Increasing Genomic-Enabled Prediction Accuracy by Modeling Genotype × Environment Interactions in Kansas Wheat

Diego Jarquín, Cristiano Lemes da Silva, R. Chris Gaynor, Jesse Poland, Allan Fritz, Reka Howard, Sarah Battenfield,* and Jose Crossa*

Abstract

Wheat (*Triticum aestivum* L.) breeding programs test experimental lines in multiple locations over multiple years to get an accurate assessment of grain yield and yield stability. Selections in early generations of the breeding pipeline are based on information from only one or few locations and thus materials are advanced with little knowledge of the genotype × environment interaction (G × E) effects. Later, large trials are conducted in several locations to assess the performance of more advanced lines across environments. Genomic selection (GS) models that include G × E covariates allow us to borrow information not only from related materials, but also from historical and correlated environments to better predict performance within and across specific environments. We used reaction norm models with several cross-validation schemes to demonstrate the increased breeding efficiency of Kansas State University's hard red winter wheat breeding program. The GS reaction norm models line effect (L) + environment effect (E), L + E + genotype environment (G), and L + E + G + (G × E) effects) showed high accuracy values (>0.4) when predicting the yield performance in untested environments, sites or both. The GS model L + E + G + (G × E) presented the highest prediction ability (*r* = 0.54) when predicting yield in incomplete field trials for locations with a moderate number of lines. The difficulty of predicting future years (forward prediction) is indicated by the relatively low accuracy (*r* = 0.171) seen even when environments with 300+ lines were included.

Core Ideas

- Incorporating environmental covariates increases genomic selection accuracy.
- G × E models can impute known lines into known environments with good accuracy.
- Breeding programs may exploit genomic selection cross-validation schemes in trial designs.

BREAD WHEAT, the predominant field crop in Kansas, represents a farm gate value of approximately $2 billion dollars per year (USDA-National Agricultural Statistical Services, 2014). However, production in Kansas is highly impacted by climatic factors, especially extreme temperature and precipitation fluctuations, which result in highly variable yield and production from year to year

D. Jarquín, Dep. of Agronomy and Horticulture, Univ. of Nebraska, Lincoln, NE 68583; C. Lemes da Silva and A. Fritz, Dep. of Agronomy, Kansas State Univ., Manhattan, KS 66506; R.C. Gaynor, The Roslin Institute and Royal School of Veterinary Studies, Univ. of Edinburgh, Easter Bush, Midlothian, UK; J. Poland, Wheat Genetics Resource Center, Dep. of Plant Pathology, Kansas State Univ., Manhattan, KS 66506; R. Howard, Dep. of Statistics. Univ. of Nebraska, Lincoln, NE; S. Battenfield, AgriPro Wheat, Syngenta, Junction City, KS 66441; J. Crossa, Biometrics and Statistics Unit, CIMMYT, El Batan, Mexico. Received 21 Dec. 2016. Accepted 1 Mar. 2017. *Corresponding authors (sarah.battenfield@syngenta.com, j.crossa@cgiar.org). Assigned to Associate Editor Roberto Tuberosa.

(Holman et al., 2011). The climatic stresses inflicted on the wheat crop may also promote various biotic stressors in different regions over different years (e.g., rust diseases, *Fusarium* head blight, and increased aphid pressure favoring the barley yellow dwarf virus). Thus environment and G × E interactions strongly impact realized yield in this region annually. Consequently, all these environmental influences significantly affect wheat breeding programs' selection for yield improvement per se.

In a breeding program, experimental lines must be tested over several years and locations to determine their expected yield performance across a wide area of adaptation (multienvironment trials). The limited amount of seed and the large number of lines to be tested in early yield trials, along with the high cost of multienvironment trials, lead to the selection of candidate lines on the basis of unbalanced or augmented experimental designs in different environments (site–year combinations). Thus candidate lines that are promoted from early generation testing on the basis of single location's or year's evaluation may not be stable in many environments and many materials that are discarded could have superior performance in other niche environments. Breeding programs have long been interested in increasing the accuracy of selection in early generation trials in which many entries have limited environmental representation of the full range of target environments.

Genomic selection was proposed as a way to use information from other related materials to predict the performance of individuals that have not yet been observed, saving time, land, or costs compared with phenotyping. It was originally suggested that as genomic marker density increased, it would be possible to estimate the variance attributed to all loci and predict line performance, as was first demonstrated by Meuwissen et al. (2001). Rapid developments in low-cost, dense genome-wide genotyping (Poland and Rife, 2012) have made it more feasible to use prediction models to reduce the amount of materials screened by breeding programs. Genomic selection uses genome-wide marker data and phenotypic information to estimate genetic breeding values from which superior candidates can be selected on the exclusive basis of genotypic data before phenotyping. To implement GS, two sets of lines are required; the first is a set of materials that have been genotyped and phenotyped, referred to as the training population (TRN). The TRN is used to calibrate the GS models and predict the genetic breeding values or phenotypic values of nonphenotyped candidates (Bassi et al., 2016). A second set of individuals made up of materials that have been genotyped but not phenotyped is the testing population (TST). The TRN and TST can be updated every year as new materials are genotyped and phenotyped by the breeding program with the aim of increasing the accuracy of its predictions over time.

Genomic selection models involving genetic markers or pedigree relationships have been used by several breeding programs for multiple traits in wheat (de los Campos et al., 2009, 2010; Crossa et al., 2010, 2011; Battenfield et al., 2016). However, fewer studies have focused on implementing GS including G × E and the unbalanced designs used in breeding programs (Burgueño et al., 2012; Lado et al., 2016). Therefore, identifying effective GS models that include G × E and other covariates (e.g., environmental covariables) could optimize resource allocation and boost genetic yield gains without significantly increasing costs.

Building on genetic-based models, additional models have been tested that include information borrowed from environments of interest to increase the models' predictive ability. Examples of these more complex GS models include models that use covariates from high-throughput phenotyping (Rutkoski et al., 2016) and environmental relationships (Jarquín et al., 2014; Heslot et al., 2014; Lado et al., 2016), and also incorporate pedigree × environment interactions (Pérez-Rodríguez et al., 2015; Velu et al., 2016) or crop models (Technow et al., 2016) to better predict line performance within specific environments. These studies indicated substantial increases in genomic prediction accuracy when the model includes G × E with the addition of environmental information as well as pedigree covariates.

Cross-validation schemes are used in genomic prediction studies to estimate accuracy when predicting different traits and environments (Burgueño et al., 2012; de los Campos et al., 2009, 2010; Crossa et al., 2010, 2011), and to mimic real situations breeders face when they have to predict lines in environments, sites, and years that have not been observed in the field. Most studies that incorporate G × E into genomic prediction use two basic random cross-validation schemes (Burgueño et al., 2012) to predict: (i) the performance of lines that have not been evaluated in any of the observed environments (CV1) and (ii) the performance of lines that have been evaluated in some environments but not in others (CV2). Another prediction problem that does not involve random cross-validation is predicting an environment (i.e., a site–year combination) that was not included in the usual set of testing environments in the evaluation system (leave-one-environment-out). In two recent studies, Jarquin et al. (2016) and Saint-Pierre et al. (2016) discussed the prediction of new sites not previously included in the usual testing sites (i.e., prediction of untested sites). However, other cross-validation schemes might be useful for testing other prediction problems (when data are available). For example, a prediction problem called forward prediction uses previous years to predict the next year. Other prediction problems of interest might be the prediction of sites that were included in different years or even cases where the prediction included sites and lines that were never used in the evaluation system.

In this study, we evalauted the genomic prediction accuracy for the grain yield of wheat lines that have been evaluated in the Kansas State University (KSU) hard red winter wheat breeding program for different sites and years. The main objective was to obtain prediction accuracy of cross-validation schemes that would answer questions related to several genomic prediction problems. We

used different sizes of the TRN sets for examining the use of resources more efficiently in a highly unbalanced and heterogeneous breeding program for predicting crop performance on a trial (environments) basis. We studied several genomic prediction problems such as genomic prediction of sites, years, and site–year combinations (environments) (leave-one- out); forward prediction (future years); prediction of newly developed lines not evaluated in any environment; and prediction of lines that were tested in some environments but not in others. We provided results for these prediction problems by using four cross-validation schemes together with six prediction models. Some of these models use only line information and environment information, others include genomic information, and still others incorporate G × E interaction effects (where E denotes either "environments" or "sites," depending on the model). The cross-validation methods consisted of random cross-validations (CV1 and CV2, respectively), as well as two more cases where no information on the environments or sites to be predicted appears in the TRN (CV0 included the prediction of tested lines in untested sites or environemnts by leaving one environment out or leaving one site out) and another case where untested lines were predicted in untested environments and or sites (CV00).

We evaluated six different prediction models [Model 1 (M1)–Model 6 (M6)] using the four different cross-validations schemes in an extensive dataset consisting of 1378 breeding lines evaluated from 2009 to 2014 at 31 different environments (site–year combinations) in the state of Kansas that are commonly used in the standard multienvironment trials of KSU's hard red winter wheat breeding program.

Materials and Methods

Genetic Materials

Yield data on 1378 wheat breeding lines were obtained from Kansas State University's Manhattan breeding program for different locations in advanced stages of yield testing from 2009 to 2014. These lines are predominantly classified as hard red winter bread wheat and are adapted to central and eastern Kansas. Breeding materials represented preliminary, advanced, and intrastate yield testing, ranging from $F_{5:7}$ to $F_{5:10}$, with a few released cultivars from the region, serving as check varieties. The materials were tested in an unbalanced manner, as the data originated from historical breeding selection trials, where only best materials would be advanced to further trials. Thus a limited number of selected individuals were tested across environments (site–year combination) in this dataset. Table 1 shows a total of 31 environments, with each of them including a different number of lines evaluated (sample size). For example, in 26 environments, more than 100 lines were tested (100+); in 17 environments, more than 200 lines were tested (200+); and in 14 environments, more than 300 lines were included (300+). The results are

also presented according to the number of lines within the environment (100, 200, or 300+, respectively).

Check varieties are commonly used to allow for comparisons between or among individuals in different environments in plant breeding. Therefore, most environments share some entries in common across years, and some environments may share many more entries because of the breeding pipeline. Table 2 (31 × 31) shows the number of lines in each of the 31 environments (site–year combinations) in the diagonal, and the off-diagonals demonstrate commonalities between environments. The cells of the upper off-diagonal (i.e., the section above the diagonal) has the number of lines in common between the two environments and the cells of the bottom off-diagonal has the number of lines not in shared between the two environments. Supplemental Fig. S1 depicts lines (black vertical lines) for the particular line × environment combinations that were observed in each of the environments.

Yield Trials and Experimental Design

For the advanced materials, linear mixed models were used to analyze breeding trials including the design effect, replicate, and subblock within replicated as random effect and the effect of the lines as random (nongenetic BLUPs), assuming that the lines are uncorrelated. Breeding yield trial data were analyzed in varying designs on the basis of the amount of seed and various objectives based on breeding stage. Preliminary yield trials ($F_{5:7}$) were conducted using a modification of augmented design (Federer and Raghavarao, 1975) with one replicate of each experimental line per location. In this design, whole-plot checks are planted across whole rows and columns in the field, and sub-block checks are randomly assigned within blocks. The individual yield is then adjusted using a row–column design (Lin and Poushinsky, 1986). Preliminary yield trials were also conducted at seven locations across Kansas. Materials advanced from preliminary trials were promoted to advanced yield trials ($F_{5:8}$) using two replicated α-lattice designs (Patterson and Williams, 1976). Finally, in the most advanced stage of testing, lines in the the Kansas intrastate nursery ($F_{5:9}$ or beyond) were planted in three replicated, randomized complete block designs (Cochran and Cox, 1957) at approximately 17 locations per year in Kansas (Table 1). In all these generations, yield was measured in plots measuring 1.5 by 4.5 m. All individual yield trials were analyzed with Agrobase Generation II software [Agrobase Generation II 2014, Agronomix, Winnipeg, MB, Canada; https://www.agronomix.com/ (accessed 8 May 2017)] according to their respective experimental designs, resulting in site–year BLUPs for lines tested within each site × trial × year combination.

DNA Extraction and Genotyping

Genotyping of all advanced materials in the KSU wheat breeding program began in 2011. Initially, it was done retrospectively using historical stored seed that was also genotyped from yield trials as long ago as 2005. Genotyping of the preliminary yield trial materials was also conducted

Table 1. Distribution, sample size, means, and yield data of 1378 wheat breeding lines from the Kansas State University Manhattan breeding program at the advanced stages of yield testing, 2009–2014.

Site	Harvest year	Environment code	Environment number	Sample size (n lines tested)	Mean	SD	Min	Max
						t ha^{-1}		
Belleville	2009	Belleville09	1	56	3.29	0.36	2.39	4.14
Gypsum	2009	Gypsum09	2	55	4.36	0.36	3.59	5.07
Hutchinson	2009	Hutchinson09	3	56	3.49	0.46	1.98	4.26
Barber	2010	Barber10	4	179	2.38	0.41	1.42	3.62
Belleville	2010	Belleville10	5	179	2.73	0.58	0.90	4.17
Gypsum	2010	Gypsum10	6	179	2.80	0.43	0.88	4.21
Lane	2010	Lane10	7	179	3.80	0.57	2.10	5.54
Summer	2010	Sumner10	8	179	3.00	0.36	1.76	3.87
Gypsum	2011	Gypsum11	9	125	2.82	0.43	1.18	3.74
Hutchinson	2011	Hutchinson11	10	125	2.04	0.54	0.66	3.35
Summer	2011	Sumner11	11	104	1.67	0.32	0.83	2.63
Belleville	2012	Belleville12	12	340	3.42	0.64	1.81	4.98
Gypsum	2012	Gypsum12	13	350	2.52	0.41	1.34	3.68
Hutchinson	2012	Hutchinson12	14	349	2.87	0.61	0.03	4.43
Manhattan	2012	Manhattan12	15	349	3.41	0.48	1.69	4.78
McPherson	2012	McPherson12	16	350	3.08	0.55	0.55	4.31
Barber	2013	Barber13	17	60	2.26	0.19	1.84	2.70
Belleville	2013	Belleville13	18	287	2.74	0.41	1.76	3.96
Ellsworth	2013	Ellsworth13	19	60	2.65	0.31	1.84	3.37
Gypsum	2013	Gypsum13	20	442	3.49	0.60	1.62	5.57
Hutchinson	2013	Hutchinson13	21	441	3.29	0.66	1.55	5.26
Lane	2013	Lane13	22	157	1.75	0.40	0.65	2.79
Manhattan	2013	Manhattan13	23	442	3.20	0.34	1.91	4.28
McPherson	2013	McPherson13	24	442	2.96	0.47	1.51	4.26
Summer	2013	Sumner13	25	442	2.24	0.47	0.03	3.41
Belleville	2014	Belleville14	26	377	1.48	0.35	0.09	2.41
Lane	2014	Lane14	27	288	2.37	0.94	0.00	4.53
Gypsum	2014	Gypsum14	28	256	2.52	0.55	0.66	3.99
Manhattan	2014	Manhattan14	29	378	3.77	0.43	2.37	4.88
McPherson	2014	McPherson14	30	378	1.78	0.41	0.33	3.02
Summer	2014	Sumner14	31	378	1.95	0.35	0.77	2.68

annually at the line derivation stage ($F_{5:6}$) from that point forward. DNA was extracted from bulked leaf tissue using the BioSprint 96 DNA Plant Kit (Qiagen) with the BioSprint 96 Workstation (Qiagen). DNA was quantified and normalized, digested with two restriction enzymes (*PstI* and *MspI*), ligated with barcoded adapters, and then sequenced following the genotyping-by-sequencing two-enzyme protocol described in Poland et al. (2012).

Sequence reads were analyzed using the TASSEL version 4 de novo pipeline to identify single nucleotide polymorphisms (Bradbury et al., 2007). Single nucleotide polymorphisms were converted to numeric allele classes (1, 0, −1 for homozygous major, heterozygous, and homozygous minor alleles, respectively) using the R package 'GSwGBS' (Gaynor, 2015). Markers were filtered to exclude those with a minor allele frequency smaller than 0.01 and more than 20% missing values across genotypes. In addition, lines with more than 50% missing molecular marker values were discarded. Mean marker imputation was conducted using the R package 'rrBLUP' (Endelman, 2011), where missing markers were simply imputed as the mean value among all lines for that marker.

Statistical Yield Analysis

Best linear unbiased predictors were obtained via Agrobase Generation II (Agronomix) for each site × year combination to account for the experimental design effects. Environments were filtered for low numbers of entries and excessive coefficient of variation values (CV > 12%) before assessing GS predictions. Yield BLUPs for each environment were used in the GS models.

Statistical Prediction Models

A series of models was used in this study for performing predictions: two models that included only phenotypic information on the training sets, two GS models that included main effects of the markers, and two models that added the G × E component using the reaction norm model as described by Jarquín et al. (2014). These models used the random effect approach for all components.

Table 2. Number of wheat lines tested in 31 environments (shown in bold in the diagonal), number of common lines in a pair of environments (shown above the diagonal), and number of lines not in common in a pair of environments (shown below the diagonal).

Env.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	**56**	55	56	5	5	5	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	55	**55**	1	4	4	4	4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	**56**	5	5	5	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	225	226	225	**179**	179	179	179	179	20	20	1	9	9	9	9	9	2	0	2	2	2	2	2	2	2	1	1	1	1	1	1
5	225	226	225	0	**179**	179	179	179	20	20	1	9	9	9	9	9	2	0	2	2	2	2	2	2	2	1	1	1	1	1	1
6	225	226	225	0	0	**179**	179	179	20	20	1	9	9	9	9	9	2	0	2	2	2	2	2	2	2	1	1	1	1	1	1
7	225	226	225	0	0	0	**179**	179	20	20	1	9	9	9	9	9	2	0	2	2	2	2	2	2	2	1	1	1	1	1	1
8	225	226	225	0	0	0	0	**179**	20	20	1	9	9	9	9	9	2	0	2	2	2	2	2	2	2	1	1	1	1	1	1
9	181	180	181	20	20	20	20	20	**125**	125	104	9	19	19	19	19	6	0	6	6	6	0	6	6	6	1	1	1	1	1	1
10	181	180	181	20	20	20	20	20	0	**125**	104	9	19	19	19	19	6	0	6	6	6	0	6	6	6	1	1	1	1	1	1
11	160	159	160	281	281	281	281	281	21	21	**104**	1	11	11	11	11	4	0	4	4	4	4	4	4	4	0	0	0	0	0	0
12	396	395	396	501	501	501	501	501	447	447	442	**340**	340	339	339	340	56	33	56	56	56	33	56	56	56	14	14	14	14	14	14
13	406	405	406	511	511	511	511	511	437	437	432	10	**350**	349	349	350	60	33	60	60	60	33	60	60	60	14	14	14	14	14	14
14	405	404	405	510	510	510	510	510	436	436	431	11	1	**349**	349	349	59	33	59	59	59	32	59	59	59	14	14	14	14	14	14
15	405	404	405	510	510	510	510	510	436	436	431	11	1	2	**349**	349	59	32	59	59	59	32	59	59	59	14	14	14	14	14	14
16	406	405	406	511	511	511	511	511	437	437	432	10	1	1	1	**350**	60	33	60	60	60	33	60	60	60	14	14	14	14	14	14
17	116	115	116	235	235	235	235	235	173	173	156	288	290	291	291	290	**60**	60	33	60	60	33	60	60	60	13	13	13	13	13	13
18	343	342	343	466	466	466	466	466	412	412	391	561	571	570	572	571	281	**287**	33	287	286	157	287	287	287	31	31	31	31	31	31
19	116	115	116	235	235	235	235	235	173	173	156	288	290	291	291	290	0	281	**60**	60	60	33	60	60	60	13	13	13	13	13	13
20	498	497	498	617	617	617	617	617	555	555	538	670	672	673	673	672	382	155	382	**442**	441	157	442	442	442	47	47	47	47	47	47
21	497	496	497	616	616	616	616	616	554	554	537	669	671	672	672	671	381	156	381	441	**441**	157	441	441	441	47	47	47	47	47	47
22	213	212	213	336	336	336	336	336	282	282	261	431	441	440	442	441	151	130	151	157	157	**157**	157	157	157	19	19	19	19	19	19
23	498	497	498	617	617	617	617	617	555	555	538	670	672	673	673	672	382	155	382	442	1	285	**442**	442	442	47	47	47	47	47	47
24	498	497	498	617	617	617	617	617	555	555	538	670	672	673	673	672	382	155	382	442	1	285	0	**442**	442	47	47	47	47	47	47
25	498	497	498	617	617	617	617	617	555	555	538	670	672	673	673	672	382	155	382	442	1	285	0	0	**442**	47	47	47	47	47	47
26	433	432	433	554	554	554	554	554	500	500	481	689	699	698	698	699	411	602	411	725	724	496	725	725	725	**377**	255	287	377	377	377
27	312	311	312	433	433	433	433	433	379	379	360	568	578	577	577	578	290	481	290	604	603	375	604	604	604	123	**256**	166	256	256	256
28	344	343	344	465	465	465	465	465	411	411	392	600	610	609	609	610	322	513	322	636	635	407	636	636	636	91	212	**288**	288	288	288
29	434	433	434	555	555	555	555	555	501	501	482	690	700	699	699	700	412	603	412	726	725	497	726	726	726	1	122	90	**378**	378	378
30	434	433	434	555	555	555	555	555	501	501	482	690	700	699	699	700	412	603	412	726	725	497	726	726	726	1	122	90	378	**378**	378
31	434	433	434	555	555	555	555	555	501	501	482	690	700	699	699	700	412	603	412	726	725	497	726	726	726	1	122	90	378	378	**378**

Baseline Model

The response of the jth genotype in the ith environment (y_{ij}) could be described as $y_{ij} = \mu + E_i + L_j + EL_{ij} + e_{ij}$, where μ is the overall mean; E_i ($i = 1,\ldots,I$) denotes the random effect of the ith environment and assuming $E_i \overset{iid}{\sim} N(0,\sigma_E^2)$ with $N(.,.)$ denoting a normal density, where iid stands for independent and identically distributed observations, and σ_E^2 represents the variance component of the environments; L_j represents the random effect of the jth line ($j = 1,\ldots, J$) such that $L_j \overset{iid}{\sim} N(0,\sigma_L^2)$, where σ_L^2 is the variance of the line; EL_{ij} describes the random interaction effect between the ith environment and the jth line with $EL_{ij} \overset{iid}{\sim} N(0,\sigma_{EL}^2)$; and σ_{EL}^2 as the line × environment interaction variance; e_{ij} is the random error term where $e_{ij} \overset{iid}{\sim} N(0,\sigma_e^2)$ with σ_e^2 as the residual variance. The main limitation of this model is that it does not allow one to borrow information between lines because assumptions of being independent and identically distributed were made for these. All the implemented models can be derived from model (1) either by subtracting terms and/or adding further assumptions.

Environment + Line Model (Model 1)

This model was obtained from the baseline model after retaining the first three components plus the error term and their corresponding assumptions of these random effects as

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

with environments as the site–year combinations as described before. A graphic representation of this component can be found in Fig. 1A, where the environments are considered to be independent and identically distributed.

Site + Line Model (Model 2)

In Model 1 (M1), environments were considered to be site–year combinations; however, since most of the sites were observed in several years, we attempted to recover information from sites in different years by considering the site effect instead the environment effect. Under this consideration, the year effect was treated as negligible to allow borrowing information between environments (site–year combinations) coming from same site but observed in different years (Fig. 1B). Thus the term y_{kj} denotes the response of the jth line observed in the kth site ($S_k, k = 1, 2, \ldots, K$) and can be described as:

$$y_{kj} = \mu + S_k + L_j + e_{kj} \qquad [2]$$

with $S_k \overset{iid}{\sim} N(0,\sigma_S^2)$ and σ_S^2 being the variance of sites.

Environment + Genomic Main Effects Model (Model 3)

Markers are introduced in M1 by a genomic representation of the random effect of line (L_j) with its genomic surrogate, such that $g_j = \sum_{m=1}^{p} x_{jm} b_m$, a linear combination between p markers and their corresponding marker effects; marker values were coded as before. Marker effects are considered to be random effects, such that $b_m \overset{iid}{\sim} N(0,\sigma_b^2)$ for ($m = 1,\ldots,p$) and σ_b^2 is the marker effect variance. Using the properties of the multivariate normal distribution, the vector $\mathbf{g} = (g_1,\ldots,g_J)'$ containing the genomic values of all the lines follows a multivariate normal density with a zero mean and a covariance matrix

$$Cov(\mathbf{g}) = \mathbf{G}\sigma_g^2, \text{ where } \mathbf{G} = \frac{XX`}{p} \text{ is the genomic relationship}$$

matrix (VanRaden, 2008), $\mathbf{X}$ is the centered and standardized genotype matrix, and σ_g^2 is the genomic variance equivalent to p times the variance of the markers (i.e., $\sigma_g^2 = p \times \sigma_b^2$).

Therefore, after adding g_j, M1 becomes Model 3 (M3), which includes environment and line effects plus genomic main effects (this last component is also known as the GBLUP model) and can be expressed as:

$$y_{ij} = \mu + E_i + L_j + g_j + e_{ij} \qquad [3]$$

where $\mathbf{g} = \{g_j\} \sim N(0,\mathbf{G}\sigma_g^2)$ and the other terms are as defined previously. In this case, the line component remains in the model to account for imperfect information and model misspecification caused by imperfect linkage disequilibrium.

Site + Genomic Main Effects Model (Model 4)

This is similar to M3 but the response variable and the environmental component are replaced with the response and the site random effects terms shown in Model 2 (M2). Thus Model 4 (M4) is:

$$y_{kj} = \mu + S_k + L_j + g_j + e_{kj} \qquad [4]$$

Interaction Model for Environments (Model 5)

This random effects model (Model 5, M5) accounts not only for the genomic main effects, as was the case in M3, but also attempts to include G × E interactions by including covariance structures, as shown by Jarquín et al. (2014). In this case, EL_{ij}, the interaction component of the baseline model, can be replaced by Eg_{ij}, a component that conceptually represents the interaction between each molecular marker and each environment (further details given below). Under these assumptions, M3 can be extended as follows:
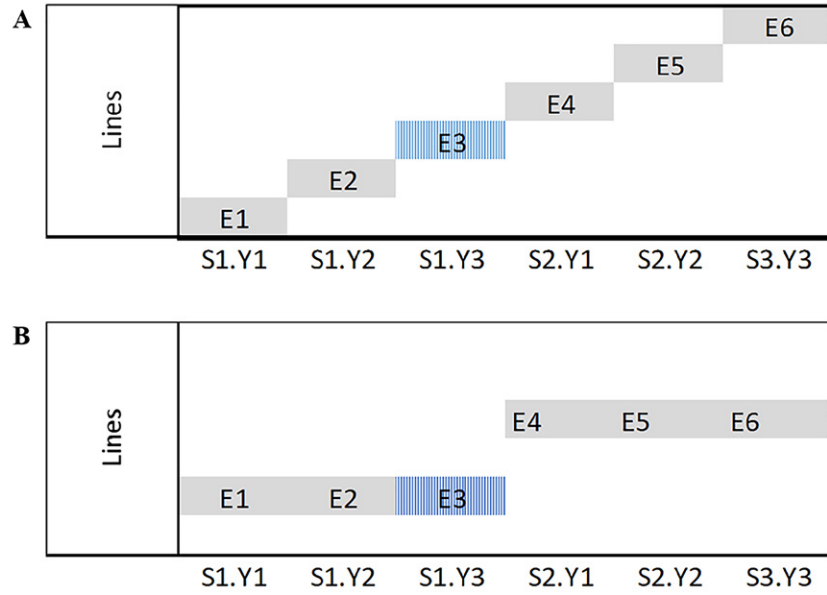
Fig. 1. Graphic representation of experimental trials at three sites (S) tested in two different years (Y) under different assumptions about the year effect. (A) Site–year combinations (S1.Y1, S1.Y2, S1.Y3, S2.Y2, S2.Y3, S2.Y3) are treated as independent outcomes for the same site. (B) When years present a null effect over certain sites, the environmental conditions observed across years could be considered as equivalent, allowing information to be borrowed within sites predicting, for example, environment E3 (blue shading).

$$y_{ij} = \mu + E_i + L_j + g_j + Eg_{ij} + e_{ij} \qquad [5]$$

with $Eg = \{Eg_{ij}\} \sim N\left(0, \left(Z_g G Z_g'\right) \circ \left(Z_E Z_E'\right) \sigma_{Eg}^2\right)$, where Z_g and Z_E are the incidence matrices for lines and environments, respectively; σ_{Eg}^2 is the variance component of the Eg_{ij} interaction component; and $\circ$ stands for the Hadamard or Schur product between two matrices. It denotes the element-to-element product between two matrices.

Interaction Model for Sites (Model 6)

Like the last model, Model 6 (M6) replaces the interactions term between each marker and each environment by the interactions between each marker and each site. Thus M6 is:

$$y_{kj} = \mu + S_k + L_j + g_j + Sg_{kj} + e_{kj} \qquad [6]$$

where $Sg = \{Sg_{kj}\} \sim N\left(0, \left(Z_g G Z_g'\right) \circ \left(Z_S Z_S'\right) \sigma_{Sg}^2\right)$, Z_S is the incidence matrix for sites, σ_{Sg}^2 is the variance component of the Sg_{kj} interaction component, and the other terms are as previously defined.

Description of Prediction Problems Using Various Cross-Validation Strategies

Model predictive ability was assessed on a trial basis (Jarquin et al., 2016). Predictive ability was computed as the correlation between observed and predicted values within the same environment (location–year), no matter how the TST sets were comprised under the different cross-validation schemes. The main objective was to study prediction problems of interest to the KSU hard

Table 3. Strategies (prediction problem by cross-validation scheme combinations) used for mimicking real scenarios that wheat breeders may face in the field.

Prediction problem	Cross-validation schemes			
	CV2†	CV1	CV0	CV00
Prediction of lines in incomplete trials (sparse testing)	1	–	–	–
Predicting new lines (not being observed in any environment yet)	–	2	–	–
Predicting lines in untested environments	–	–	3	4
Predicting lines in untested sites	–	–	5	6
Predicting lines in untested years	–	–	7	8

† CV2, predicting tested lines in tested environments; CV1, predicting untested lines in tested environments; CV0, predicting tested lines in untested environments; CV00, predicting untested lines in untested environments.

red winter wheat breeding program. The main prediction problems consisted of studying the prediction accuracy of environments, years, and sites, as well as the prediction accuracy of newly developed lines (i.e., lines that were never evaluated in any environment, CV1) and of lines that were evaluated in some environments but not in others (CV2). Prediction of environments, years, and sites was achieved by using the cross-validations CV0, which consisted of predicting environments that were never previously observed, and CV00, which predicted both lines and environments never previously tested.

Table 3 shows the eight strategies used for specific combinations of the four cross-validation schemes (CV2, CV1, CV0, and CV00) and two levels of genotypes and environments (tested or untested) for genotypes and environments, as well as four problems to be solved: prediction of incomplete field trials, prediction of 'newly'

developed (untested) lines, prediction of unobserved sites, and prediction of unobserved years. CV2 is used to solve the problem of predicting a certain portion of tested genotypes in a certain portion of tested environments; CV1 refers to cases where certain individuals were not observed in any tested environment, whereas other lines were tested in these environments; CV0 refers to cases where some tested genotypes that were observed in some tested environments are predicted in environments that were not previously used (untested environments); and CV00 refers to cases where unobserved individuals (untested genotypes) are predicted in environments not previously used (untested environments).

Assessing Prediction Accuracy for Each Cross-Validation Strategy

Four basic cross-validation schemes were implemented to mimic real application problems that breeders may face in the field: CV1, CV2, CV0, and CV00 (Fig. 2A–D). Random fivefold partitions of the entire population were used for CV1 and CV2 and prediction accuracy was the average correlation between predicted and observed values of the lines within the same environment for 50 random fivefold partitions (see the detailed description below). For CV0 and CV00, the correlations between observed and predicted values are computed for (i) leave-one-environment-out, (ii) leave-one-site-out, and for (iii) forward prediction of one future year using all the previous years. In all cases, correlations were computed only between predicted and observed values within the same environment.

The random CV2 strategy mimics the problem of predicting incomplete field trials where some experimental lines have been evaluated in some environments but not in others. For example, Fig. 2A shows that the aim is to predict those lines that appear in gray color in the top left panel, having observed the performance of these lines in other environments (bottom left panel), other lines in these same environments (top right panel), and other lines in other environments (bottom right panel). In this case, breeders may wish to estimate the performance of the unobserved lines in a target environment using phenotypic information on these lines and others observed in other environments together with information on the performance of other lines tested in the current environment. Knowing how the target environment affected other lines may help the prediction process, since the unobserved genotypes could be affected in a similar way. Site–year yield BLUPs were assigned randomly to fivefold partitions and each partition was predicted using the remaining four in a proportion of 20:80 (20% of the phenotypic records were predicted using the remaining 80%). This process was repeated 50 times, then the correlations between the predicted and observed values within the same environment were computed for all environments; means (and SD) across replicates for each environment were computed.

As previously described, the random CV1 scheme mimics the problem of predicting the crop performance of new experimental lines that have not yet been observed in any of the tested environments. Fig. 2B depicts the prediction of the lines in gray color in those environments in the top left panel; no information from these lines was observed in the other environments, since these were masked as missing (bottom left panel). The genetic similarities between lines in TRN and TST play a major role in the models' predictive ability. Lines were assigned randomly to fivefold partitions (with 20% of the lines being predicted using the phenotypic information of the remaining 80% of the genotypes) such that all the phenotypes from the same line appeared in same fold. Each partition was predicted using the remaining four, one at a time. The correlations between predicted and observed values within same environment were computed for all 50 fivefold random partitions.

The CV0 strategy aims to predict the crop performance of th experimental lines in new (untested) environments. Fig. 2C shows that the aim is to predict those lines in gray color observed in those environments grouped in the top left panel, although there are no data from any line tested before in these same environments (top right panel). This method is used when there are no phenotypic records of any line being observed in the target environment (Jarquin et al., 2016; Saint-Pierre et al., 2016). The success of this strategy's predictive ability will depend partly on whether the environmental conditions in the unobserved environment fall within the range of conditions in the environments making up TST as well as the performance and number of phenotypic records from the same genotypes in the TST observed in other environments. Here, the predictions are made by leaving one environment out and using the remaining environments as the TST. The correlations between the predicted and observed values for each environment were then computed. The procedure does not involve random partitioning and thus it was implemented only once.

The CV00 strategy is similar to the previous scheme; however, the lines to be predicted were never observed before (lines are untested). This is shown in Fig. 2D with the same objective as before: to predict those lines in gray color observed in the environments that appear in the top left panel where the only source of information is a different set of lines observed in other environments (bottom right panel). This model may apply to the situation where most of the materials (except the checks) are new or have not been observed before in any field and their performance for the next year needs to be estimated. The accuracy of this strategy's predictive ability will depend on the genetic similarities shown between lines in the TST and those in the TRN sets.

Models M1 to M6 represented in Eq. [1–6] were implemented following the eight strategies described in Table 3. In models that included the site component (S, site only) instead of the environment component (E, site–year combination) the correlation between the predicted and observed values were computed within environments and not within sites. If there was valuable information on sites, this could be recovered from predicting
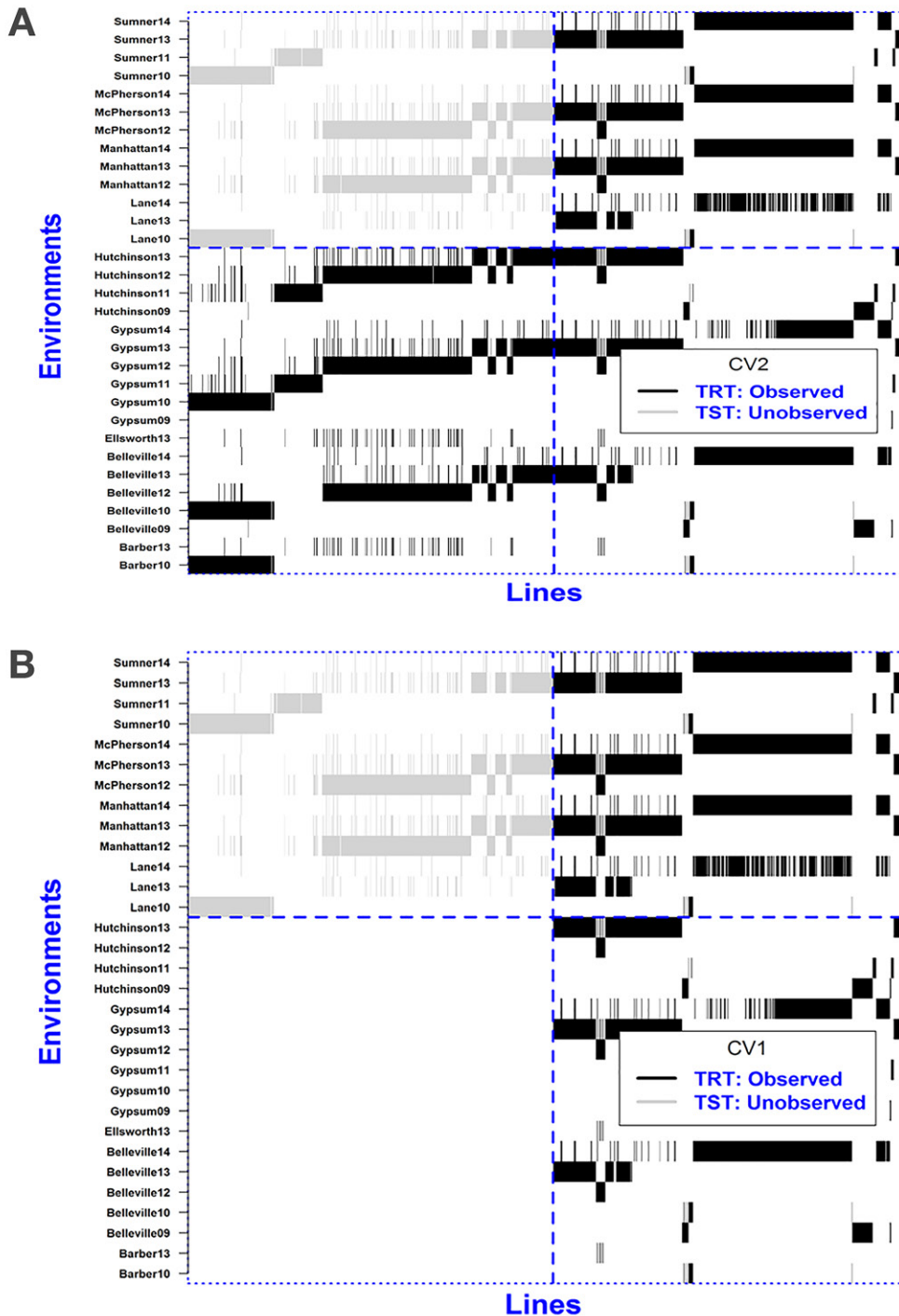
Fig. 2. (continued on next page) Graphic representation of the four cross-validation (CV) schemes used in this study (CV1, CV2, CV0, and CV00) to assess predictive ability of models on a trial basis (environment). Lines appear along the *x* axis and environments are indicated on the *y* axis. In this example, for all cases, the testing population (TST) is the block of lines that appear in gray in the environments at the top. (A) CV2 (incomplete field trials) attempt to predict the TST using information from other lines observed in all environments plus the same lines from TST but observed in the remaining environments. Black vertical lines represent combinations of lines and environments observed in the field. Gray lines denote the prediction of a portion of tested lines that were never observed in some tested environments. (B) CV1 (newly developed lines) discards the lines from TST observed in other environments from the training set. Black lines represent combinations of lines and environments observed in the field. Gray lines denote the prediction of a portion of tested lines that were never observed in some tested environments. (C) CV0 (observed lines in unobserved environment) discards information from all lines observed in the environments to be predicted from the training set. Black lines represent combinations of lines and environments observed in the field. Gray lines represents the prediction of a portion of tested lines in untested environments. (D) CV00 (unobserved lines in unobserved environments) discards not only information on all lines observed in the environments to be predicted but also lines from TST observed in other environments. Black lines represent combinations of lines and environments observed in the field. Gray lines denote the prediction of untested lines in untested environments.
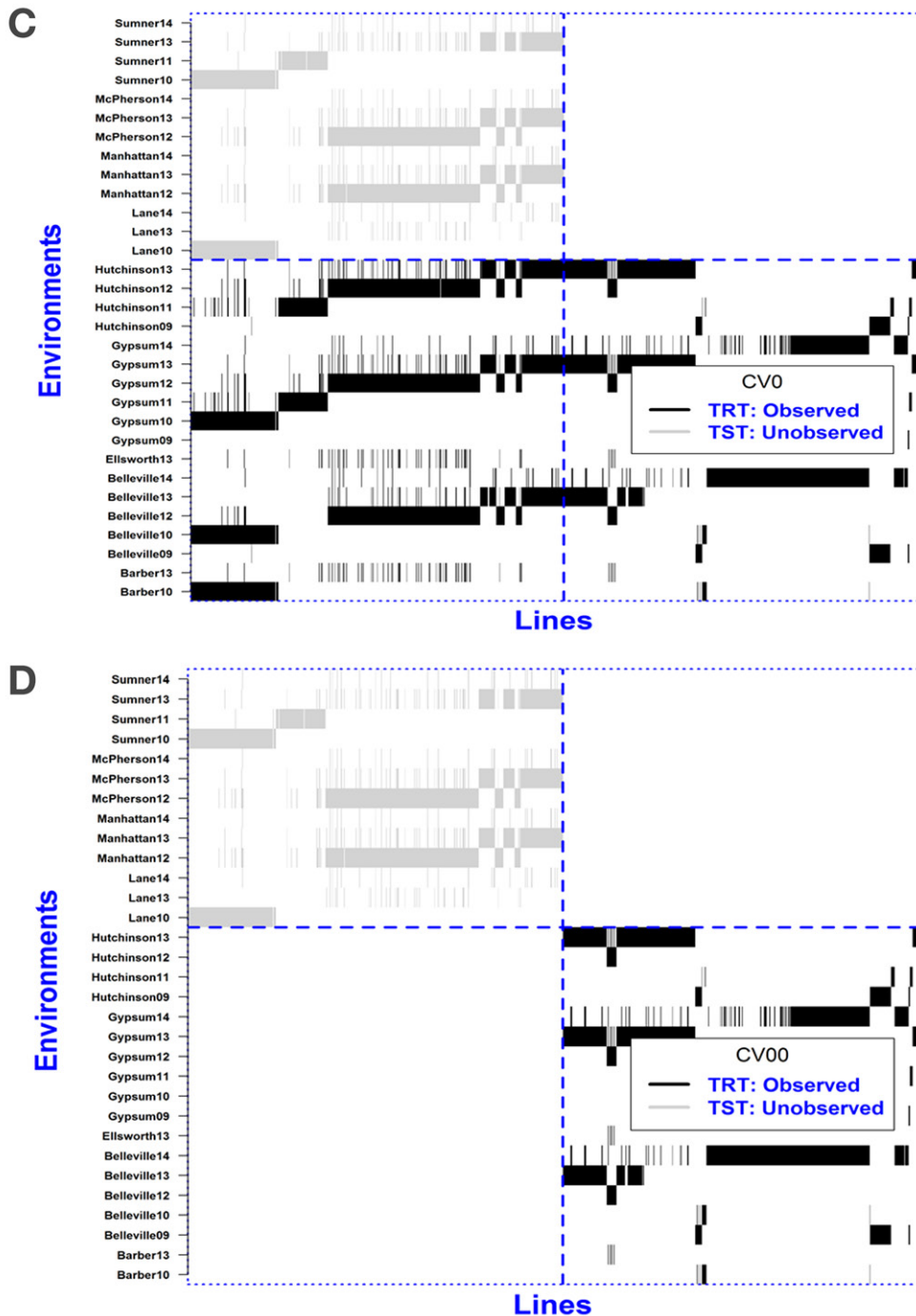
Fig. 2. Continued.

the performance for a year at the same site using another year. For predicting new environments via the CV0 and CV00 strategies, predictions were made by leaving one environment out at the time (Strategies 3 and 4 in Table 3) and not leaving sites out, unless explicity stated. For example, using M1, M3, and M5 to predict environment E3, Fig. 1A shows the portion of data that should be set as missing, which corresponds to S1.Y3 (Site 1 × Year 3 combination). On the other hand, with M2, M4, and M6, the data portion of phenotypes to be set as missing is the same as before but we assume that the other environments (S1.Y1 and S1.Y2) are a part of a more general

environment where all partitions (three in this explanatory example) shared similar environmental conditions without considering time constraint or issues (Fig. 1B).

Two more studies were proposed to evaluate the potential of GS methods for making predictions in real scenarios. The first one attempts to predict all environments coming from the same site. The aim was to predict the crop performance of lines in new sites but not including the years when the experiments were established. These are strategies 5 and 6 in Table 3, which use the CV0 and CV00 cross-validation strategies, respectively. A graphic representation of this scheme considering Environments

Table 4. Estimated variance components (for six models (M1–M6) and percentage of within-environment or within-site variance accounted for by each random effect of the corresponding model using the full data for the grain yield of wheat lines.

Models	Estimated variance component							Percentage of within-environment or within-site variance				
	E†	S	L	G	G × E	G × S	Res.	L	G	G × E	G × S	Res.
M1 = L + E	124.66	–	17.84	–	–	–	46.96	27.53	–	–	–	72.47
M3 = L + E + G	119.65	–	11.25	7.08	–	–	46.95	17.23	10.84	–	–	71.92
M5 = L + E + G + (G × E)	115.25	–	11.27	4.59	14.55	–	34.02	17.49	7.12	22.59	–	52.80
M2 = L + S	–	63.14	34.93	–	–	–	105.61	24.85	–	–	–	75.15
M4 = L + S + G	–	52.72	12.73	22.72	–	–	105.17	9.05	16.16	–	–	74.79
M6 = L + S + G + (G × S)	–	45.15	15.30	16.14	–	27.12	78.42	11.17	11.78	–	19.80	57.25

† E, environment; L, line; S, site; G, genomic (marker); G × E, genomic × environment; G × S, genomic × site; Res., residual

1 to 3 [environments observed in the same site (Site 1) but in different years (Years 1, 2, and 3)] to predict all environments in Site 2 (S2.Y1, S2.Y2, and S2.Y3) appears in Fig. 1B. As before, correlations were computed on a trial basis (i.e., correlations between predicted and observed performance were computed only within environments).

The second study involves so-called forward prediction (Strategies 7 and 8 in Table 3). This consists of predicting the environments in any given year (TST population) using all the environments from the previous years (TRN population). For example, data from 2009 are the TRN population used for predicting the year 2010 (the TST population); data from 2009 and 2010 combined are the TRN population used to predict the year 2011 (the TST population). This forward cross-validation prediction scheme provides a more realistic idea of GS potential to deal with unobserved years. As in the other cases, correlations for each environment were computed after poststratifying predicted values.

Results and Discussion

Phenotypic Results

The average number of phenotypic records for the 31 environments was 258; the sample size of the environments with the smallest and largest number of individuals was 55 and 442, respectively. Of the 31 environments, 26, 17, and 14 had more than 100, 200, and 300 phenotypic records; thus only five environments contained fewer than 100 records (only 57 on average). The mean grain yield for all environments was 2.77 t ha^{-1}, whereas the maximum and minimum values were 1.48 and 4.36 t ha^{-1}, and the standard deviation was 0.46 t ha^{-1}.

Large portions of variance were attributed to the environment or site, depending on the model of data analysis (Table 4). As expected, the variability explained by the environment component in M1, M3, and M5 was larger than the variance explained by the other components. For M2, M4, and M6, the variability explained by the site component was close to 50% of the variance explained by the environments. In this case, the remaining variability was left in the residual component. Half of the environment variance component is accounted for by the site and the remaining by the line, genomic, and

residual variances. Since predictive ability is assessed on a trial basis, it is important to compare the percentage of within-environment (site–year) or within-site (location) variance shown for the interaction components. In this case, G × E and genome × site effects explained a larger percentage of variability than the linear effects of the lines and the markers forM5 and M6.

Genomic Selection Performance

Results from the different cross-validation strategies and models for different groups of environments with different sample sizes appear in Table 5, Table 6, Table 7, Table 8 and Fig. 3. In all cases, four different groups of sample sizes were considered. "All" displays the mean across all environments despite the number of phenotypic records per environment. The other cases show the mean across environments with more than 100, 200, and 300 lines included in each of them.

For the strategy of predicting incomplete field trials using the CV2 scheme, Table 5 shows an average predictive ability of 0.381 considering only lines and environments. A small increase was achieved when markers were included using M3; however, M5, which accounts for the G × E interaction, improved by about 16%, shown by Fig. 3. In all these cases, the predictive ability improved slightly in environments with larger sample sizes. With models M2, M4 and M6 improvements in predictive ability were observed in environments with larger samples; however, the results were never as good as those achieved by their counterparts (M1, M3, and M5).

The strategy of predicting newly developed lines using the CV1 strategy showed (Table 5) that M5 produced better results than all the other models. Model 5 improved predictive ability between 52 and 82% for the different sample sizes compared with the baseline model, M3, which did not include the G × E component. As before, models based on sites (i.e., M2, M4, and M6) showed similar patterns; however, they never outperformed models that considered environments as site–year combinations. This may indicate a strong effect of the year factor on sites.

Predicting unobserved environments using the CV0 strategy with models M1, M3, and M5 produced results that improved the prediction accuracy of sites with larger sample sizes (Table 6, Fig. 3). However, no advantages of

Table 5. Mean correlation and the SD (in parentheses) between predictive and observed values for wheat trials from two random cross-validation for groups of environments with different sample sizes and six prediction models (M1–M6).

Purpose	CV	Sample size†	M1‡	M3	M5	M2	M4	M6
Predicting lines in incomplete trials	CV2§	All	0.381 (0.018)	0.386 (0.018)	0.445 (0.023)	0.339 (0.021)	0.312 (0.018)	0.299 (0.024)
		100+	0.397 (0.013)	0.403 (0.012)	0.459 (0.018)	0.358 (0.015)	0.343 (0.013)	0.314 (0.014)
		200+	0.465 (0.009)	0.472 (0.008)	0.540 (0.013)	0.419 (0.013)	0.408 (0.010)	0.363 (0.015)
		300+	0.465 (0.008)	0.472 (0.008)	0.540 (0.013)	0.419 (0.012)	0.408 (0.010)	0.363 (0.015)
Predicting newly developed lines	CV1	All	−0.089 (0.071)	0.143 (0.033)	0.261 (0.038)	−0.029 (0.075)	0.144 (0.021)	0.165 (0.024)
		100+	−0.077 (0.060)	0.171 (0.026)	0.286 (0.031)	−0.024 (0.064)	0.165 (0.019)	0.18 (0.021)
		200+	−0.057 (0.051)	0.242 (0.019)	0.368 (0.022)	−0.021 (0.051)	0.218 (0.015)	0.219 (0.019)
		300+	−0.057 (0.049)	0.242 (0.018)	0.368 (0.021)	−0.021 (0.049)	0.218 (0.015)	0.219 (0.019)

† All, 31 environments were included; 100+, the 26 environments where 100 or more lines were tested; 200+, the 17 environments where 200 or more lines were tested; 300+, the 14 environments where 300 or more lines were tested.

‡ L, line effect; E, environment (site–year combination) effect; G, main effect of genomic markers; G × E, genotype × environment interaction; S, site effect; G×S, genotype × site interaction; M1, L + E; M3, L + E + G; M5, L + E + G + (G × E); M2, L + S; M4, L + S + G; M6, L + S + G + (G × S)

§ CV2, prediction of incomplete field trials, where some lines were tested in some environments but not in others; CV1, prediction of newly developed lines, with lines that have not yet been tested in any field trial.

Table 6. Correlation between predictive and observed values for two cross-validation scenarios for wheat trials leaving one environment out: prediction of crop performance in unobserved environments using information on lines from other environments (CV0) and when none of these lines and environments have ever been tested before (CV00) for groups of environments with different sample sizes and six prediction models (M1–M6).

Purpose	CV	Sample size†	M1‡	M3	M5	M2	M4	M6
Predicting environments	CV0	All	0.405	0.399	0.398	0.359	0.318	0.239
		100+	0.414	0.412	0.404	0.375	0.348	0.247
		200+	0.478	0.476	0.466	0.439	0.415	0.292
		300+	0.471	0.47	0.457	0.432	0.409	0.308
	CV00	All	−0.016	0.023	0.028	−0.025	0.043	0.015
		100+	−0.008	0.014	0.025	−0.04	0.031	0.002
		200+	−0.012	0.04	0.057	−0.02	0.035	0.007
		300+	−0.003	0.027	0.045	−0.022	0.026	0.006

† All, 31 environments were included; 100+, the 26 environments where 100 or more lines were tested; 200+, the 17 environments where 200 or more lines were tested; 300+, the 14 environments where 300 or more lines were tested.

‡ L, line effect; E, environment (site–year combination) effect; G, main effect of genomic markers; G × E, genotype × environment interaction; S, site effect; G × S, genotype × site interaction; M1, L + E; M3, L + E + G; M5, L + E + G + (G × E); M2, L + S; M4, L + S + G; M6, L + S + G + (G × S).

Table 7. Correlation between predictive and observed values for two cross-validation scenarios for wheath trials leaving one site out: prediction of crop performance in unobserved sites using information on lines from other sites (CV0) and when none of these lines and sites has ever been tested before (CV00) for groups of environments with different sample sizes and six prediction models (M1–M6).

Purpose	CV	Sample size†	M1‡	M3	M5	M2	M4	M6
Predicting sites	CV0	All	0.404	0.398	0.393	0.364	0.320	0.352
		100+	0.413	0.412	0.401	0.382	0.351	0.375
		200+	0.481	0.478	0.465	0.449	0.420	0.437
		300+	0.473	0.472	0.456	0.437	0.406	0.422
	CV00	All	−0.009	0.031	0.032	0.012	0.042	0.038
		100+	0.000	0.031	0.035	0.015	0.035	0.028
		200+	0.005	0.060	0.060	0.001	0.045	0.044
		300+	0.000	0.052	0.047	0.001	0.041	0.041

† All, 31 environments were included; 100+, the 26 environments where 100 or more lines were tested; 200+, the 17 environments where 200 or more lines were tested; 300+, the 14 environments where 300 or more lines were tested.

‡ L, line effect; E, environment (site–year combination) effect; G, main effect of markers; G × E, genotype × environment interaction; S, site effect; G × S, genotype × site interaction; M1, L + E; M3, L + E + G; M5, L + E + G + (G × E); M2, L + S; M4, L + S + G; M6, L + S + G + (G × S).

Table 8. Correlation between predictive and observed values for two cross-validation scenarios in wheat trials with forward prediction of years.: prediction of crop performance in future unobserved year using information on lines from previous years (CV0) and when none of these lines and years have ever been tested before (CV00) for groups of environments with different sample sizes and six prediction models (M1–M6) (M1 = L+E, M3 = L+E+G, M5 = L+E+G+ (G × E), M2 = L+S, M4 = L+S+G and M6 = L+S+G+G×S; L: line effect, E: Environment [site-by-year combination] effect, G: main effect of markers, G × E: genotype × environment interaction, S: site effect; G×S: genotype × site interaction).

Purpose	CV	Sample size†	M1‡	M3	M5	M2	M4	M6
Predicting years	CV0	All	0.141	0.112	0.127	0.056	0.070	0.041
		100+	0.169	0.128	0.144	0.078	0.088	0.056
		200+	0.151	0.141	0.159	0.087	0.093	0.056
		300+	0.168	0.151	0.171	0.105	0.103	0.073
	CV00	All	0.01	0.018	0.019	0.024	0.041	0.013
		100+	−0.001	0.01	0.02	−0.007	0.028	0.002
		200+	−0.001	0.033	0.052	−0.004	0.031	0.004
		300+	−0.012	0.021	0.04	−0.011	0.022	0.004

† All, 31 environments were included; 100+, the 26 environments where 100 or more lines were tested; 200+, the 17 environments where 200 or more lines were tested; 300+ includes the 14 environments where 300 or more lines were tested.

‡ L, line effect; E, environment (site–year combination) effect; G, main effect of markers; G × E, genotype × environment interaction; S, site effect; G × S, genotype × site interaction; M1, L + E; M3, L + E + G; M5, L + E + G + (G × E); M2, L + S; M4, L + S + G; M6, L + S + G + (G × S).
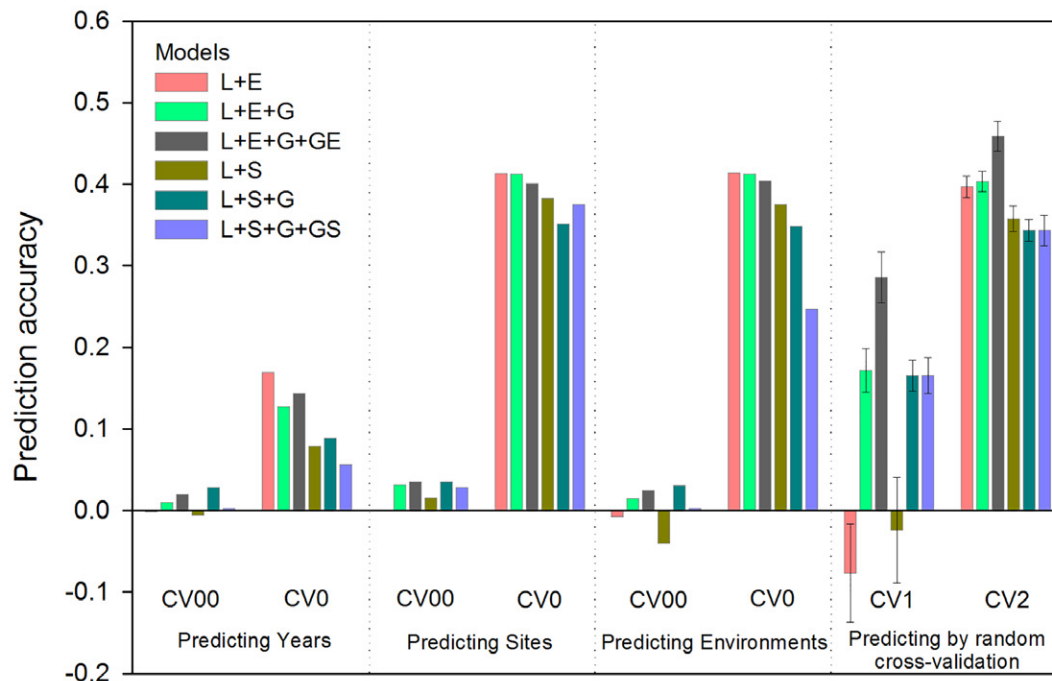
Fig. 3. Average prediction accuracy obtained for the four cross-validation schemes and six models. Values of locations with more than 100 entries are depicted; however, correlations were obtained for all environments.

including the G × E component were observed. Here, most lines had already been observed in other environments, so we could consider them as replicates of lines in the unobserved environments; thus there was plenty of information on these lines, which enabled good predictions for the missing environments. Slight differences on these patterns were observed with those models based on sites (M2, M4, and M6); however, these values never improved the prediction accuracy of previous models. On the other hand, CV00 gave very low values in general for all models, further demonstrating that predictive models perform best when they borrow information from closely related genotypes and environments.

The prediction of sites with the CV0 strategy (Table 7) produced similar predictive ability values as those obtained when predicting environments with M1, M3, and M5; M2, M4, and M6 also improved their performance. However, in spite of these improvements, the predictive ability of M2, M4, and M6 was never as good as their counterparts' predictive ability using this strategy. Using the CV00 strategy, the predictive abilities were technically null.

Forward prediction (Table 8) showed, on average, moderate results when predicting the following year after including previous years in the TST sets. For CV0, the best results were observed with M1 and M5; M3 gave intermediate values, whereas M2, M4, and M6 performed very poorly. On the other hand, none of the six models showed moderate values when CV00 was used for comprising the training sets.

Discussion

The importance of year-to-year variation on the environmental conditions prevailing in the Great Plains would

make the implementation of GS difficult without accounting for G × E. This study highlighted this issue, since environment models were generally more predictive than site models. This may not be true for traits under production in regions that have more stable climates where factors such as soil characteristics play a larger relative role. In those cases, the site model might be superior.

The clearest directive that can be taken from this study is the efficiency that can be gained in early-generation yield testing. Families in early generations could be split with half of the family members being tested in one set of sites and the other half at other sets of sites. This would reduce the number of harvested plots by half while providing good information on performance. The results also suggest that the breeder could test half the members of each family and predict the other half. The number of plots that can be harvested is the primary limiting factor in the size of the KSU hard red winter wheat breeding program. Implementing this strategy allow the program to expand, in terms of the crosses and lines per cross that are evaluated, without increasing the amount of plots that need to be harvested.

Sites previously observed in the breeding pipeline may be selected to best represent one or several target environment types or key stressors in which to predict new lines, which also would allow for *in silico* breeding for either more selected niche targets or broader yield stability. Assuming allele frequencies within a breeding program do not change too radically over a 5- to 10-yr window, it can be speculated that a relatively small set of materials grown at each site in each year could provide a way to classify environments and identify which environments are most representative of yield performance over time. If one were to use the most predictive environments,

our results imply that GS could be a good tool for predicting untested material. This study shows the importance of having an adequate number of lines tested at each site to gain the benefits from GS; thus good training sets are important for the intent of applying GS as an integral part of the breeding effort. Interestingly, the results indicated that there was little advantage to having 300 lines at a site compared to 200. Optimizing this trade-off in the number of entries per site for best prediction accuracy will assist in trial design for incorporation of G × E GS in the breeding program. Additionally, this decision is likely to be crucial when considering other traits such as disease resistance and end-use quality, which are highly heritable and predictable (Battenfield et al., 2016) and should be incorporated in this strategy as well.

The results of this study indicated that even under the complexity and challenge of predicting new environments, sites, or both that have not been previously used in the testing system, good prediction accuracies (around $r = 0.46$) can be achieved when more than 300 lines are evaluated in each environment, highlighting the importance of continued phenotyping in GS applications. An important practical result is that assessing and quantifying G × E improves the prediction accuracy of newly unobserved lines predicted in new environments. Results indicated that predicting the performance in new years in Kansas is the most challenging situation; however, increasing the number of tested lines in each environment to more than 300 increased the prediction accuracy. Again, the results indicate the importance of having enough phenotypic information to be used for predicting unobserved lines in future years. Finally, the prediction of lines that were never tested in environments that were never used is not feasible even for the case of testing more than 300 lines in each tested environment.

Conclusions

Large year-to-year variability is commonplace in the Great Plains of the United States (Holman et al., 2011), which makes phenotypic selection for breeding difficult. This also presents a significant challenge for implementing GS in a two-step approach, where a single BLUP value for each line from a very unbalanced dataset is used to train the model. Here, we demonstrate that using information from environments (site–years) modeled along with genomic data increases predictive ability in GS models in comparisons of variance components for all models and comparisons of GS accuracy between CV0 and CV1 or CV2. Thus, we recommend that efforts in yield GS in breeding programs be designed to utilize information borrowed from related environments either within the current year or from historically observed testing sites.

Designing breeding programs to exploit cross-validation accuracy in genomic and environment-enabled prediction models will assist in resource allocation within the KSU hard red winter wheat breeding program. Here, we have demonstrated (with CV2) that the amount of plots could be reduced by implementing incomplete replication designs, subsequently increasing the number of lines evaluated. The CV2 strategy would impute all lines into all environments, reducing the land used, workload in harvesting and trial preparation, and increasing efficiency per plot. Trends were also found in these data that indicated that having around 200 lines per site was ideal for GS accuracy, which will assist in planning the number of sites and lines tested in the breeding program.

We also demonstrated that new experimental lines may be predicted prior to phenotyping by using historically known environment, as simulated in CV1, which is highly desired by breeders. As genotyping costs continue to decrease and the cost for yield plots either remains constant or increases (Poland and Rife, 2012), screening hundreds or thousands of candidate lines initially through GS becomes a more attractive tool for breeders. In our study, CV1 showed intermediate predictability when more than 300 lines were evaluated in each environment and when genomic models exploit the G × E information (0.368). This can be used as a tool to make an initial cull of materials that could potentially be tested in yield trials.

Overall, designing the breeding pipeline to take advantage of various cross-validation strategies can allow the breeding program to grow while still using the same amount of plots. The use of *in silico* breeding can help breeders make decisions without adding extra workload to the program and without the burden of increasing expenditure. The results of this study indicate the the importance of having enough phenotypic data on lines tested in environments for achieving intermadiate to high prediction accuracy, as well as the need to model and exploit the information existing in G × E interactions.

Supplemental Information

Supplemental Figure S1. Allocation of experimental lines (*x* axis) by environments (*y* axis); environments were defined as site–year combinations. Black vertical lines indicate the particular combinations of lines × environments that were observed in the field.

Conflict of Interest

The authors declare no conflicts of interest.

References

Bassi, F.M., A.R. Bentley, G. Charmet, R. Ortiz, and J. Crossa. 2016. Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci., 242:23–36. doi:10.1016/j.plantsci.2015.08.021

Battenfield, S.D., C. Guzmán, R.C. Gaynor, R.P. Singh, R.J. Peña, S. Dreisigacker, et al. 2016. Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. Plant Genome 9. doi:10.3835/plantgenome2016.01.0005

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. Tassel: Software for association mapping of complex

traits in diverse samples. Bioinformatics 23:2633–2635. doi:10.1093/bioinformatics/btm308

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 52:707–719. doi:10.2135/cropsci2011.06.0299

Cochran, W.G., and G.M. Cox. 1957. Experimental designs, 2nd ed. John Wiley & Sons: Oxford, UK.

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

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

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

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

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

Federer, W.T., Raghavarao, D. 1975. On augmented designs. Biometrics 3, 29:35. doi:10.2307/2529707

Gaynor, R.C. 2015. GSwGBS: An R package genomic selection with genotyping-by-sequencing. Genomic selection for Kansas wheat. K-State Research Exchange, Manhattan, KS.

Heslot, N., D. Akdemir, M.E. Sorrells, and J.L. Jannink. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor. Appl. Genet. 127:463–480. doi:10.1007/s00122-013-2231-5

Holman, J.D., A.J. Schlegel, C.R. Thompson and J.E. Lingenfelser. 2011. Influence of precipitation, temperature, and 56 years on winter wheat yields in western Kansas. Crop Manag. 10 (1). doi:10.1094/cm-2011-1229-01-rs.

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

Jarquin, D., Specht,J. and A. Lorenz. (2016). Prospects of genomic prediction in the USDA Soybean Germplasm Collection: Historical

data creates robust models for enhancing selection of accessions. G3 (Bethesda) 6: 2329–2341. doi:10.1534/g3.116.031443.

Lado, B., P.G. Barrios, M. Quincke, P. Silvac, and L. Gutiérrez. 2016. Modeling genotype × environment interaction for genomic selection with unbalanced data from a wheat breeding program. Crop Sci. 56:2165–2179. doi:10.2135/cropsci2015.04.0207

Lin, C.S., and G. Poushinsky. 1986. A modified augmented design (Type 2) for rectangular plots. Can. J. Plant Sci. 65:743–749. doi:10.4141/cjps85-094

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

Patterson, H.D., and E.R. Williams. 1976. A new class of resolvable incomplete block designs. Biometrika 63:83–92. doi:10.1093/biomet/63.1.83

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

Poland, J.A., and T.W. Rife. 2012. Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92–102. doi:10.3835/plantgenome2012.05.0005

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:E32253. doi:10.1371/journal.pone.0032253

Rutkoski, J., J. Poland, S. Mondal, E. Autrique, L. González-Pérez, J. Crossa, ,et al. (2016). Predictor traits from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3 (Bethesda) 6(3):2799–2808.

Saint Pierre, C., J. Burgueño, J. Crossa, G.F. Dávila, P.F. López, E.S. Moya, et al. 2016. Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. Sci. Rep. 6:27312. doi:10.1038/srep27312

Technow, F., C.D. Messina, L.R. Totir, and M. Cooper. 2016. Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PLoS ONE 10(6):e0130855. doi:10.1371/journal.pone.0130855

USDA National Agricultural Statistical Services. 2014. Kansas wheat history. http://www.nass.usda.gov/Statistics_by_State/Kansas/Publications/Crops/whthist.pdf (accessed 8 May 2017).

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980

Velu, G., J. Crossa, R.P. Singh, Y. Hao, P. Rodriguez-Perez, A.K. Joshi, et al. 2016. Genomic prediction for grain zinc and iron in wheat. Theor. Appl. Genet. 129:1595. doi:10.1007/s00122-016-2726-y