returns for models utilizing **D** (Fig. 2). Increasing from 5,093 to 10,153 dominance markers resulted in a prediction ability increase of 1.1% points on average, while increasing from 10,153 to 15,280 markers resulted in an additional increase of only 0.5%, and adding markers beyond 15,280 did not result in a consistent increase in prediction ability. From these results, we selected the dominance marker matrix with 10,153 markers to represent hybrid genetic effects for all subsequent environment-specific prediction modeling.

Environment main effects

Weather variables were summarized into 5-, 10-, 15-, and 30-day windows along with 23 variables representing soil parameters. We compared the use of summary values of weather variables with differing temporal resolution to model environment main effects along with the 10,153 marker dominance effects when predicting hybrid-environment mean values. All models were compared using a "leave out 1 year" CV (LO1Y) strategy, in which a single year of data was left out as the test set, to provide a challenging scenario for prediction using environment main effects.

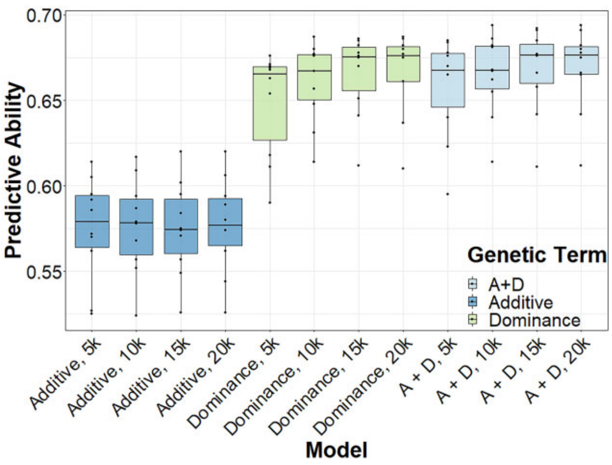

Fig. 2. Prediction ability for hybrid yield main effects using either additive (A) or dominance (D) marker coefficients or both (A + D) with differing numbers of markers (5,093, 10,153, or 15,280, or 20,373) measured using 10-fold cross-validation.

Models using the 5-day window set had the best performance of the models tested, improving environment-hybrid prediction accuracies by 30%, 19%, and 5% compared with environment labels in test years 2014, 2015, and 2016, respectively (Fig. 3). Therefore, the 5-day window size was used for weather variables in subsequent G × E models to serve as the **E** portion of the modeling efforts. Larger weather data window sizes resulted in increased prediction ability for the 2014 and 2015 data, but decreased ability for the 2016 data compared with environment labels. This indicates that addition of environmental variables helps for GP ability, but the decrease in prediction ability when predicting observations in year 2016 is likely due to differences between 2016 and previous years in terms of population composition and weather (Rogers et al. 2021).

Adding G × E interactions to prediction models

Baseline prediction ability

Downsampling training size had only a small effect on overall prediction ability. For example, removing 60% of observations from the training set resulted in an average decrease in prediction ability of 1.9% compared with holding out 10% of observations (Fig. 4). These results indicate that observed differences in prediction ability from stratified train-test sampling schemes to be discussed subsequently would be almost entirely due to the stratification itself rather than due to training set sample size. Results also demonstrated that the effect of training size was not consistent among all test environments, with some test environments having little to no change in prediction ability when reducing training population size. This is likely because the G2F experimental trials were neither balanced for genotype composition, nor evenly distributed geographically. As more data are removed from the training set, less densely sampled geographies experience faster drops in prediction ability in part due to reduced covariance to the training set (Supplementary Fig. 1).

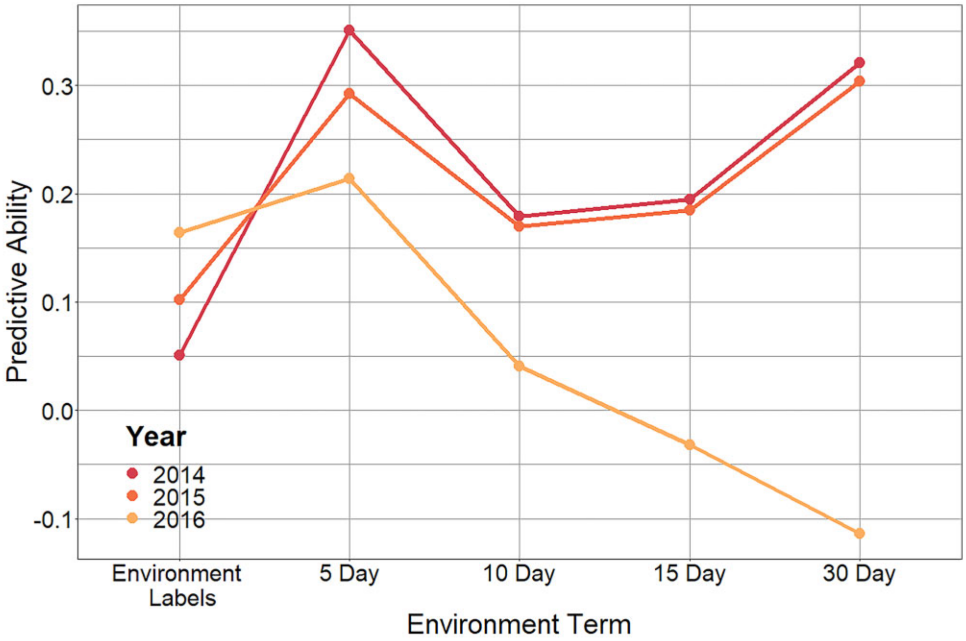

Fig. 3. Hybrid-environment yield prediction ability of models using 10,153 dominance marker coefficients and environment labels or environmental variables summarized in 5-, 10-, 15-, or 20-day windows measured by holding out a full year of data as a test set.

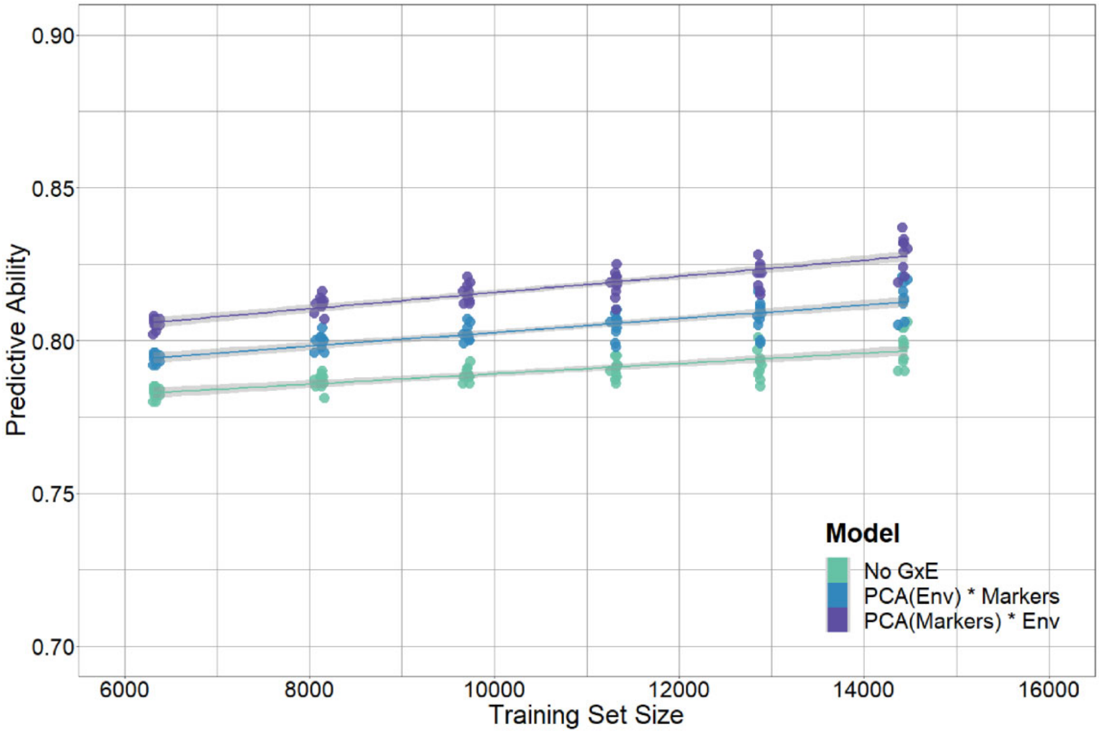

Fig. 4. Baseline prediction ability of hybrid-environment yield performance from 10-fold cross-validation using different percentages of randomly sampled training and testing data, using a model with 10,153 marker dominance coefficients, weather variables summarized in 5-day windows, and G × E effects computed from interactions of the PCs of the environment variables and all markers [PCA(Env)*Markers] or from interactions of PCs of the marker matrix and all environment variables [PCA(Markers)*Env]. Each model and percentage of train-test split was run on 10 random samples. Prediction ability was measured on values across all environments.

Does the dimension on which reduction occurs matter?

On average, the addition of the G × E term using PCA on the markers resulted in a 2.7% increase in prediction ability, while addition of the G × E term using PCA for dimension reduction on the environmental data increased prediction ability by 2.0% in the baseline cross-validation scheme (Fig. 4). These results indicate that while both additions of G × E result in an increase in prediction ability, the model reducing dimension of the genetic component better captures the G × E effects observed in the data.

This could stem from the retained PCs of the environmental data capturing a smaller percentage of the observed variance than the marker PCs (60% compared with 70%), or from PCA not efficiently weighting the environmental covariates. G × E models that utilize crop modeling to determine the most important environmental variables to include in GP models allow a priori preselection of environmental variables (Lobell et al. 2013; Heslot et al. 2014; Bustos-Korts et al. 2019; Hammer et al. 2019). In contrast, we used LASSO on the environmental covariable main effects and their interactions with markers in the hope that irrelevant environmental variables and interactions would have their effects shrunk to 0 empirically during the training model fit. Dimension reduction by PCA on the marker data is perhaps more reasonable for highly polygenic traits that are controlled by many variants with nearly equal very small effects, than it is on the environmental data, where the assumption of similar importance among variables is not grounded in biology.

Within-environment prediction results for the baseline scenario that randomly sampled different training set sizes showed that addition of the PCA(Markers)*Env G × E term increased mean-within environment prediction ability by 6.9% on average. Including this G × E term in the prediction model was more helpful for certain types of environments with greater deviation from the average G2F environment, which would have a score of approximately 0 for each factor in previous environmental analyses (Rogers et al. 2021). On the FA biplot of the first 2 factors, this average G2F environment would be located within the temperate Northeastern, Midwestern, and Southern Corn Belt groups (Rogers et al. 2021). This follows from interpretation of the G × E term as genotype-specific deviations from the expected value in a given environment (Pérez-Rodríguez et al. 2017). Addition of the G × E term resulted in a negligible effect on within-environment ability for midwestern and northern environments, but increased prediction ability across all training sizes for the dry plains environments (KSH2_2016, TXH2_2014, and TXH2_2015), and increased prediction ability for GAH1_2014 and TXH1 in 2015 and 2016 (Supplementary Fig. 1). Prediction ability was reduced in 13 environments, with a maximum reduction of 10.3% (TXH1_2014). Addition of G × E interactions did not improve predictive ability of GAH2_2016 (decrease of 6.7%) or ILH1_2016 (increase of 0.5%), indicating that the negative genetic correlation for yield between these 2 environments and other environments reported by Rogers et al. (2021) limited recovery of useful information for genetic effects in these unusual environments (Howard et al. 2019).

Does the effect of G × E interactions in prediction models depend on the relationships between training and test sets?

The G2F hybrid clusters, identified on the basis of genetic similarity by Rogers et al. (2021), include diverse sets of hybrids with some overlap in parentage, which likely provides great enough covariance to aid in prediction of the related hybrids left out in LORH. Furthermore, although hybrids were not randomized to environments, stratification by hybrid clusters did not greatly reduce the environmental covariance between training and test sets (Fig. 1c). All schemes where whole environments or years were left out (LORE, LO1Y, and the bidirectional methods) led to decreased prediction ability. Prediction ability for new years or distinct environment clusters is relatively poor (Fig. 5). Although the maize Genomes to Fields experiment includes a relatively large sample of environments, it remains insufficient for training robust models that extrapolate to new environments, at least without incorporating crop model-guided environmental variable selection. The 2 bidirectional CV schemes leaving out related environments (CV1 + LORE: leave out related environments + 10% of hybrids, and LORH + LORE: leave out related environments and hybrids) performed similarly to LORE despite the test set being comprised of hybrids that were not seen in the training data (Fig. 5).

Addition of environmental covariates and G × E terms increased prediction ability across CV1, CV2, and LORH sampling schemes, but did not aid prediction ability in more complex sampling schemes (Fig. 5). On average, CV1 across environment predictive ability for the G × E model using PC(Markers) was 80.8%, a gain of 8.0% over the G + E model (Fig. 5; Supplementary Table 3). The PC(Env) gave an increase of prediction ability of 4.6% over the G + E model, but lagged behind the PC(Markers) model, similar to observations in the baseline test. Across-environment prediction ability for CV2 for the full model was very similar to the randomly sampled model of comparable size. This method had the highest prediction ability across environments, likely because the test set hybrids were present in the training set. Burgueño et al. (2012) note that CV2 is an easier prediction problem than CV1, but that it adds time to the generation interval because field testing is required for all selection candidates. Within environments in the CV2 scenario, the PC(Marker) G × E model had an average gain of 8.3% over the G + E model, compared with the PC(Env) model's average gain of 4.6% over the G + E model (Fig. 5; Supplementary Table 4). In the scenario where single environments were left out (LO1E), addition of the G × E terms increased prediction ability slightly in comparison to the G + E model. This indicates that having enough related environments allows the G × E model to make gains in prediction accuracy. Observations of no increase in prediction ability from the G + E model to either G×E model in scenarios where test environments are entirely excluded from training sets indicate that while G×E information can improve prediction ability in some cases, it was not helpful when the environments in the test sets were not sampled directly in the training sets. This suggests that our estimates of specific marker-by-environmental covariable interaction effects cannot be extrapolated beyond the specific environments in which they were estimated.

Results from the analysis of prediction ability as a function of genetic and environment similarity between training and test sets demonstrate that mean environmental covariance between the train and test sets was very important ($p = 2.6 \times 10^{-10}$) for across environment prediction ability (Table 1). In contrast, prediction model and genetic similarity had no significant effect on mean prediction ability (Table 1). The genetic relationships among the G2F hybrids are highly complex, as demonstrated by Rogers et al. (2021). Therefore, even separating the hybrids by genetically defined clusters was not sufficient in this case to dramatically reduce prediction ability across groups.

Where does modeling G × E help the most?

Although addition of the GxE term did not increase overall prediction ability across sampling scenarios, we hypothesized that GxE effects might be useful in environments that deviate from the average G2F environment. This follows from the idea that G × E effects are considered to be environment-specific deviations from the stable genetic and environmental main effects (Lopez-Cruz et al. 2015), consequently we would then expect G × E to be most helpful in cases where the environment deviates more from the mean environment. The G2F environments represent an unbalanced sample with the number of temperate northern and

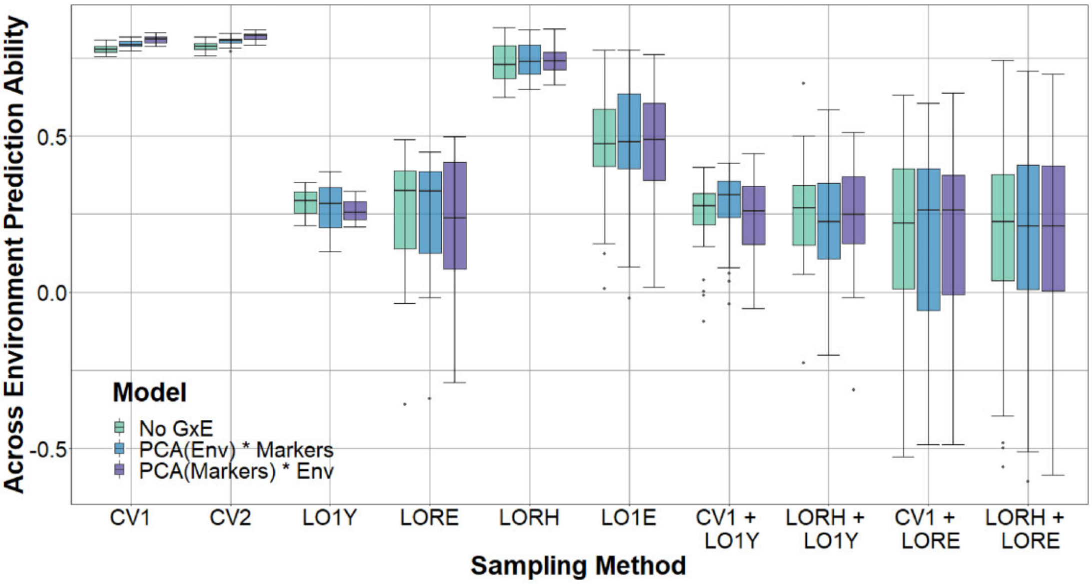

Fig. 5. Distributions of hybrid-environment yield prediction ability across cross-validation replicates for models including 10,153 marker dominance coefficients and 377 environmental covariates, along with no G × E effects or G × E effects computed using PCs of the environmental data [PCA(Env)*Markers]) or using PCs of the marker data [PCA(Markers)*Env]. Each model was evaluated in test sets selected by random sampling (CV1), partial replication across environments (CV2), leaving out a single year of data (LO1Y), stratification by environment clusters (LORE), by hybrid clusters (LORH), leaving out single environments (LO1E), and bidirectional censoring schemes leaving out both a year and 10% of hybrids (CV1 + LO1Y), a year and related hybrids (LORH + LO1Y), environment clusters with 10% of hybrids (CV1 + LORE), and environment and hybrid clusters (LORH + LORE).

Table 1. Analysis of variance of mean prediction ability as affected by training-test dominance genetic relationships, training-test environmental covariances, and a factor representing prediction model type.

Factor	df	Sum of squares	*F*-value	*P*-value
Model type	2	0.00007	0.0023	0.9977
Mean dominance genetic relationship (train, test)	1	0.00384	0.2502	0.6213
Mean environmental co-variance (train, test)	1	1.565	102.0054	<0.0001
Residual error	25	0.384		

midwestern environments outnumbering the more southern, humid environments, and dry plains environments. Gain when moving from G + E models to G × E models was present for most environments on average in simple sampling schemes (CV1, CV2, and LORH). Within environments under CV1, there was a large range of prediction abilities observed, with mean prediction abilities (averaged across folds) ranging from −0.06% (GAH2_2016) to 72.2% (WIH1_2016). Two environments, ILH1_2016 and GAH2_2016, had very low mean within-environment predictive abilities. This can be attributed to their small or negative correlations with other environments (Rogers et al. 2021). Under CV2, within-environment mean prediction abilities had a similar range to those observed in CV1, ranging from 12.3% (ILH1_2016) to 77.9% (WIH1_2016). In CV2, GAH2 has average within-environment prediction ability of 44.7%, indicating that presence of GAH2 data was helpful in improving prediction ability for an environment with little covariance to other environments in the dataset. In this scenario all of the environments had at least a small increase in prediction ability when adding G × E effects to the model, indicating that overlap of hybrids in the training and test sets aided prediction ability.

In particular G × E models aided prediction ability in the Georgia environments (GAH1) and the humid Texas environment (TXH1) (Fig. 6a). Smaller, but consistent increases were also observed in North Carolina (NCH1) environments when G × E terms were added to the model. Notably, introduction of G × E did not help in the dry plain environments (KSH1 and TXH2), which have a different type of G × E than the southeastern, humid environments.

In the case of more challenging sampling schemes, moving from a G + E model to a G × E model resulted in a slight decrease in prediction ability on average (Fig. 6b). This is likely because estimation of G × E effects in environments distinct from the test environments was not useful for extrapolation to the test environments and added noise to the model. Therefore, for the addition of G × E interactions to improve environment-specific prediction, the training environments must adequately represent the test environments.

How does bias change depending on sampling scheme?

Within-environment mean bias varied by environment, with some yield values being more likely to be under- or overestimated than others (Fig. 7). Train-test composition influenced within-environment bias more than the type of model (as was true for overall prediction ability), with the sampling schemes where the training set represented the test set well (CV1, CV2, and LORH) having within-environment bias values close to 0, and LORE having the most extreme within-environment bias values. This indicates that stronger positive genetic and environmental covariances between training and test sets help to reduce prediction bias. Addition of G × E terms tended to slightly reduced bias in environments where G × E aided prediction, but had little effect on bias in other environments, consistent with the assumptions of the G × E model estimating G × E effects as deviances of the most common differing environment.

The slope of regression of observed on predicted values also varied among environments but the distribution of these values was similar for different prediction models (Supplementary

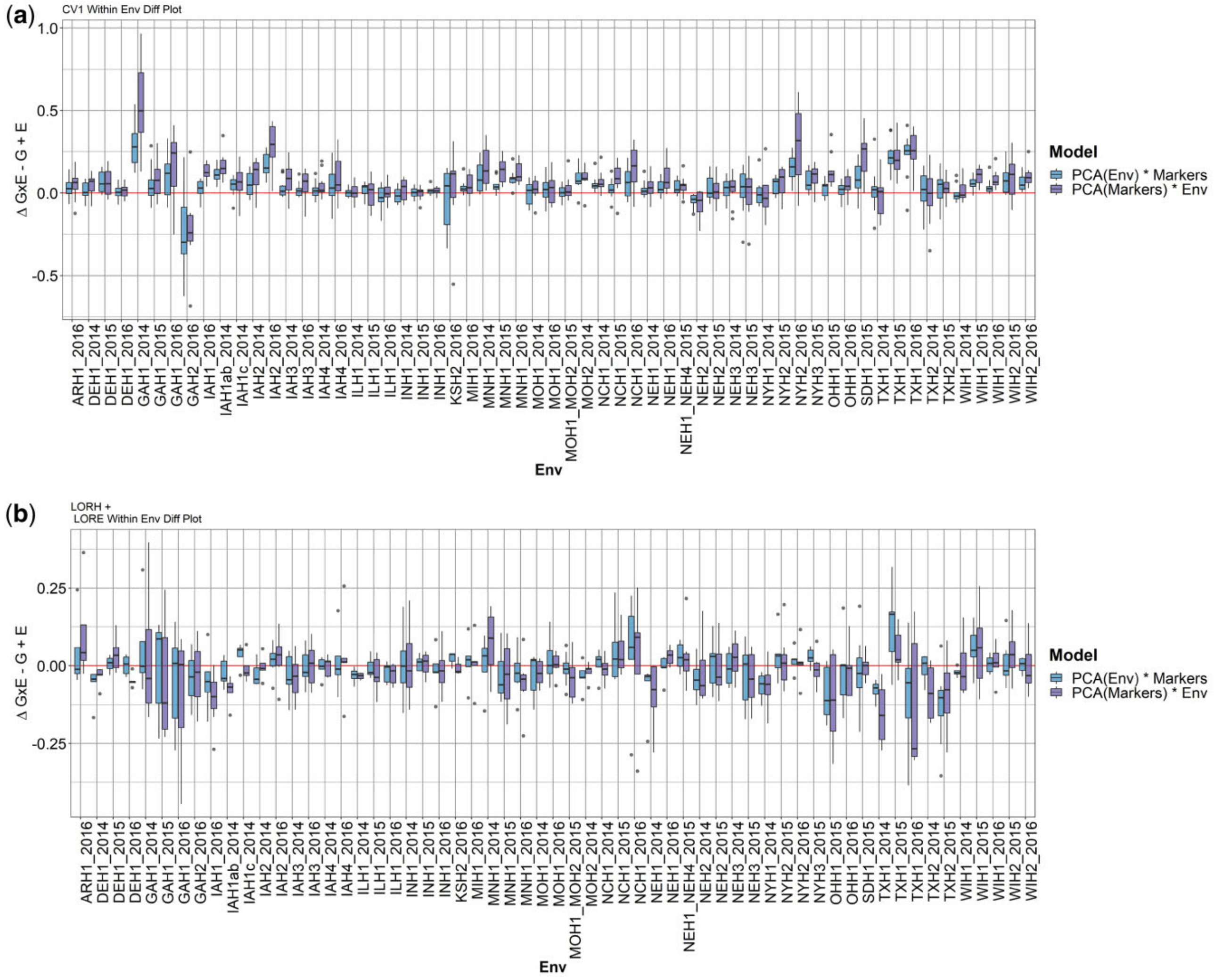

Fig. 6. Representative examples for 2 patterns of change in within-environment prediction accuracy when adding G × E interactions to the G + E model to as either PCA(Env)*Markers or PCA(Markers)*Env, using a) random cross-validation across hybrids scheme (CV1) or b) stratified sampling leaving out groups of related hybrids and related environments (LORH + LORE)

Fig. 2). The median value was slightly <1.0 for all methods and sampling schemes (indicating slight overdispersion of predictions), and the variation in the regression slopes increased when test sets had greater genetic distance from training sets (Supplementary Fig. 2).

Conclusions

Our results demonstrate that dominance effects are more important than additive effects for prediction of grain yield in this hybrid maize data set, and that utilizing both dominance and additive effects does not improve prediction ability. This comes with the caveat that our dominance matrix was parameterized such that marker calls measure heterozygosity (Vitezica et al. 2013; Muñoz et al. 2014). Other parameterizations may measure dominance such that addition of additive genetic effects would aid prediction ability.

The above results demonstrate that addition of environmental data as a measure of similarity between environments aids in environment-specific GP. The window size used matters, and can be selected empirically, although it is still unknown if identifying the optimal windows for any given covariate will substantially increase prediction ability. The parameter space for optimizing windows is vast and challenging to address in a way that is computationally feasible for the rapid turn-around between data collection and prediction of breeding values usually required for plant breeding programs. Jarquín et al. (2020) used hourly weather values summarized into an environmental relationship matrix for GP in a subset of these same experiments, but reported little gain in prediction ability even in CV1- and CV2-type scenarios, noting that this was in part because of the equal weighting of very high dimensional weather data for modeling. Our approach utilized a 5-day window size for summarizing weather variables, allowing for the use of each of these variables individually in prediction models while reducing computational burden and averaging out noise present within hourly and daily values. Additionally, the use of LASSO to model the environmental and GxE effects permits the model to learn which covariates are important to a given trait, demonstrating gain in predictive ability compared with the parameterization used by Jarquín et al. (2020). Use of these environmental covariates permits prediction of environment-specific performance, but the ability of such predictions depends greatly on the training data including environments similar to the new environment. However, environment-specific prediction is also a function of within-environment heritability, which limits prediction ability in environments where heritability is poor.

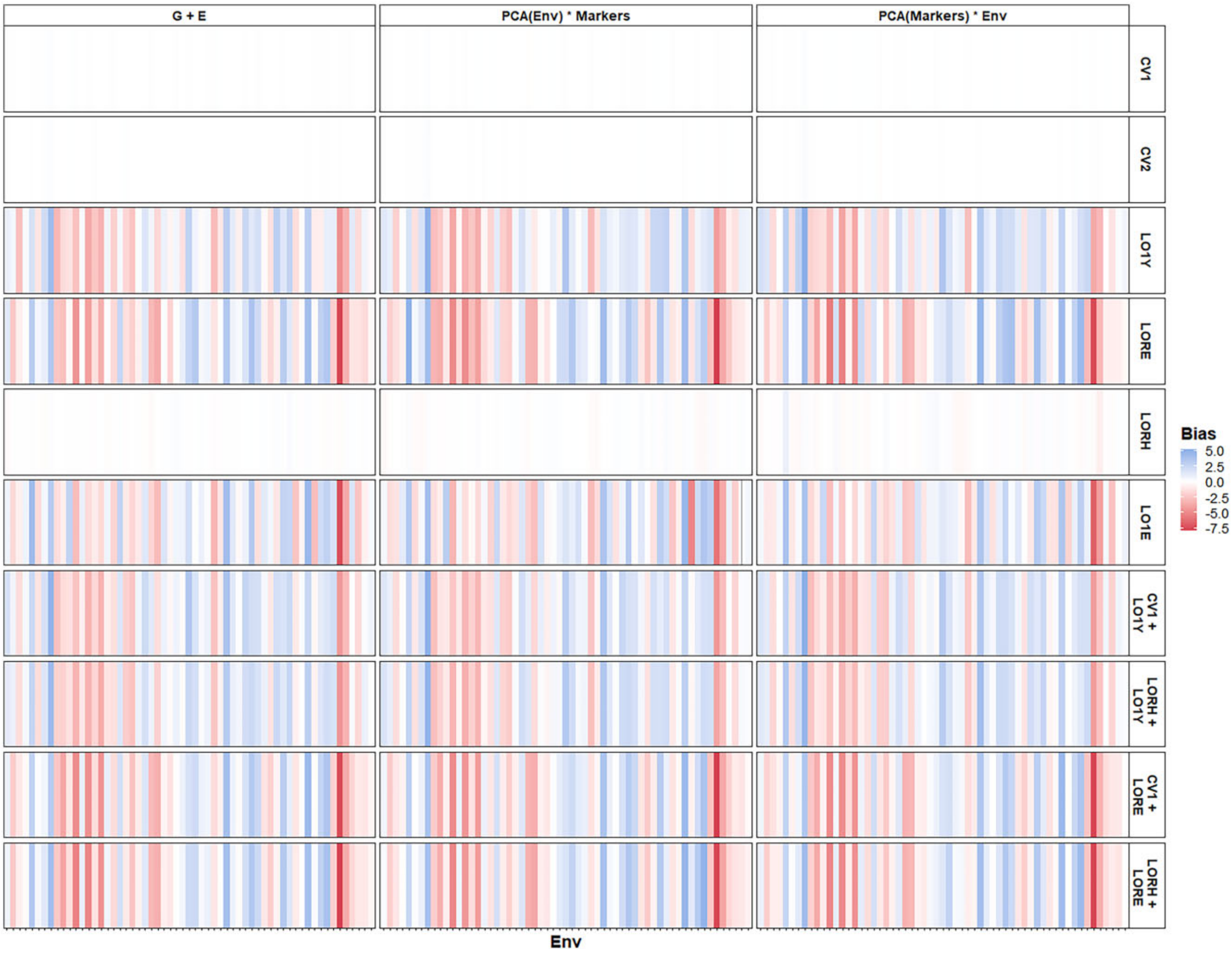

Fig. 7. Prediction bias measured as the difference between the mean of test set hybrids within 1 environment and the mean of the corresponding BLUEs at the same environments for hybrid-environment prediction models based on 10,153 marker dominance coefficients and 377 environmental variables alone (G + E) or with the addition of PC(Env)*Marker interactions or PC(Marker)*Env interactions. Each column is 1 of the 59 G2F environments, sorted alphanumerically. Each row corresponds to a cross-validation sampling scheme.

Our G × E models aim to enable environment-specific prediction such that both E and G × E effects for new environments can be estimated using available environmental data, differing from traditional GS models in that to predict performance in a new environment only marker data for the genotypes of interest and historical environmental data are needed. This eases implementation by not requiring extensive field trials to estimate G × E and E effects in an environment prior to selection germplasm for said environment. Our approach can expand to other crops and may be especially useful in crops where G × E variance is important. We found that dimension reduction still allowed G × E to aid in prediction while allowing for models that were computationally tractable under reasonable constraints (no model was allowed to run for more than 24 h, and had to be able to run using less than 70 GB of RAM). The PC(Markers)*Env approach outperformed the PC(Env)*Markers approach under similar dimensionalities, indicating that dimension reduction on the marker side summarized information in such a way that allowed for better modeling of G × E effects than the PC(Env)*Marker counterpart.

Currently modeling G × E effects in a way that approximates the biological reality is challenging, largely because of computational limitations in both memory and the turnaround time needed by plant breeders to drive crossing decisions. The M × E approach is limited to estimation of linear G × E effects, whereas many covariates likely have a threshold or other nonlinear relationship with yield. For example, presence of drought is detrimental to maize development but high amounts of rain that may cause flooding are also detrimental to development meaning that the relationship between yield and rainfall tends to be quadratic. Neural network models have become more popular in recent years because of their abilities to handle such types of nonlinear relationships, but these types of models often require very large datasets, along with high amounts of RAM and computer processing power. These requirements mean that this approach is currently untenable for most plant breeders, especially those in the public sector. For most use cases, we would argue that the M × E type approaches here can provide reasonable prediction ability using resources commonly available to academic breeding programs such that they could implement this type of GP modeling.

It is important to note that modeling G × E will not rescue a breeder from poor sampling of the target populations of environments and genotypes of interest. G × E modeling will likely be most useful for programs that have a set of target environments for future lines that are looking to direct early stage material toward a specific target population of environments. It would likely also help in merging material across a larger network of environments, such as the case of breeders who would like to move material developed in 1 geography to another, both of which already have data. Framing G × E modeling to solve specific problems in plant breeding and identifying portions of breeding pipelines where G × E would be useful for driving decisions and optimizing in-field testing will be important to integrating it into the plant breeder's toolbox.

Data availability

Original trait, environmental covariate, and marker data were taken from: https://doi.org/10.25387/g3.12636095. Scripts and specific R objects used in the analysis can be obtained from https://doi.org/10.25387/g3.17209085. Supplementary File 1 is an R markdown script to extract soil data from the USDA-NCRS Soil Survey Geographic Database (Soil Survey Staff 2021). Supplementary File 2 is a bash script to submit multiple analysis jobs to a high performance computing load sharing facility. Supplementary File 3 is an R script that accepts input parameters to select training sets and model types. Supplementary File 4 is a zipped archive containing multiple data sets used in the analysis, including the trait data subset for this study, the additive and dominance marker matrices at various densities, and the weather variables summarized to different temporal window sizes. Supplementary File 5 is an R markdown script to compute and summarize prediction ability from analysis outputs.

Supplemental material is available at figshare DOI: https://doi.org/10.25387/g3.17209085.

Acknowledgments

We would first and foremost like to acknowledge the Genomes to Fields cooperators who carried out field trials and collected data that was used in this project, especially Natalia de Leon and Shawn Kaeppler (University of Wisconsin), and Cinta Romay (Cornell University), who was instrumental to generation of the genotype data used in this study. For computation, we acknowledge the computing resources provided on Henry2, a high-performance computing cluster operated by North Carolina State University. We also thank Lisa Lowe (NC State University) for their assistance with learning to use the Henry2 cluster, development of job submission scripts, and optimization of resource requests, which was provided through the Office of Information Technology HPC services at NC State University.

Funding

ARR was supported by the Genetics program at North Carolina State University and by the United States Department of Agriculture - Agriculture Research Service.

Conflicts of interest statement

None declared.

Literature cited

Adee E, Roozeboom K, Balboa GR, Schlegel A, Ciampitti IA. Drought-tolerant corn hybrids yield more in drought-stressed environments with no penalty in non-stressed environments. Front Plant Sci. 2016;7:1534–1539.

Bandeira e Sousa M, Cuevas J, de Oliveira Couto EG, Pérez-Rodríguez P, Jarquín D, Fritsche-Neto R, Burgueño J, Crossa J. Genomic-enabled prediction in maize using kernel models with genotype x environment interaction. G3 (Bethesda). 2017;7(6):1995–2014.

Beaudette D, Skovlin J, Roecker S, Brown A. 2021. soilDB: Soil Database Interface. R Package Version 2.6.10. https://cran.r-project.org/ (Accessed: 2021 December 29).

Burgueño J, de los Campos G, Weigel K, Crossa J. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 2012;52(2):707–719.

Bustos-Korts D, Boer MP, Malosetti M, Chapman S, Chenu K, Zheng B, van Eeuwijk FA. Combining crop growth modeling and statistical genetic modeling to evaluate phenotyping strategies. Front Plant Sci. 2019;10:1491–1421.

Comstock RE, Moll RH. Genotype-environment interactions. In: WD Hanson, HF Robinson, editors. Statistical Genetics and Plant Breeding. Washington (DC): The National Academies Press; 1963. p. 164–197. doi:10.17226/20264.

Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, Hausmann NJ, Wright D, Graham G. Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. Crop Pasture Sci. 2014;65(4):311–336.

Costa-Neto G, Fritsche-Neto R, Crossa J. Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity (Edinb). 2021;126(1):92–106.

Crossa J, de los Campos G, Maccaferri M, Tuberosa R, Burgueño J, Pérez-Rodríguez P. Extending the marker' environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Sci. 2016;56(5): 2193–2209.

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de Los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, et al. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017; 22(11):961–975.

Cuevas J, Crossa J, Montesinos-López OA, Burgueño J, Pérez-Rodríguez P, de Los Campos G. Bayesian genomic prediction with genotype x environment interaction kernel models. G3 (Bethesda). 2017;7(1):41–53.

Cuevas J, Granato I, Fritsche-Neto R, Montesinos-Lopez OA, Burgueño J, Bandeira e Sousa M, Crossa J. Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. G3 (Bethesda). 2018;8(4):1347–1365.

Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics. 2013; 193(2):347–365.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193(2):327–345.

Edwards JW. Genotype × environment interaction for plant density response in maize (*Zea mays* L.). Crop Sci. 2016;56:1493–1505.

Granato I, Cuevas J, Luna-Vázquez F, Crossa J, Montesinos-López O, Burgueño J, Fritsche-Neto R. BGGE: a new package for genomic-enabled prediction incorporating genotype x environment interaction models. G3 (Bethesda). 2018;8(9):3039–3047.

Hammer G, Messina C, Wu A, Cooper M. Biological reality and parsimony in crop models—why we need both in crop improvement! In Silico Plants. 2019;1(1):21.

Heslot N, Akdemir D, Sorrells ME, Jannink J-L. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet. 2014;127(2):463–480.

Hickey JM, Chiurugwi T, Mackay I, Powell W; Implementing Genomic Selection in CGIAR Breeding Programs Workshop Participants. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat Genet. 2017;49(9): 1297–1303.

Howard R, Gianola D, Montesinos-López O, Juliana P, Singh R, Poland J, Shrestha S, Pérez-Rodríguez P, Crossa J, Jarquín D. Joint use of genome, pedigree and their interaction with environment for predicting the performance of wheat lines in new environments. G3 (Bethesda). 2019;9(9):2925–2934.

Isik F, Holland JB, Maltecca C. Genetic Data Analysis for Plant and Animal Breeding. New York (NY): Springer; 2017.

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet. 2014;127(3):595–607.

Jarquín D, de Leon N, Romay C, Bohn M, Buckler ES, Ciampitti I, Edwards J, Ertl D, Flint-Garcia S, Gore MA, et al. Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. Front Genet. 2020;11:592769.

Krishnamurthy SL, Sharma PC, Sharma DK, Ravikiran KT, Singh YP, Mishra VK, Burman D, Maji B, Mandal S, Sarangi SK, et al. Identification of mega-environments and rice genotypes for general and specific adaptation to saline and alkaline stresses in India. Sci Rep. 2017;7(1):14.

Lasky JR, Upadhyaya HD, Ramu P, Deshpande S, Hash CT, Bonnette J, Juenger TE, Hyma K, Acharya C, Mitchell SE, et al. Genome-environment associations in sorghum landraces predict adaptive traits. Sci Adv. 2015;1(6):e1400218.

Li X, Guo T, Mu Q, Li X, Yu J. Genomic and environmental determinants and their interplay underlying phenotypic plasticity. Proc Natl Acad Sci U S A. 2018;115(26):6679–6684.

Li X, Guo T, Wang J, Bekele WA, Sukumaran S, Vanous AE, McNellie JP, Tibbs-Cortes LE, Lopes MS, Lamkey KR, et al. An integrated framework reinstating the environmental dimension for GWAS and genomic selection in crops. Mol Plant. 2021;14(6):874–887.

Lobell DB, Hammer GL, McLean G, Messina C, Roberts MJ, Schlenker W. The critical role of extreme heat for maize production in the United States. Nat Clim Change. 2013;3(5):497–501.

Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink J-L, Singh RP, Autrique E, de los Campos G. Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3 (Bethesda). 2015;5(4):5: 569–582.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–753.

Millet EJ, Kruijer W, Coupel-Ledru A, Alvarez Prado S, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, van Eeuwijk F, Tardieu F, et al. Genomic prediction of maize yield across European environmental conditions. Nat Genet. 2019;51(6): 952–956.

Montesinos-López A, Montesinos-López OA, Gianola D, Crossa J, Hernández-Suárez CM. Multi-environment genomic prediction of plant traits using deep learners with dense architecture. G3 (Bethesda). 2018;8(12):3813–3828.

Monteverde E, Gutierrez L, Blanco P, Pérez de Vida F, Rosas JE, Bonnecarrère V, Quero G, McCouch S. Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (*Oryza sativa* L.) grown in subtropical areas. G3 (Bethesda). 2019;9(5):1519–1531.

Monteverde E, Rosas JE, Blanco P, Pérez de Vida F, Bonnecarrère V, Quero G, Gutierrez L, McCouch S. Multienvironment models increase prediction accuracy of complex traits in advanced breeding lines of rice. Crop Sci. 2018;58(4):1519–1530.

Muñoz PR, Resende MFR, Gezan SA, Resende MDV, de los Campos G, Kirst M, Huber D, Peter GF. Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics. 2014; 198(4):1759–1768.

Park T, Casella G. The Bayesian Lasso. J Am Stat Assoc. 2008; 103(482):681–686.

Pauli D, Chapman SC, Bart R, Topp CN, Lawrence-Dill CJ, Poland J, Gore MA. The quest for understanding phenotypic variation via integrated approaches in the field environment. Plant Physiol. 2016;172(2):622–634.

Pérez-Rodríguez P, Crossa J, Rutkoski J, Poland J, Singh R, Legarra A, Autrique E, de los Campos G, Burgueño J, Dreisigacker S, et al. Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. Plant Genome. 2017;10(2):1–15.

Pérez-Rodríguez P, de los Campos G. 2010. BGLR: a statistical package for whole genome regression and prediction. https://cran.r-project.org/ (Accessed: 2021 December 29).

R Core Team. 2020. R: *A Language and Environment for Statistical Computing*. Vienna (Austria): R Foundation for Statistical Computing.

Rogers AR, Dunne JC, Romay C, Bohn M, Buckler ES, et al. The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. G3 (Bethesda). 2021;11:jkaa050.

Saint Pierre C, Burgueño J, Crossa J, Fuentes Dávila G, Figueroa López P, Solís Moya E, Ireta Moreno J, Hernández Muela VM, Zamora Villa VM, Vikram P, et al. Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. Sci Rep. 2016;6(1):27312.

Soil Survey Staff. Web Soil Survey. Natural Resources Conservation Service, United States Department of Agriculture; 2021. [accessed 2021 Dec 16]. http://websoilsurvey.sc.egov.usda.gov/.

Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc. 1996;58(1):267–288.

van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Soft. 2011;45(3):1–67.

Vitezica ZG, Varona L, Legarra A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics. 2013;195(4):1223–1230.

Voss-Fels KP, Cooper M, Hayes BJ. Accelerating crop genetic gains with genomic selection. Theor Appl Genet. 2019;132(3):669–686.

Yang J, Jin Z-B, Chen J, Huang X-F, Li X-M, Liang Y-B, Mao J-Y, Chen X, Zheng Z, Bakshi A, et al. Genetic signatures of high-altitude adaptation in Tibetans. Proc Natl Acad Sci U S A. 2017;114(16): 4189–4194.

Communicating editor: A. Lipka