agronomy

MDPI

Article

Increasing Predictive Ability by Modeling Interactions between Environments, Genotype and Canopy Coverage Image Data for Soybeans

Diego Jarquin [1,*], Reka Howard [2], Alencar Xavier [3] and Sruti Das Choudhury [4]

1 Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
2 Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA; rekahoward@unl.edu
3 Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA; alenxav@gmail.com
4 School of Natural Resources and Department of Computer Science & Engineering, University of Nebraska-Lincoln, Lincoln, NE 68583, USA; srutidc@gmail.com
* Correspondence: diego.jarquin@gmail.com; Tel.: +1-402-613-9013

Received: 26 February 2018; Accepted: 16 April 2018; Published: 17 April 2018

Abstract: Phenomics is a new area that offers numerous opportunities for its applicability in plant breeding. One possibility is to exploit this type of information obtained from early stages of the growing season by combining it with genomic data. This opens an avenue that can be capitalized by improving the predictive ability of the common prediction models used for genomic prediction. Imagery (canopy coverage) data recorded between days 14–71 using two collection methods (ground information in 2013 and 2014; aerial information in 2014 and 2015) on a soybean nested association mapping population (SoyNAM) was used to calibrate the prediction models together with the inclusion of several types of interactions between canopy coverage data, environments, and genomic data. Three different scenarios were considered that breeders might face testing lines in fields: (i) incomplete field trials (CV2); (ii) newly developed lines (CV1); and (iii) predicting lines in unobserved environments (CV0). Two different traits were evaluated in this study: yield and days to maturity (DTM). Results showed improvements in the predictive ability for yield with respect to those models that solely included genomic data. These relative improvements ranged 27–123%, 27–148%, and 65–165% for CV2, CV1, and CV0, respectively. No major changes were observed for DTM. Similar improvements were observed for both traits when the reduced canopy information for days 14–33 was used to build the training-testing relationships, showing a clear advantage of using phenomics in very early stages of the growing season.

Keywords: genomic prediction; genotype by environment interaction; interaction models; canopy coverage; cross-validation schemes

1. Introduction

To meet the needs of feeding an increasing world population [1] and to satisfy people's dietary needs, it is important to increase food production. Soybean is a major oil seed in the United States; more than 90% of the oil seed production comes from soybean programs [2]. Since soybeans have a low production cost, they have the potential to increase yield, and are produced mostly as part of a crop rotation with maize, where it is essential to increase soybean yield to reduce the per-bushel production cost. Furthermore, the harvestable land area cannot be significantly increased for soybean, so we must sustainably improve the yield potential of soybean to enhance food production, thus there is a need to develop methods that enable us to increase soybean yield. Nowadays, modern platforms can be used for monitoring large planted regions intensively in time and space to deliver accurate

information about the specific (physiological) conditions of the plants, thereby allowing a better characterization of the response of these genotypes to specific stress stimuli. Thus, we can characterize genotypes using high dimensional marker information and high dimensional phenotypic information.

Genomic Prediction (GP) techniques have become an important part of plant breeding programs due to their advantages when compared to traditional phenotypic and pedigree-based selections [3]. GP is a technique that aids selection for yield and quality related traits, and it has been shown [4] that it has the potential to lead to a threefold increase in genetic gain when compared to marker assisted selection.

In GP, the marker and phenotypic information of individuals in the training set are used to model the relationship between the phenotype and genotype, then the model is used to predict the phenotype for individuals in the testing set for which only marker information is available. GP was first introduced by Meuwissen et al. [5], and since then an extended number of models have been developed for phenotype prediction incorporating marker information [6–8]. Early applications of GP for the selection of soybean varieties were introduced by Jarquin et al. [9].

The prediction accuracy of GP methods can be improved by including the genotype by environment (G × E) component by borrowing information from related materials and correlated environments. Jarquin et al. [10] utilized the reaction norm model for genomic prediction where the genetic and environmental values were replaced by the regression on the markers, and in the interaction between the markers and the environmental covariates, respectively. Dealing with high-throughput phenotypic information, several authors [11–13] have shown improvements in predictive ability with the inclusion of these sources of information in the models for wheat and maize. Montesinos-Lopez et al. [14] showed that accounting for the band (hyper-spectral image data)-by-environment interaction also improved yield predictability in wheat when compared with those models that did not include this component in the models.

Herein, we extended the reaction norm model for prediction using canopy coverage image data. Xavier et al. [15] incorporated canopy coverage image data into a selection scheme by studying the large effect QTL associated with canopy coverage. A QTL (Quantitative Trait Loci) is a particular region in the genome that is statistically associated with the variation of one or several traits. Since canopy coverage is a trait with high estimated heritability and correlation with grain yield [15], it has the potential to improve prediction models when this information is included into the model.

We used the soybean nested association mapping (soyNAM) population data to implement our prediction models. The soyNAM data used consisted of phenotypic yield data on 5600 F5-derived recombinant inbred lines, over 4000 single nucleotide polymorphism (SNP) markers, and ground-based imagery and/or aerial imagery data, depending on the year.

This article is organized as follows. First, we provide a brief description of the soyNAM phenotypic and marker data that were used for the predictions. Then, we describe how the canopy coverage image data were collected, and why it has the potential to increase prediction accuracy compared to traditional GP models when included into the prediction models. In the next section, the statistical models and cross validation (CV) schemes implemented for genomic-enabled prediction are described. We compared nine prediction models with three different CV schemes for yield and days to maturity (DTM) for the soyNAM data set. Additionally, we compared the effects of predictive ability with models that included canopy data captured during the early stages of the growing season (days 14–33) instead of the whole growing season. Finally, we discuss the results, and some future research avenues.

2. Material and Methods

2.1. SoyNAM Phenotypic and Genotypic Data

The predictions were conducted using a soybean nested association mapping population (SoyNAM) containing 5600 recombinant inbred lines (RILs) created by crossing a common high

yielding parent (IA3023) to 40 parents. For a detailed description of the structure of the soyNAM population the reader can refer to Xavier et al. [16]. For the analysis, 39 families and 5143 RILs were kept due to uncertainty in the phenotypic and/or genotypic data.

Phenotypic data were collected on grain yield (in kg/ha), days to maturity (in days after planting), plant height (in cm), lodging (score 1–5), seed size (mass of 100 seed in grams), seed composition (in % in the grain) of protein, oil, and fiber content. For this study, however, the focus was on grain yield and days to maturity.

A total of 5303 single nucleotide (SNP) markers were used to design the SoyNAM 6K BreadChip array [17]. Markers with a minor allele frequency of less than 0.15 were removed, and missing markers were imputed using a random forest algorithm [18]. After quality control, the final set of markers was comprised of 4240 SNPs.

For more information about the soyNAM parents, RILs, the experimental design, the collected phenotypes and corresponding genotypes, the reader can refer to Xavier et al. [16].

2.2. Canopy Coverage and Imagery Data Collection

Grain yield in soybean is influenced by genetic and environmental factors and their interactions. Among the environmental factors, drought, salinity and temperature have the largest impact on the agronomical traits (i.e., yield) [19]. Many of these factors are hard to measure, therefore, even though they have a large influence on the trait, it is not cost effective to include them in prediction models. Canopy light interception (LI) is one of the important factors that influences yield and growth in soybean, and is difficult to measure [20]. Canopy coverage, which is the proportion of ground area covered by the soybean plant, is a trait that can be more easily measured, and can be used as a replacement for canopy LI measurements. High-throughput phenotyping platforms allow us to capture information about a trait in a non-destructive way. It also permits the collection of time-series measurements at a relatively low cost that is essential to follow the growth change of a plant [21], and aids in quantifying the genotype by environmental interaction.

The canopy data used for the analyses were obtained either as ground-based imagery (in 2013 and 2014) or as aerial imagery data (in 2014 and 2015) as described by Xavier et al. [15]. Xavier et al. noted that the correlation between the ground-based data and the aerial data was high (r^2 = 0.87) for 2014, thus in 2015 only aerial imagery data were collected, and in our case, we only used data from the aerial platform for 2014. Thus, our models were calibrated using ground data for 2013 and aerial data for 2014 and 2015. The data were collected at different time points at regular intervals ranging from two to eight weeks after planting to provide information on the different phases of growth and crop physiology. The collected data were classified to determine the canopy coverage, which was defined as the proportion of image pixels that were canopy pixels to the total number of pixels for any given field plot. The canopy coverage data of the ground-based imageries were determined using the software SigmaScan Pro® (SYSTAT, San Jose, CA, USA), which implements the method described by Karcher and Richardson [22]. The canopy coverage data of aerial imageries were obtained by the software ENVI 5.0™ (Harris Geospatial Solutions, Boulder, CO, USA) using a binomial model.

2.3. Relationship between Canopy Coverage Data and Grain Yield

In order to use canopy coverage data successfully to increase the predictive ability of the prediction models, it should show some type of relationship with the traits that are predicted. In this study, for each year (2013–2015)–by–acquisition method (ground and aerial) combination, we computed the correlation between the interpolated canopy values (days 14–71) and the observed grain yield values (Figure 1). The four data sets showed different patterns; however, in all cases the correlation decreased as the number of days increased. The 2013-ground and 2014-ground sets reached the highest correlation (0.27 and 0.34) on days 33 and 32, respectively. For the aerial method, the highest correlations (0.33 and 0.44) were reached on days 45 and 41, respectively. To assess the usefulness of canopy coverage data in the early stages of the vegetative growth, the information from days 14–30

(for the four data sets) was included in the prediction models, and the results were compared with those obtained using canopy information from the whole season.

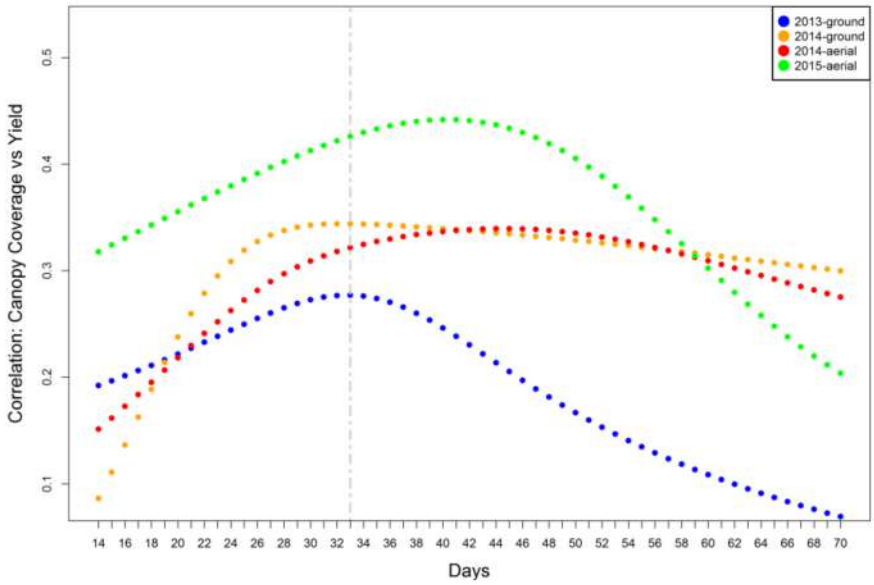


Figure 1. Correlation between the interpolated daily canopy coverage values and yield for data sets including 2013-ground, 2014-ground, 2014-aerial, and 2015-aerial canopy coverage data.

2.4. Statistical Prediction Models

In this study we evaluated nine prediction models: six main effects models that included combinations of line, environment, marker genotype, and canopy coverage image information; seven models with two-way interaction(s) among the components; and two models with a three-way interaction between environments, marker genotypes, and the canopy coverage data. The interaction components were modeled using the reaction norm model [10]. All models assumed that the components were random effects. For all of the models, we considered two responses: grain yield and DTM. Since the canopy coverage imagery data were either ground based or collected from the air, we incorporated two types of canopy coverage image data. In the statistical models described below, the term CC (Canopy Coverage) denotes the canopy coverage where the ground-based imagery data were included from 2013 and the aerial canopy coverage image data were included from 2014 and 2015.

2.5. Main Effects Models

2.5.1. Model 1: Environment + Line

The response of the j^{th} genotype in the i^{th} environment $\{y_{ij}\}$ can be written as:

$$y_{ij} = \mu + E_i + L_j + e_{ij} \tag{1}$$

where μ is the overall mean, E_i ($i = 1, \ldots, I$) is the random effect of the i^{th} environment assuming $E_i \overset{iid}{\sim} N\left(0, \sigma_E^2\right)$ with $N(.,.)$ denoting a normal density, where iid stands for independent and identically distributed observations, and σ_E^2 represents the variance component of environments; L_j represents the random effect of the j^{th} line ($j = 1, \ldots, J$) assuming $L_j \overset{iid}{\sim} N\left(0, \sigma_L^2\right)$, where σ_L^2 is the variance of the line; and e_{ij} is the random error term where $e_{ij} \overset{iid}{\sim} N\left(0, \sigma_e^2\right)$ with σ_e^2 as the residual variance. All the models described below include the terms specified for Model 1 with additional components.

2.5.2. Model 2: Environment + Line + CC

The only difference between this model and Model 1 is that the CC component is also included. The model can be described as follows:

$$y_{ij} = \mu + E_i + L_j + CC_{ij} + e_{ij} \quad (2)$$

where μ, E_i, L_j and e_{ij} are defined as before, and CC_{ij} is the canopy coverage imagery data. It is defined as $CC_{ij} = \sum_{k=14}^{70} P_{ijk} c_k$ where P_{ijk} is the k^{th} canopy measurement for genotype j in environment i, c_k is the effect of the k^{th} canopy measurements, and $c_k \overset{iid}{\sim} N(0, \sigma_c^2)$. In general, $CC \overset{iid}{\sim} N\left(0, \frac{\boldsymbol{PP'}}{K}\sigma_{\boldsymbol{CC}}^2\right)$, where $\boldsymbol{P}$ is the centered and standardized (per columns) canopy matrix. For later derivations we define $\boldsymbol{\omega} = \frac{\boldsymbol{PP'}}{K}$ as the canopy coverage covariance structure between pairs of lines × environment combinations. The entries of the matrix describe the canopy coverage similarities between pairs of lines × environment combinations. The number of total canopy measurements for a genotype in an environment after the extrapolation of missing daily measures was $K = 57 = 70 - 13$.

2.5.3. Model 3: Environment + Line + Marker

In this model the response of the j^{th} genotype in the i^{th} environment $\{y_{ij}\}$ can be modeled as a linear function of the environment effect, line effect, marker effect, and a random residual:

$$y_{ij} = \mu + E_i + L_j + g_j + e_{ij} \quad (3)$$

where μ, E_i, L_j and e_{ij} are defined as before, and g_j is the linear combination of p markers and the corresponding marker effects such that $g_j = \sum_{m=1}^{p} x_{jm} b_m$. The marker effects are assumed to be normally distributed such that $b_m \overset{iid}{\sim} N(0, \sigma_b^2)$ for $m = 1, \ldots, p$, where σ_b^2 is the marker effect variance. If we write the genomic values in a vector form so that $\boldsymbol{g} = (g_1, \ldots, g_J)$, then the covariance matrix can be expressed in the form $Cov(\boldsymbol{g}) = \boldsymbol{G}\sigma_{\boldsymbol{g}}^2$, where $\boldsymbol{G} = \frac{\boldsymbol{XX'}}{p}$ is the genomic relationship matrix [23], $\boldsymbol{X}$ is the centered and standardized genotype matrix, and $\sigma_{\boldsymbol{g}}^2 = p \times \sigma_b^2$. In summary, we can write that $\boldsymbol{g} = \{g_j\} \sim N\left(\boldsymbol{0}, \boldsymbol{G}\sigma_{\boldsymbol{g}}^2\right)$.

2.5.4. Model 4: Environment + Line + Marker + CC

This model is the combination of Model 2 and Model 3 as both the marker and CC canopy component are added in addition to the environment and line components:

$$y_{ij} = \mu + E_i + L_j + g_j + CC_{ij} + e_{ij} \quad (4)$$

where all of the terms are defined before (Models 2 and 3).

2.6. Two-Way Interaction Models

2.6.1. Model 5: Environment + Line + CC + (CC × Environment Interaction)

This model is an extension of Model 2 where the interaction term between canopy and environments is added:

$$y_{ij} = \mu + E_i + L_j + CC_{ij} + CCE_{ij} + e_{ij} \quad (5)$$

where all of the terms except CCE_{ij} are defined as before (for Model 2). CCE_{ij} is the canopy × environment measurement interaction term with $\boldsymbol{CCE} = \{CCE_{ij}\} \sim N\left(\boldsymbol{0}, (\boldsymbol{\omega})^{\circ}(\boldsymbol{Z_E Z_E'})\sigma_{\boldsymbol{CCE}}^2\right)$, where $\boldsymbol{Z_E}$ is the incidence matrix for the environments, which connects the environments to the phenotypes, $\sigma_{\boldsymbol{CCE}}^2$ is the variance component for the interaction term, $\boldsymbol{\omega}$ is defined previously, and '$^{\circ}$' stands for the Hadamard or Schur (element-by-element or cell-by-cell) product between two matrices.

2.6.2. Model 6: Environment + Line + Marker + (Marker × Environment Interaction)

This model accounts for the environment, line, and genomic main effects as per Model 3, but it also incorporates the genotype × environment interaction via co-variance structures, as per the environment × canopy interaction in Model 5. In this case, the model can be written as

$$y_{ij} = \mu + E_i + L_j + g_j + gE_{ij} + e_{ij} \quad (6)$$

where $\boldsymbol{gE} = \{Eg_{ij}\} \sim N\left(\mathbf{0}, \left(\boldsymbol{Z_g G Z_g'}\right)^{\circ}(\boldsymbol{Z_E Z_E'})\sigma^2_{\boldsymbol{gE}}\right)$, where $\boldsymbol{Z_g}$ and $\boldsymbol{Z_E}$ are the incidence matrices for the lines and environments, respectively, $\sigma^2_{\boldsymbol{gE}}$ is the variance component of the gE_{ij} interaction component, and **G** is the additive relationship matrix defined previously.

2.6.3. Model 7: Environment + Line + Marker + CC + (Marker × CC Interaction)

This model is an extension to Model 5 as it does not only account for the main effect of the environment, line, marker and the canopy coverage, but also includes the interaction between the marker and the canopy coverage:

$$y_{ij} = \mu + E_i + L_j + g_j + CC_{ij} + gCC_{ij} + e_{ij} \quad (7)$$

where all of the terms except gCC_{ij} are defined as before (for Model 6). gCC_{ij} is the marker × canopy measurement interaction term with $\boldsymbol{gCC} = \{gCC_{ij}\} \sim N\left(\mathbf{0}, \left(\boldsymbol{Z_G G Z_G'}\right)^{\circ}(\boldsymbol{\omega})\sigma^2_{\boldsymbol{gCC}}\right)$, where $\boldsymbol{Z_E}$ is the incidence matrix for the genotypes, which connects the genotypes and the phenotypes, $\sigma^2_{\boldsymbol{gCC}}$ is the variance component for the genotype × canopy coverage interaction term.

2.6.4. Model 8: Environment + Line + Marker + CC + (Marker × Environment Interaction) + (CC × Environment Interaction)

This model is a combination of Models 6 and 7 as it accounts for the main effect of environment, line, marker, and canopy coverage, and the interactions between the marker and environment and between the environment and the canopy coverage:

$$y_{ij} = \mu + E_i + L_j + g_j + CC_{ij} + gE_{ij} + CCE_{ij} + e_{ij} \quad (8)$$

where all of the terms are the same as defined previously.

2.7. Three-Way Interaction Models

Model 9: Environment + Line + Marker + CC + (Marker × Environment Interaction) + (CC × Environment Interaction) + (Marker × CC × Environment Interaction)

This three-way interaction model is an extension of Model 8 with the addition of the marker × canopy coverage × environment interaction:

$$y_{ij} = \mu + E_i + L_j + g_j + CC_{ij} + gE_{ij} + CCE_{ij} + gCCE_{ij} + e_{ij} \quad (9)$$

where the main effects and two-way interaction terms are defined the same as previously, and $\boldsymbol{gCCE} = \{gCCE_{ij}\} \sim N\left(\mathbf{0}, (\boldsymbol{Z_E Z_E'})^{\circ}\left(\boldsymbol{Z_G G Z_G'}\right)^{\circ}(\boldsymbol{\omega})\sigma^2_{\boldsymbol{gCCE}}\right)$, where $\boldsymbol{Z_E}$, $\boldsymbol{Z_G}$, $\boldsymbol{G}$, and $\boldsymbol{\omega}$ are defined as before, and $\sigma^2_{\boldsymbol{gCCE}}$ is the variance component for the three-way interaction term.

2.8. Description of Cross-Validation Schemes Implemented for Assessing Predictive Ability

The performance of the models was compared based on the Pearson correlation coefficient between the predicted phenotypic values and the observed phenotypic values. Three cross validation

techniques (CV2, CV1, and CV0) were implemented and compared when predicting yield and DTM. The cross-validation schemes mimicked real plant breeding situations.

CV2 is the case where some lines were observed in some environments but not in others, and we attempted to predict the performance of these unobserved line × environment combinations. This scheme mimicked the situation of predicting incomplete field trials. The graphical representation of CV2 is shown in Figure 2a. Figure 2a is a simplified representation of CV2 where we observed values for five lines in five environments, and the goal was to predict the phenotype of the lines that were not observed in a particular environment. Y_{ij} represents the phenotypic value for the i^{th} line in the j^{th} environment in the training set where $i = 1, \ldots, 5$ and $j = 1, \ldots, 5$, and NA represents the lines for which the phenotype needed to be predicted (testing set).

CV1 refers to the case where the performance of lines was evaluated in some environments, and some different genotypes were predicted in the same environments. The graphical representation of CV1 is shown in Figure 2b. Here, the performance of a new developed line (line 3) needed to be predicted.

CV0 is the cross-validation scheme where the performance of lines was evaluated in some environments, and the goal was to predict the performance of the already tested lines in untested environments. To determine the training and testing sets for the CV0 prediction, the leave-one-environment out scheme was implemented, and since there was no random partitioning involved in the CV0 method, it was only implemented once.

For both the CV1 and CV2 schemes, five-fold random partitions (80% of data was used as the training set, and the remaining 20% was used as the testing set) were used to determine the training and testing sets, and the partitioning was repeated 50 times.

	E1	E2	E3	E4	E5
Line 1	Y_{11}	NA	Y_{13}	Y_{14}	Y_{15}
Line 2	Y_{21}	Y_{22}	NA	Y_{24}	Y_{25}
Line 3	Y_{31}	Y_{32}	Y_{33}	Y_{34}	NA
Line 4	Y_{41}	Y_{42}	Y_{43}	NA	Y_{45}
Line 5	NA	Y_{52}	Y_{53}	Y_{54}	Y_{55}

(**a**)

	E1	E2	E3	E4	E5
Line 1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}
Line 2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
Line 3	NA	NA	NA	NA	NA
Line 4	Y_{41}	Y_{42}	Y_{43}	Y_{44}	Y_{45}
Line 5	Y_{51}	Y_{52}	Y_{53}	Y_{54}	Y_{55}

(**b**)

	E1	E2	E3	E4	E5
Line 1	Y_{11}	Y_{12}	NA	Y_{14}	Y_{15}
Line 2	Y_{21}	Y_{22}	NA	Y_{24}	Y_{25}
Line 3	Y_{31}	Y_{32}	NA	Y_{34}	Y_{35}
Line 4	Y_{41}	Y_{42}	NA	Y_{44}	Y_{45}
Line 5	Y_{51}	Y_{52}	NA	Y_{54}	Y_{55}

(**c**)

Figure 2. Graphic representation of the Cross-Validation schemes: (**a**) CV2, incomplete field trials; (**b**) CV1, prediction of newly developed lines; and (**c**) CV0, predicting crop performance in new environments. NA, stands for a not available record (missing value).

3. Results and Discussion

3.1. Analysis of Variance Components

The estimated variance components for the nine models for yield are shown in Table 1, and for DTM in Table 2. For both traits, a large amount of the total variation was accounted by the environmental variation. However, the environmental variation was reduced when the marker and canopy information were included into the model. Furthermore, the residual variation was reduced when the interaction components were added in the models. For yield, the proportion of the variability explained by the residual term was reduced from 30.7% to 20.9%. The DTM also showed a reduction of the residual variance from 22.9% to 10.7%.

Table 1. Estimated variance components for the nine models for yield.

Model	No.	Estimated Variance Components								
		E	L	G	GE	CC	CCE	GCC	GCCE	R
E + L	1	59.0	10.3							30.7
E + L + G	2	49.9	2.4	14.9						32.8
E + L + G + GE	3	46.5	4.0	10.4	10.4					28.7
E + L + CC	4	51.7	10.1			7.0				31.3
E + L + CC + CCE	5	63.4	8.9			0.1	0.1			27.5
E + L + G + CC	6	48.1	1.6	13.7		8.8				27.8
E + L + G + CC + GCC	7	42.8	1.9	15.7		8.1		3.4		28.1
E + L + G + CC + GE + CCE	8	55.1	2.4	8.3	7.8	5.0	0.4			20.9
E + L + G + CC + GE + CCE + GCCE	9	44.3	2.8	11.7	7.8	5.0	0.6		2.8	25.0

E = environment, L = line, G = genotype, GE = genotype × environment interaction, CC = canopy information, CCE = canopy × environment interaction, GCC = genotype × environment × canopy interaction, GCCE = genotype × canopy × environment interaction, R = residual term.

Table 2. Estimated variance components for the nine models for days to maturity.

Model	No.	Estimated Variance Components								
		E	L	G	GE	CC	CCE	GCC	GCCE	R
E + L	1	45.1	32.0							22.9
E + L + G	2	19.8	5.2	61.2						13.8
E + L + G + GE	3	17.9	6.4	61.0	2.7					12.0
E + L + CC	4	31.5	37.5			3.0				28.0
E + L + CC + CCE	5	35.8	37.1			0.1	0.1			26.9
E + L + G + CC	6	15.2	5.4	63.4		1.1				14.9
E + L + G + CC + GCC	7	12.3	6.0	64.6		1.2		1.3		14.5
E + L + G + CC + GE + CCE	8	18.8	5.7	52.1	2.2	0.0	10.4			10.7
E + L + G + CC + GE + CCE + GCCE	9	13.3	6.0	54.1	2.3	0.0	12.8		0.7	10.9

E = environment, L = line, G = genotype, GE = genotype × environment interaction, CC = canopy information, CCE = canopy × environment interaction, GCC = genotype × environment × canopy interaction, GCCE = genotype × canopy × environment interaction, R = residual term.

3.2. Assessment of Predictive Ability

In this study, nine models were evaluated in terms of prediction accuracy. The main objective was to identify whether incorporating canopy coverage data into the prediction models would increase the prediction accuracy, and whether the inclusion of interaction terms would further improve the models. Jarquin et al. [10] showed the advantage of including the genotype × environment interaction into the prediction models, and here we wanted to determine whether the inclusion of interaction terms among the environment, the genotype, and the canopy coverage measurement would enhance the prediction accuracy.

Out of the nine models, only three did not include any canopy coverage: Model 1: Environment + Line, Model 3: Environment + Line + Marker, and Model 6: Environment + Line + Marker + (Marker × Environment Interaction). We evaluated the models using SNP marker information, canopy coverage measurement, and phenotypic information collected on yield and DTM on the SoyNAM recurrent inbred lines. The phenotypic information and the canopy coverage measurements were collected

in 2013, 2014, and 2015. All of the prediction accuracies were calculated implementing the CV2, CV1, and CV0 cross-validation schemes.

Figures 3a and 4a show the results for the nine models using CV2 for yield and DTM, respectively. Similarly, Figures 3b and 4b present the results for CV1, and Figures 3c and 4c for CV0.

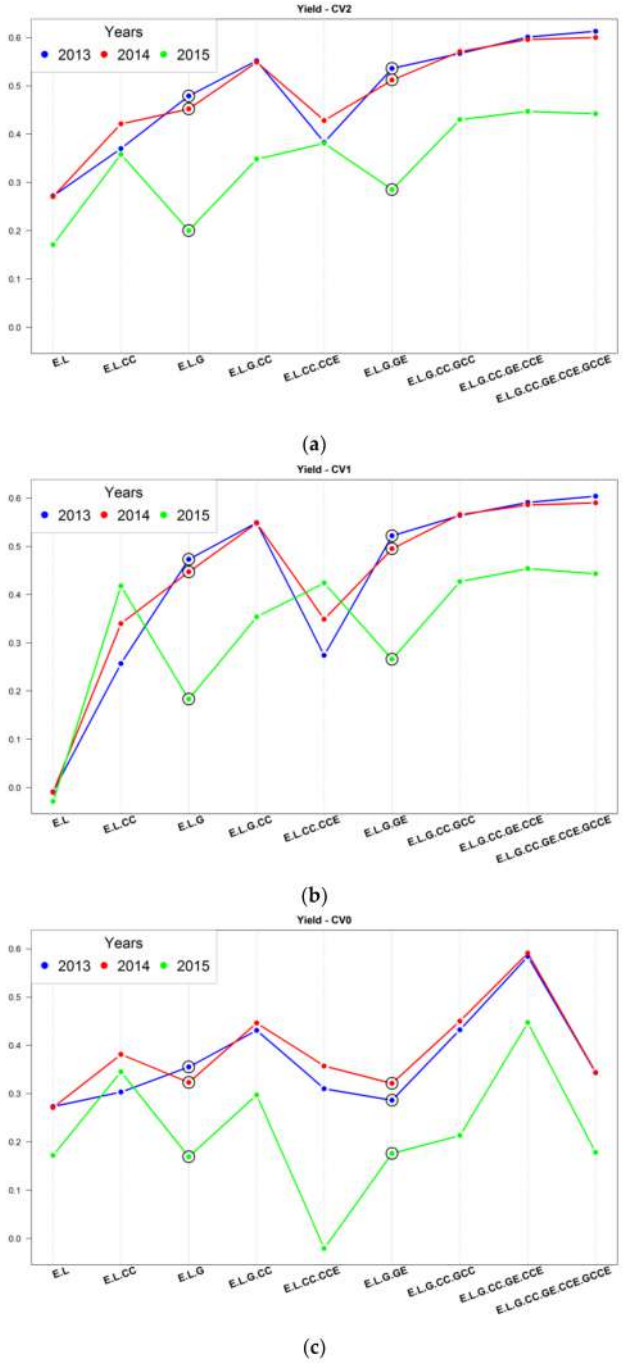


Figure 3. Average prediction accuracy obtained for the nine models for yield using the CV2 (**a**); CV1 (**b**); and CV0 (**c**) cross-validation schemes. The three different colors represent the years for which the prediction was carried out. E = environment, L = line, G = genotype, GE = genotype × environment interaction, CC = canopy information, CCE = canopy × environment interaction, GCC = genotype × environment × canopy interaction, GCCE = genotype × canopy × environment interaction.

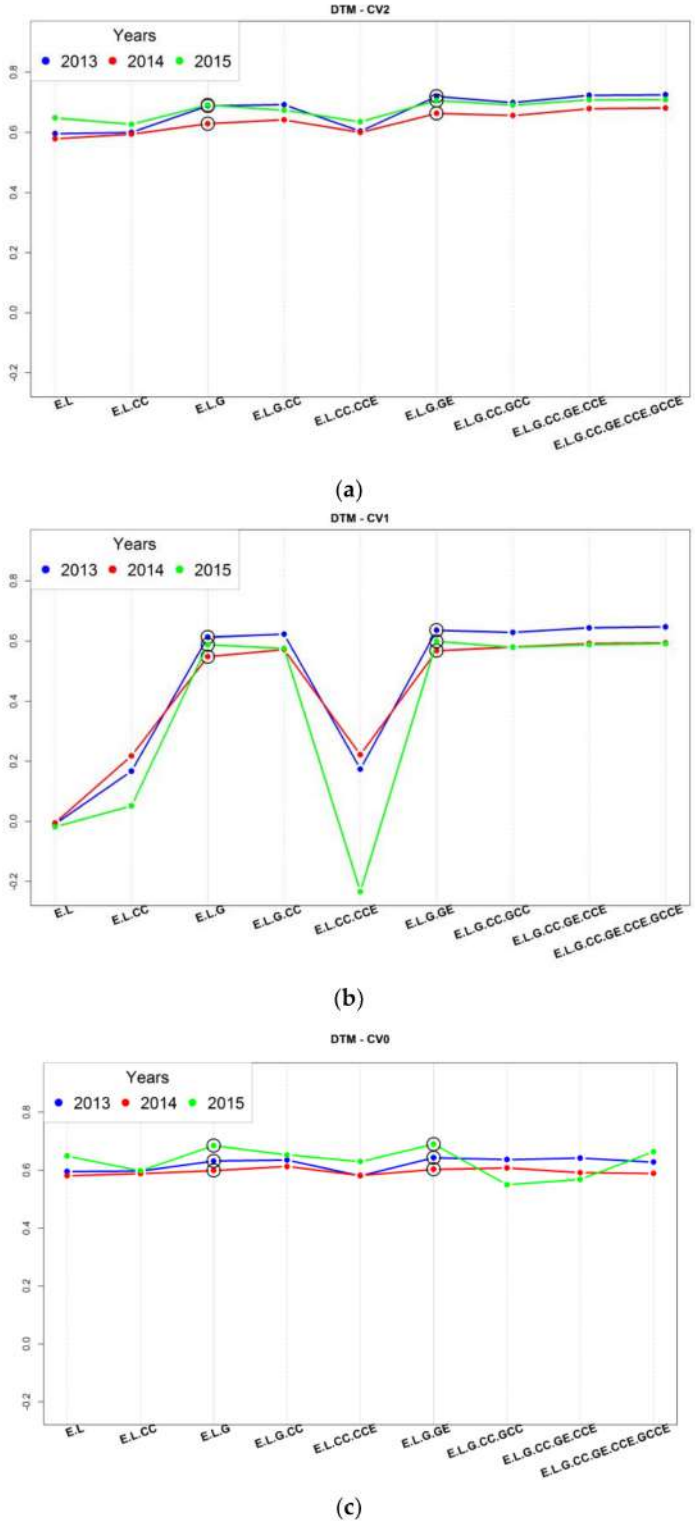


Figure 4. Average prediction accuracy obtained for the nine models for days to maturity (DTM) using the CV2 (**a**); CV1 (**b**); and CV0 (**c**) cross-validation schemes. The three different colors represent the years for which the prediction was carried out.

The performance of the CV1 scheme depended mostly on the genetic similarities between the training and testing sets while CV2 and CV0 also accounted for the environmental variations via replicates of the same genotypes observed in other environments.

When we considered yield as the trait to be predicted, for the CV1 and CV2 cross-validation schemes, models 7–9 performed better than the models that did not include the marker, canopy data and environmental interactions. In these cases, the correlation between the predicted and observed values was around 0.61 for 2013 and 2014, and 0.42 for 2015. These values were slightly increased when more terms were included in the model. The common feature among the models was that all of these models (7–9) included the genotypic information, the canopy coverage information, and at least one interaction component involving canopy coverage and marker data. For CV0, model 8 reached the same levels of predictive ability than the CV2 and CV1 schemes.

For all of the cross-validation techniques, most of the models had the lowest mean prediction accuracy when the prediction was carried out for 2015. In addition, similar patterns were shown when contrasting results for 2013 vs. 2014. In general, all of the cross-validation schemes showed significant variations among the models in terms of predictive ability. However, in all cases, predictive ability was improved when the interactions involving canopy data were included.

When we evaluated the models for DTM, and when CV2 and CV0 were implemented, we did not see a large variation among the models in terms of accuracy of prediction. The prediction accuracy of all models was within 0.55 and 0.65. For CV1, models 1, 2, and 5 had significantly lower prediction accuracy than the rest of the models. These were the only models that did not include genotype as a main effect term or an interaction with environments. For DTM, we could not identify a year that had the highest prediction accuracy across the cross-validation schemes. For all of the three cross-validation techniques, model 6—the Environment + Line + Marker + (Marker × Environment interaction)—performed the best, but for CV0 and CV2 the difference between model 6 and some of the other models was not significant.

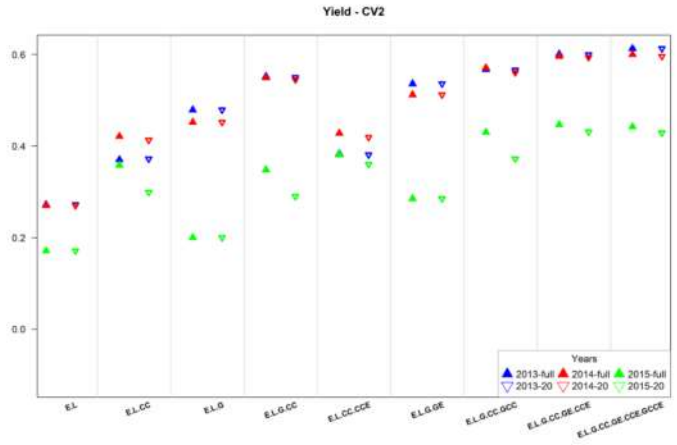
When predicting yield, we clearly saw the advantage of including the canopy coverage measurements, and the interactions among markers, environment, and canopy coverage measurement. The highest predictive ability for CV2 and CV1 were delivered by model 9 (three-way interaction model) while for CV0, model 8 produced the highest values.

For DTM, the advantage of including these terms was not as evident as for yield, but including the canopy coverage measurement as a main effect and as part of interactions did not hurt the model performance when the markers were also present in the model.

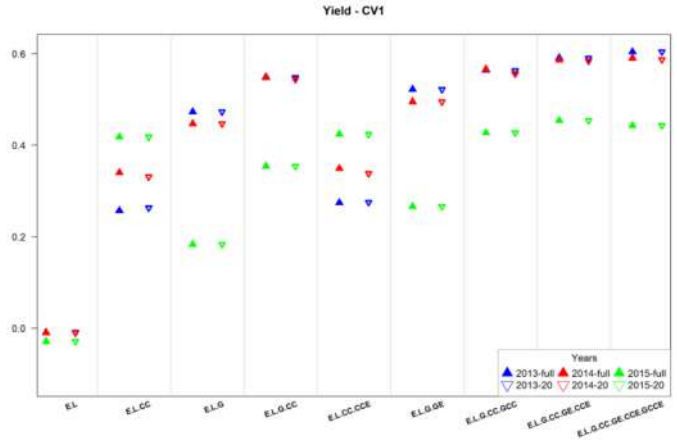
3.3. Effectiveness of Canopy Data from Early Stages

As mentioned before, another objective of this study was to compare the usefulness of canopy data collected during the early stages (days 14–33) of the growing season with the whole data set (days 14–71). Figures 5 and 6 show the correlation between the predicted and observed values for the nine models using the reduced (days 14–33) and the whole canopy sets (days 14–71) for yield and DTM, respectively. For yield (Figure 5), in schemes CV2 and CV1, the whole and reduced sets showed similar results. The whole set was always slightly better in terms of prediction accuracy than the reduced set. With CV0, in most of the cases, the reduced set provided better results than the whole set. For DTM (Figure 6), the whole and reduced sets performed the same for CV2 and CV1, while for CV0 there were a few cases where the reduced set slightly outperformed the results of the full set.

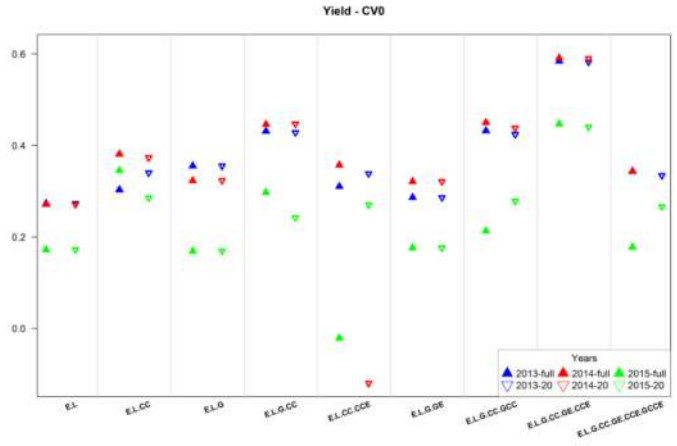
Since the results of the reduced set of canopy values measured between days 14 and 33 were similar to those obtained using the information from the whole set (days 14–71), we are confident about the applicability of this technique to improve the predictive ability of the genomic prediction models using information from the early stages of the growing season. An important practical implication of these results is that we could reach the same degree of predictive ability using canopy coverage data collected during the very early stages of the growing season instead of collecting the canopy information throughout the whole growing season.



(**a**)



(**b**)



(**c**)

Figure 5. Average prediction accuracy obtained for the nine models for yield using the CV2 (**a**); CV1 (**b**); and CV0 (**c**) cross-validation schemes for the complete (days 14–71) and reduced (days 14–33) canopy sets. The three different colors represent the years for which the prediction was carried out.

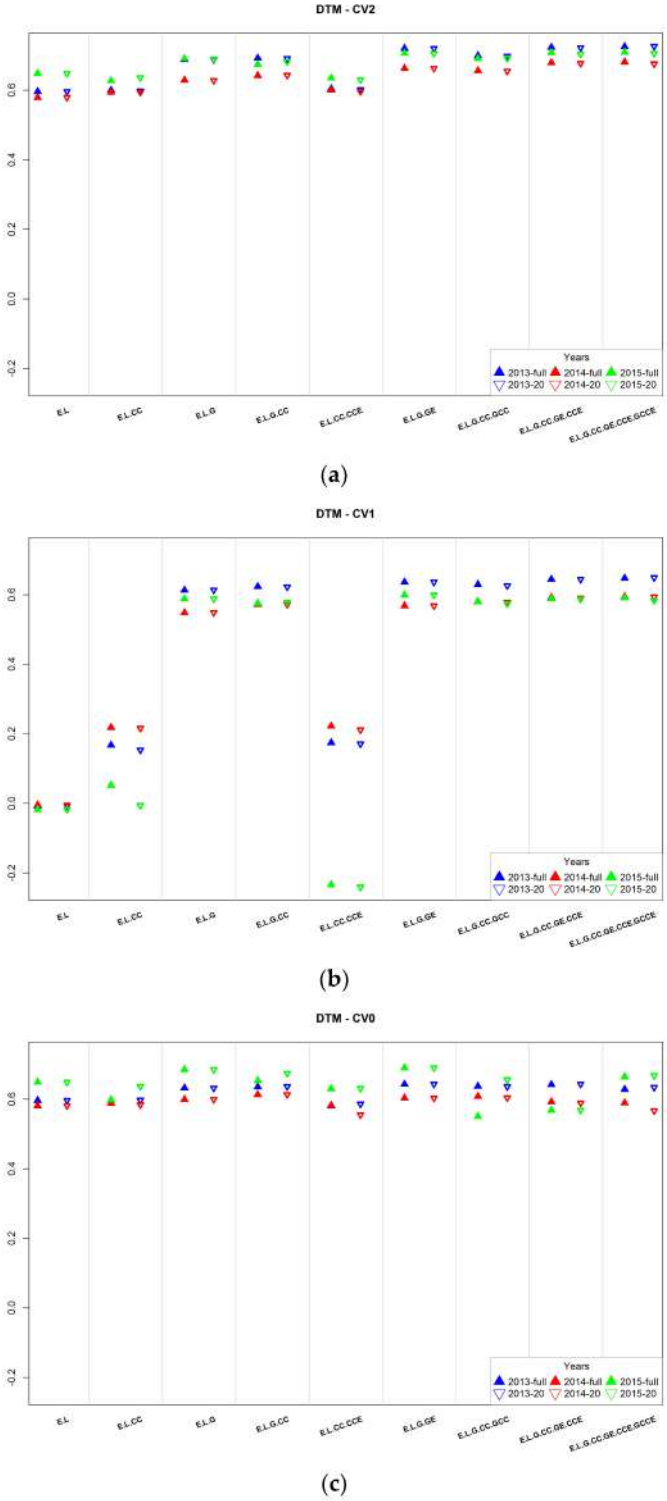


Figure 6. Average prediction accuracy obtained for the nine models for DTM using the CV2 (**a**); CV1 (**b**); and CV0 (**c**) cross-validation schemes for the complete (days 14–71) and reduced (days 14–33) canopy sets. The three different colors represent the years for which the prediction was carried out.

4. Conclusions

In our study we evaluated nine prediction models, from which six included the canopy coverage information either as a main effect or as an interaction effect with the genotype and/or environment

information. The models were compared in terms of prediction accuracy for three different cross validation schemes (CV2, CV1, and CV0) and for two different traits (yield and DTM). We compared the model performance in two situations; when all of available canopy image data were fitted, and when the canopy image data was used only for days 14–33 (20 days). Our results indicated no significant difference between model performances when all of the available canopy information was utilized versus when the canopy data for only 20 days of the growing season was included into the models. When we compared the models with the common genomic prediction model (Model 3) for yield we observed a 27–123% improvement for CV2, a 27–148% improvement for CV1, and a 65–165% improvement for CV0 depending on the model. For DTM the improvements were 2-8% for CV2, 3–8% for CV1, and 1–2% for CV0 depending on the model.

Since the ultimate goal of performing predictions of unobserved genotypes is to save time and resources developing new improved cultivars, planting those genotypes for collecting canopy data might not be helpful to shorten breeding cycles to avoid spending money and resources to phenotype these lines. However, as shown, the integration of canopy coverage data and marker data improved model performance. Thus, this might allow a more accurate selection of superior breeding lines as pointed out by Crain et al. [12]. From a more practical perspective, the use of the canopy coverage information taken during the early stages of the growing season could aid the selection process when no phenotypes for yield and/or DTM are available in the case where breeders prefer phenotypic selection. In this scenario, when an extreme hydro-climatic (hurricane, tornado, etc.) event occurs, it might partially or completely destroy the cultivars before the harvest season, leaving no chance to perform phenotypic selection. Thus, the use of canopy coverage data might allow us to recover valuable information about these destroyed cultivars, therefore improving the accuracy of selection. Some future work should attempt to investigate the use of canopy information to predict the performance of future generations. In this case, our proposed solution is to predict the canopy values of untested lines, and use these predicted values as covariates for predicting yield and/or DTM.

Acknowledgments: We did not receive any financial support for this project.

Author Contributions: Diego Jarquin developed the concept idea and prediction models. Reka Howard developed the concept idea and implemented the analysis. Alencar Xavier contributed ideas for the analysis and provided the canopy image data. Sruti Das Choudhury contributed ideas to the manuscript and reviewed the writing. All authors wrote the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

GP	Genomic prediction
CV0	cross-validation predicting the performance of previously tested lines in untested locations
CV1	cross-validation evaluating the performance of new developed lines, lines that have not been evaluated in any of the observed environments
CV2	cross-validation evaluating the performance of lines that have been evaluated in some environments but not in others, incomplete field trials
SoyNAM	soybean nested association mapping

References

1. Whitford, R.; Fleury, D.; Reif, J.C.; Garcia, M.; Okada, T.; Korzun, V.; Langridge, P. Hybrid breeding in wheat: Technologies to improve hybrid wheat seed production. *J. Exp. Bot.* **2013**, *64*, 5411–5428. [CrossRef] [PubMed]
2. Zulauf, C.; Coppess, J.; Paulson, N.; Schnitkey, G. U.S. Oilseeds: Production and Policy Comparison. *Farmdoc Daily* **2017**, *7*, 28.
3. De los Campos, G.; Gianola, D.; Rosa, G.J.M.; Weigel, K.A.; Crossa, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* **2010**, *92*, 295–308. [CrossRef] [PubMed]

4. Heffner, E.L.; Lorenz, A.J.; Jannink, J.L.; Sorrells, M.E. Plant breeding with genomic selection: Potential gain per unit time and cost. *Crop Sci.* **2010**, *50*, 1681–1690. [CrossRef]
5. Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829. [PubMed]
6. Endelman, J.B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **2011**, *4*, 250–255. [CrossRef]
7. Gianola, D.; Fernando, R.L.; Stella, A. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* **2006**, *173*, 1761–1776. [CrossRef] [PubMed]
8. Habier, D.; Fernando, R.L.; Kizilkaya, K.; Garrick, D.J. Extension of the bayesian alphabet for genomic selection. *BMC Bioinf.* **2011**, *12*, 186. [CrossRef] [PubMed]
9. Jarquín, D.; Kocak, K.; Posadas, L.; Hyma, K.; Jedlicka, J.; Graef, G.; Lorenz, A. Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genom.* **2014**, *15*, 740. [CrossRef] [PubMed]
10. Jarquín, D.; Crossa, J.; Lacaze, X.; Cheyron, P.D.; Daucourt, J.; Lorgeou, J.; Piraux, F.; Guerreiro, L.; Pérez, P.; Calus, M. A reaction norm model for genomic selection using highdimensional genomic and environmental data. *Theor. Appl. Genet.* **2014**, *127*, 595–607. [CrossRef] [PubMed]
11. Aguate, F.M.; Trachsel, S.; Pérez, L.G.; Burgueño, J.; Crossa, J.; Balzarini, M.; Gouache, D.; Bogard, M.; Campos, G.D.L. Use of Hyperspectral Image Data Outperforms Vegetation Indices in Prediction of Maize Yield. *Crop Sci.* **2017**, *57*, 2517–2524. [CrossRef]
12. Crain, J.; Mondal, S.; Rutkoski, J.; Singh, R.P.; Poland, J. Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. *Plant Genome* **2018**. [CrossRef] [PubMed]
13. Montesinos-Lopez, O.A.; Montesinos-Lopez, A.; Crossa, J.; De los Campos, G.; Alvarado, G.; Mondal, S.; Rutkoski, J.; Gonzalez-Perez, L.; Burgueño, J. Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. *Plant Methods* **2017**, *13*. [CrossRef] [PubMed]
14. Montesinos-López, A.; Montesinos-Lopez, O.A.; Cuevas, J.; Mata-López, W.A.; Burgueño, J.; Mondal, S.; Huerta-Espino, J.; Singh, R.P.; Autrique, E.; Gonzalez-Perez, L.; et al. Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper spectral image data. *Plant Methods* **2017**, *13*, 62. [CrossRef] [PubMed]
15. Xavier, A.; Hall, B.; Hearst, A.A.; Cherkauer, K.A.; Rainey, K.M. Genetic Architecture of Phenomic-Enabled Canopy Coverage in Glycine max. *Genetics* **2017**, *206*, 1081–1089. [CrossRef] [PubMed]
16. Xavier, A.; Muir, W.M.; Rainey, K.M. Assessing Predictive Properties of Genome-Wide Selection in Soybeans. *G3 (Bethesda)* **2016**, *6*, 2611–2616. [CrossRef] [PubMed]
17. Song, Q.; Yan, L.; Quigley, C.; Jordan, B.D.; Fickus, E.; Schroeder, S.; Song, B.; An, Y.C.; Hyten, D.; Nelson, R.; et al. Genetic Characterization of the Soybean Nested Association Mapping Population. *Plant Genome* **2017**, *10*. [CrossRef] [PubMed]
18. Stekhoven, D.J.; Buhlmann, P. Missforest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef] [PubMed]
19. Wang, W.; Vinocur, B.; Altman, A. Plant responses to drought, salinity and extreme temperatures: Towards genetic engineering for stress tolerance. *Planta* **2003**, *218*, 1–14. [CrossRef] [PubMed]
20. Purcell, L.C. Soybean canopy coverage and light interception measurements using digital imagery this paper is published with the approval of the director of the Arkansas Agric. Exp. Stn. (manuscript number 99–107). *Crop Sci.* **2000**, *40*, 834–837. [CrossRef]
21. Fahlgren, N.; Gehan, M.A.; Baxter, I. Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. *Curr. Opin. Plant Biol.* **2015**, *24*, 93–99. [CrossRef] [PubMed]
22. Karcher, D.E.; Richardson, M.D. Batch analysis of digital images to evaluate turfgrass characteristics. *Crop Sci.* **2005**, *45*, 1536–1539. [CrossRef]
23. VanRaden, P.M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **2008**, *91*, 4414–4423. [CrossRef] [PubMed]