9-27-2021

Development of a Genomic Prediction Pipeline for Maintaining Comparable Sample Sizes in Training and Testing Sets across Prediction Schemes Accounting for the Genotype-by-Environment Interaction

Reyna Persa

Martin Grondona

Diego Jarquin

agriculture

MDPI

Article

Development of a Genomic Prediction Pipeline for Maintaining Comparable Sample Sizes in Training and Testing Sets across Prediction Schemes Accounting for the Genotype-by-Environment Interaction

Reyna Persa [1], Martin Grondona [2] and Diego Jarquin [1,*]

[1] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68588, USA; reynapersa@gmail.com
[2] Advanta Seeds, College Station, TX 77845, USA; martin.grondona@advantaseeds.com
* Correspondence: diego.jarquin@gmail.com

Abstract: The global growing population is experiencing challenges to satisfy the food chain supply in a world that faces rapid changes in environmental conditions complicating the development of stable cultivars. Emergent methodologies aided by molecular marker information such as marker assisted selection (MAS) and genomic selection (GS) have been widely adopted to assist the development of improved genotypes. In general, the implementation of GS is not straightforward, and it usually requires cross-validation studies to find the optimum set of factors (training set sizes, number of markers, quality control, etc.) to use in real breeding applications. In most cases, these different scenarios (combination of several factors) vary just in the levels of a single factor keeping fixed the levels of the other factors allowing the use of previously developed routines (code reuse). In this study, we present a set of structured modules that are easily to assemble for constructing complex genomic prediction pipelines from scratch. Also, we proposed a novel method for selecting training-testing sets of sizes across different cross-validation schemes (CV2, predicting tested genotypes in observed environments; CV1, predicting untested genotypes in observed environments; CV0, predicting tested genotypes in novel environments; and CV00, predicting untested genotypes in novel environments). To show how our implementation works, we considered two real data sets. These correspond to selected samples of the USDA soybean collection (D1: 324 genotypes observed in 6 environments scored for 9 traits) and of the Soybean Nested Association Mapping (SoyNAM) experiment (D2: 324 genotypes observed in 6 environments scored for 6 traits). In addition, three prediction models which consider the effect of environments and lines (M1: E + L), environments, lines and main effect of markers (M2: E + L + G), and also the inclusion of the interaction between makers and environments (M3: E + L + G + G×E) were considered. The results confirm that under CV2 and CV1 schemes, moderate improvements in predictive ability can be obtained with the inclusion of the interaction component, while for CV0 mixed results were observed, and for CV00 no improvements were shown. However, for this last scenario, the inclusion of weather and soil data potentially could enhance the results of the interaction model.

Keywords: genotype-by-environment interaction (G×E); genomic prediction (GP); genomic prediction pipeline; genomic selection (GS); similar sample sizes for cross-validation schemes; SoyNAM; USDA soybean collection

check for updates

Citation: Persa, R.; Grondona, M.; Jarquin, D. Development of a Genomic Prediction Pipeline for Maintaining Comparable Sample Sizes in Training and Testing Sets across Prediction Schemes Accounting for the Genotype-by-Environment Interaction. *Agriculture* **2021**, *11*, 932. https://doi.org/10.3390/agriculture11100932

Received: 23 July 2021
Accepted: 7 September 2021
Published: 27 September 2021

1. Introduction

The world confronts several challenges for satisfying the increased demands to feed the growing human population, which is projected to grow close to 10 billion by 2050 [1,2]; however, not only the population is increasing but also the natural resources (e.g., forest, soil, land, and water availability, etc.) have been drastically affected due to environmental problems such as deforestation and land degradation [1]. In addition, in agriculture, the elite genotypes (high yield performance) have been negatively impacted due to the more often and more intense environmental perturbances. To guarantee the food requirements and confront these challenges, the new varieties yet to develop (in the very near future) will require to come up with a better resilience for a wide range of adaptation [3]. For this, new strategies and methodologies for selecting genotypes to face these environmental challenges [4] with high yield potential should be developed.

Breeders have implemented traditional breeding methods for selecting the best phenotypes to increase genetic gains [5,6]. However, phenotyping all genotypes in a wide range of environmental conditions is challenging because it requires a large number of plots in field experiments, it is also time-consuming of manual labor, and in general, there is a reduced availability of sources such as land, water, and seed. A more elaborated traditional breeding method considers the use of the pedigree information derived from the genetic relationship between the genotypes in the population [7–9]. Pedigree based selection has been successful delivering predictions of the estimated breeding values of unobserved genotypes; however, its implementation presents some challenges [10]. For example, it requires to keep track of the genetic relationships between all genotypes in training and testing sets. Also, it does not account for the Mendelian segregation within populations from a pair of genotypes limiting the rates of genetic progress that can be accomplished in a given period of time [11].

Hence, the traditional selection methods based on phenotypic and pedigree information may not be the most suitable options for increasing genetic gains in short periods of time. Specially, because it is not easy to a priori estimate the recombination amount of the genome that comes from each of the parental lines [6] complicating the selection process. The development of modern sequencing technologies offered the opportunity of characterizing genotypes based on their genomic information [11]. A widely used alternative to the traditional selection methods is the Marker Assisted Selection (MAS)which uses genomic information [12]. It considers a reduced set of influential molecular markers also known as quantitative trait loci (QTL) [13,14] to assist during the selection process. The main objective of MAS is to help to select the best candidate genotypes best candidate genotypes by predicting their phenotypic performance using the most influential genomic variants. Furthermore, this methodology has demonstrated to be more effective than the pedigree-based selection method [12–14]. However, this method also presents some limitations specially when the traits are controlled by a large number of genes with small effects (complex traits) such as yield [15] limiting/reducing the accuracy of the selection.

To overcome the limitations of MAS, the implementation of an emergent methodology called GS became popular in the last decade in plant and animal breeding applications across species and traits. Conceptually, this method uses the information on all available molecular markers for selection purposes [11]. This methodology was first proposed by Bernardo [16] and later on Meuwissen [17] introduced a new framework to confront the challenge of dealing with a large set of markers (p) and a reduced number of phenotypic records (n) available for model fitting.

GS enables the prediction of the performance of genotypes at the early stages of the breeding programs using abundant molecular marker information of new/untested genotypes and a relative small number of genotypes with phenotypic and genomic information for model calibration. Such that, the predicted values can be used for selecting the best candidate genotypes that would perform well on advanced phenotyping stages [18]. Another advantage of GS is that breeders can reduce phenotyping costs by employing predictions as surrogates of phenotypes. At beginning, GS was used in plant breeding for performing within-environments predictions only [18–21]. In general, breeders establish extensive field experiments for testing new cultivars in a wide range of environmental conditions and release stable genotypes that outperforms current elite cultivars [4]. However, usually different response patterns are observed when same genotypes are observed in different environments showing a change in the relative ranking from one environment to another complicating the selection process [22]. The occurrence of a change in the response patterns is also known as the presence of the genotype-by-environment interaction (G×E).

Several studies highlighted the impacts of accounting for the G×E in GS models [23–25] when performing predictions in multi environments. Several applications have been developed to conduct the predictions of genotypes in single and multi-environments [26–32]. However, to our knowledge, no comprehensive implementations/examples showing how the genomic prediction pipelines are built have been released. Among the tasks that a GS pipeline considers we have the implementation of quality control on genomic data, the assignation of training and testing sets for different cross-validation schemes, the construction of the different linear predictors, and the model fitting.

The main objective of this study is to provide an example of how a genomic prediction pipeline is built when considering different cross-validation schemes while preserving comparable sample sizes in training and testing sets. For this, we considered two soybean data sets. The first one (D1) corresponds to a sample of the USDA soybean collection with information on 324 genotypes tested in 6 environments (not all genotypes tested in all environments) and 9 traits. The second dataset (D2) corresponds to a sample of the SoyNAM experiment with information on 324 genotypes tested in 6 environments and 6 traits (all genotypes scored for all traits in all environments). The pipeline was built considering elemental modules that perform simple tasks and their implementation is controlled by changes in a parameter input file. Also, the outputs of the early stages of the pipeline become the input of more advanced stages allowing the assemble of complex structures in an easy way. Potentially, users will be able to easily modify and adapt the proposed pipeline to conduct their own data analyses (data sets for a desired set of parameters).

2. Materials and Methods

2.1. Phenotypic and Genomic Data

In this research, two different soybean datasets were used to show how the pipeline is implemented and these correspond to a sample of the USDA soybean collection, and a sample of the SoyNAM experiment.

Data set 1 (D1). Sample of the USDA soybean collection.

The USDA soybean collection is comprised of 14,430 genotypes that were collected in many locations around the world and observed in 4 different locations (States; Illinois, Kentucky, Minnesota, and Missouri) in the USA from 1963 to 2003. Not all the genotypes were observed in all the location-by-year combinations (environments). Further details of the USDA soybean collection can be found in Bandillo [33]. The evaluation of the soybean genotypes in the US locations was gradually conducted. For this reason, the connectivity rate of genotypes across environments is very low. In this study, conveniently we selected a reduced set of genotypes (324) that were observed in 6 environments (MN945, IL945, IL0102, MS989, MS2000_2, and MN0102) and showed moderate levels of connectivity. We selected genotypes that were observed in at least 2 environments and with complete information on all 9 traits (grain yield, plant height in centimeters, lodging 1–5, days to physiological maturity - DysToR8, oil content, protein content, seed weight of 100 seeds, early shattering 1–5, and stem term score 1–5). Figure 1 illustrates the levels of connectivity of the genotypes across environments for this data set (D1). Out of the 100% of the total (324 × 6 = 1944) potential cells (all genotypes observed in all environments) only 33.6% (654) of these combinations were observed (vertical gray lines in Figure 1) in fields.

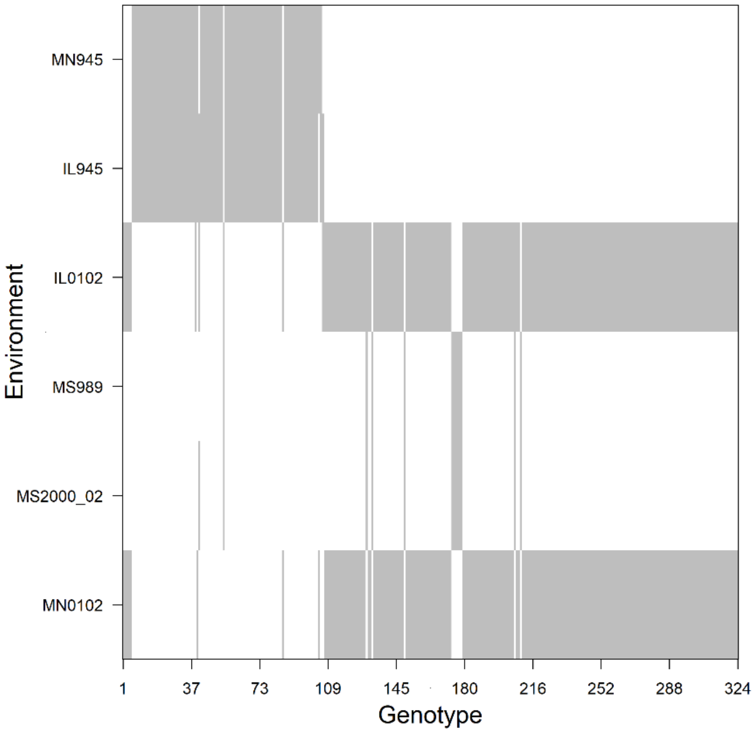

Figure 1. Graphical representation of the allocation of genotypes (x-axis) in environments (y-axis) of 324 soybean genotypes selected from the USDA soybean collection observed in 6 environments. Vertical gray lines represent the genotypes-in-environments combinations that were observed while the white lines correspond to unobserved combinations. The information of the observed combinations (vertical gray lines) is used for establishing training-testing partitions.

Data set 2 (D2). Sample of the SoyNAM project.

A random sample of the SoyNAM project was selected for the second dataset. The SoyNAM project is comprised of 40 biparental populations (140 individuals per population) sharing a common hub parent (IA3023) crossed with elite parents (17), plant introductions (8), and parents with exotic ancestry (15). [34,35] provide a detailed description of the SoyNAM data set. Briefly, the resulting 5400 accessions derived of the 40 biparental populations were observed in 18 location × year combinations (environments). In our case, conveniently we selected a random sample of 324 genotypes observed in 6 environments (IA_2012, IL_2011, IL_2012, IN_2012, NE_2011, and NE_2012) and measured for 6 traits (yield, moisture, protein content, oil content, fiber, and seed size). In this case, all the 324 genotypes were observed in all the 6 environments and were scored for all 6 traits.

2.2. Models

The main objective of this research is to provide a genomic prediction pipeline easy to adapt to different realistic scenarios of interest in breeding programs. Here, a set of 3 models was considered to show how to compose the linear predictors to use in different cross-validation schemes (4). Alternative models can be obtained by changing the assumptions of the model terms. Later we describe how to perform these changes in an easy manner.

2.2.1. M1. E + L

Consider that y_{ij} represents the performance of the i^{th} genotype observed in the j^{th} environment for a given trait (e.g., grain yield) and it can be described as the sum of a constant common effect across lines and environments (μ), a fixed effect due to the environmental stimuli (E_j) corresponding to the j^{th} environment, a random effect (L_i) corresponding to the i^{th} line such that $L_i \sim N(0, \sigma_L^2)$, and a random effect (ε_{ij}) capturing the non-explained variability by the previous model terms with $\varepsilon_{ij} \sim N(0, \sigma^2)$. Collecting the previous assumptions, we have that the linear predictor becomes

$$y_{ij} = \mu + E_j + L_i + \varepsilon_{ij} \tag{1}$$

One disadvantage of this model is that it does not allow the borrowing of information between genotypes complicating the prediction of the untested materials. To overcome this issue, the genomic information of the individuals in training and testing sets can be leveraged together with the phenotypic data from the observed genotypes (training set) to predict un-phenotyped individuals. Details of this approach are provided in model M2.

2.2.2. M2. E + L + G

In the previous model M1, the L_i term is used to describe the effect of the i^{th} genotype and it relies on phenotypic information only. Now, consider that this term can be also described by a linear combination between p molecular markers and their corresponding marker effects such as $g_i = \sum_{k=1}^{p} x_{ik} b_k$, where b_k corresponds to the marker effect of the k^{th} SNP (x_{ik}). When the number of molecular markers (p) surpass the number of data points (n) available for model fitting, it is impossible to obtain a unique solution for the marker effects because it involves the inversion of non-full rank matrices. In these cases, further assumptions about the marker effects should be considered under the statistical framework. Several alternatives have been proposed to overcome this issue and some of these are based on penalized regressions (Ridge Regression, LASSO, ELASTICNET, etc.) and Bayesian approaches (Bayesian Ridge Regression, Bayesian LASSO, BayesA, BayesB, etc.) Meuwissen [17] proposed a set of models for those cases where the number of genomic variants (p) was larger than the number of data points (n) available for model fitting. A compressive review of the available genomic models to deal with this issue can be found in [11].

In our case, the marker effects were considered independent and identically distributed (IID) outcomes from a normal distribution centered on zero with a common variance $\left(\sigma_b^2\right)$ [36,37] such that $b_k \sim N\left(0, \sigma_b^2\right)$. From results of the multivariate normal distribution, the vector of genomic effects $\boldsymbol{g} = \{g_i\} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_g^2\right)$ where $\mathbf{G} = \frac{\boldsymbol{XX}'}{p}$, $\boldsymbol{X}$ is the standardized matrix (by columns) of marker SNPs and $\sigma_g^2 = p\sigma_b^2$ is the corresponding variance component. To avoid model miss specification due to imperfect genomic data the L_i term is also included in the model together with the genomic effect g_i. Considering the previous assumptions, we have that the resulting model becomes

$$y_{ij} = \mu + E_j + L_i + g_i + \varepsilon_{ij} \tag{2}$$

An advantage of this model is that it allows the borrowing of information between tested and untested genotypes permitting the prediction of materials yet to be observed. However, a disadvantage of this model is that across environments it returns the same genomic value g_i for the i^{th} genotype. To allow specific genomic values of genotypes in environments, the reaction norm model [25] was also implemented. This model decomposes the genomic effect as the sum of a common effect (intercept) of the genotypes across environments plus specific effects (slopes) for each environment. Further details of this model are provided next.

2.2.3. M3. E + L + G + G×E

Consider the inclusion of the gE_{ij} model term to describe the specific response of the i^{th} genotype in the j^{th} environment $\left(g_i E_j\right)$. Jarquin [25] proposed to model the vector of genomic effects in interaction with environments via co-variance structures as $gE = \{gE_{ij}\} \sim N\left(0, Z_g G Z_g' \circ Z_E Z_E' \sigma_{gE}^2\right)$, where Z_g and Z_E are the corresponding incidence matrices that connect phenotypes with genotypes and environments, respectively, "$\circ$" represents the cell-by-cell product between two matrices also know as Hadamard or Shur product, and σ^2_{gE} is the corresponding variance component. The resulting linear predictor is

$$y_{ij} = \mu + E_j + L_i + g_i + gE_{ij} + \varepsilon_{ij} \tag{3}$$

2.3. Cross-Validation Schemes

To assess the ability of the different prediction models for delivering accurate results, four prediction scenarios that are of interest for breeders were considered. These prediction scenarios attempt to mimic different realistic prediction problems that breeders might face [38] at different stages of the breeding pipeline. Figure 2 presents the four different cross-validation scenarios using as example a hypothetical population of 48 genotypes to be observed in 6 environments. The different colors (vertical lines) correspond to a fivefold assignation of either phenotypes (CV2, and CV0) or genotypes (CV1, and CV00).

CV2 (tested genotypes in observed environments), corresponds to the scenario of predicting incomplete field trials where some genotypes have been observed in some environments but not in others. In this case, the genotypes of interest are probably observed in other environments and also other genotypes have been already observed in the environment(s) of interest. A random fivefold assignation, represented with different colors (black, gray, red, yellow and blue) in the top left panel of Figure 2, was considered. Here the phenotypic data was randomly assigned to each one of the five folds (colors) maintaining folds of similar size (~20% of the observations). Then four folds are employed for model calibration when predicting the remaining fold, and this procedure is sequentially repeated for all five folds (one at a time).

CV1 (untested genotypes in observed environments), mimics the scenario of predicting genotypes that have not been observed yet at any of the environments and the goal is to predict the performance of these genotypes in environments where other genotypes were already observed. Here, a fivefold cross-validation was implemented by assigning around 20% of the genotypes to folds (bottom left panel in Figure 2) such that all the phenotypic records of a genotype are assigned to the same fold (color) avoiding to encounter phenotypes of the same genotype in different folds. In the bottom left panel in Figure 2, across environments (horizontal lines) the phenotypes of the 48 genotypes have the same color, and those genotypes with the same color belong to the same fold. Same as before, four folds are considered for model training when predicting the remaining fold. This prediction procedure is sequentially repeated for each one of the five folds (one at a time).

CV0 (tested genotypes in unobserved environments), represents the prediction scenario of predicting the mean performance of genotypes in hypothetical unobserved environments. It considers phenotypic information of same and from other genotypes observed in other environments (training set). In this case, the conventional prediction procedure consists of leaving one environment out and then use the remaining environments for model calibration when predicting the excluded environment. This procedure is sequentially repeated for each environment (one at a time). However, in our case, we introduced an alternative way to conduct the prediction of unobserved environments in an attempt for preserving similar sample sizes for training and testing sets than in previous schemes. In this way, it is possible to compare the results of the different cross-validation scenarios with similar sample sizes. The top right panel of Figure 2 illustrates an example that considers the prediction of the genotypes in gray color (horizontal lines) in environment 3. In this case, there is information available of the same genotype but observed in the remaining 5 environments. Here, the same fold assignation as in the CV2 was such that it is possible to conduct a direct comparison of the accomplished predictive ability between these two cross-validation scenarios. The prediction procedure consists of sequentially predicting each one of the five folds in each environment (one at a time). This procedure is repeated for each environment (one at a time).

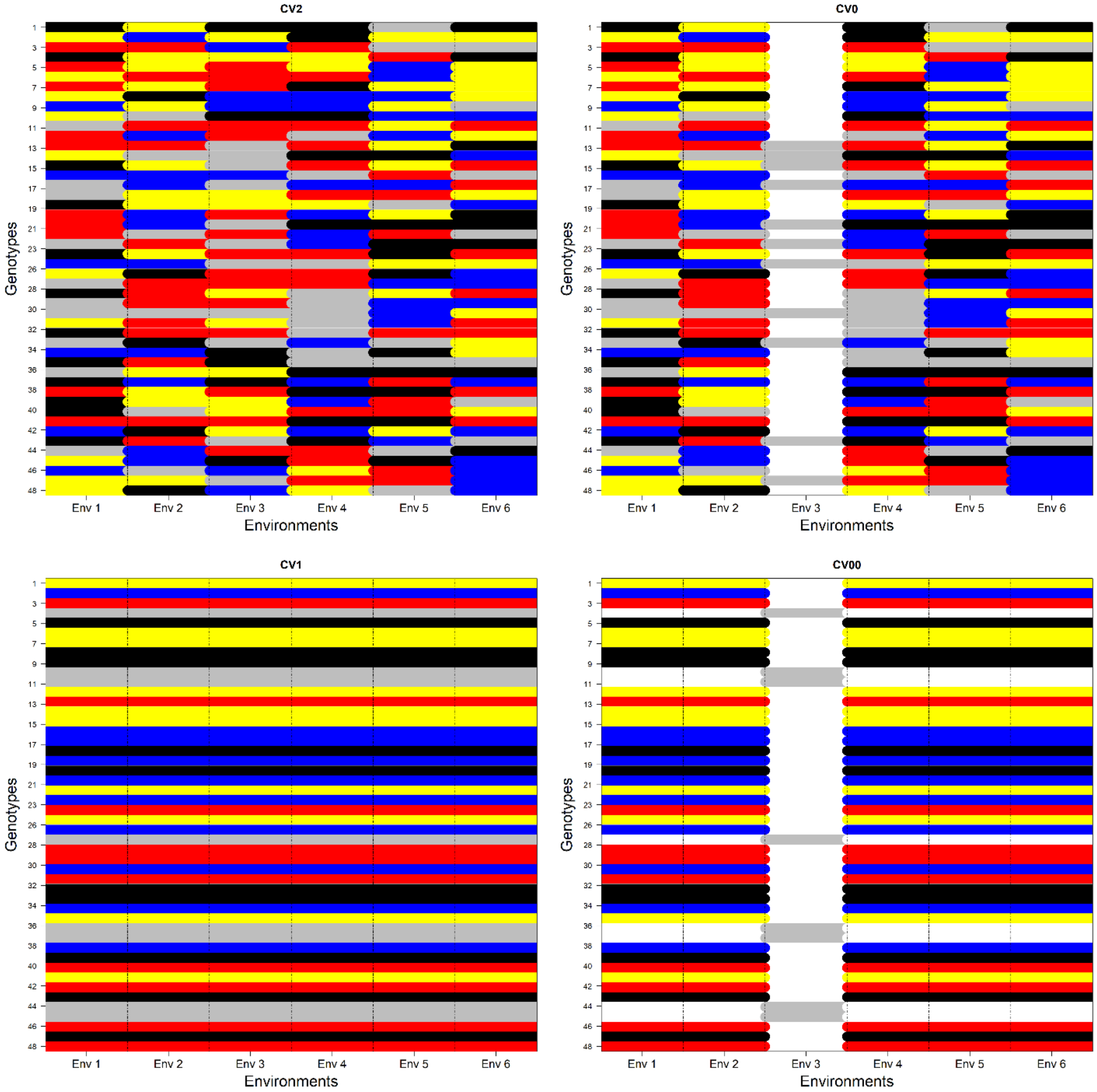

Figure 2. Graphical representation of four cross-validation schemes (CV2, predicting tested genotypes in observed environments; CV1, predicting untested genotypes in observed environments; CV0, predicting tested genotypes in unobserved environments; and CV00, predicting untested genotypes in unobserved environments) that preserve comparable training and testing set sizes. The colored horizontal lines correspond to a fivefold assignation of phenotypes (CV2; **top left**) and genotypes (CV1; **bottom left**) for composing training and testing sets (e.g., consider 4 folds for a training set: black, blue, red and yellow horizontal lines; while for testing set only one fold is considered: gray color lines). For CV0 (**top right**) and CV00 (**bottom right**), the genotypes represented with horizontal gray lines in environment 3 (Env 3) correspond to the target prediction set. In addition, for CV00 the horizontal white lines in the remaining environments (1–2, 4–6) correspond to missing phenotypic information of the target genotypes in other environments. Similar volumes of information for model training are employed for predicting comparable testing set sizes.

CV00 (untested genotypes in unobserved environments), corresponds to the case of predicting new genotypes in novel environments. The conventional method for predicting untested genotypes in unobserved environments consist of discarding from the training set the phenotypic records of those genotypes in the target environment (testing set), then predict the performance of those genotypes in the novel environment using the phenotypic information available in the calibration environments. However, this procedure poses extra challenges as the number of common genotypes increases across environments. In our hypothetical example in Figure 2, all the genotypes were observed in all environments. Thus, if the phenotypic information for those genotypes in the target environment is discarded across environments, no phenotypic records will be available for model calibration. However, with the proposed scheme (bottom right panel in Figure 2) only the phenotypic information of those genotypes in the gray fold is deleted across environments. Then the information of the remaining folds (colors) in the other environments is used for model calibration when predicting the gray fold in environment 3. The same procedure is repeated for the remaining four folds (one at a time) in environment 3. This same procedure is repeated for the other environments (one at a time).

2.4. Assessment of Predictive Ability within and across Environments

For all cross-validation schemes, the predictive ability was assessed as the within environments correlation between predicted and observed values. The average correlation across environments was computed according to [39] for accounting for uncertainty and the sample size of the environments as

$$r_{\varphi} = \frac{\sum_{i=1}^{I} \frac{r_i}{V(r_i)}}{\sum_{i=1}^{I} \frac{1}{V(r_i)}}$$

where r_i is the Pearson's correlation between predicted and observed values at the i^{th} environment, $V(r_i) = \frac{1-r_i^2}{n_i-2}$ corresponds to the sampling variance and n_i is the number of observations at the i^{th} environment.

2.5. Variance Components

To assess the relative importance of the different model terms on each one of the prediction models, a full data analysis (i.e., non-missing values are considered) is conducted for computing the corresponding variance components. Then, percentage of explained variability by the z^{th} model term is computed as the ratio between the corresponding variance component and the sum of all the variance components of the model times 100, $\left(\frac{\sigma_z^2}{\sum_{z=1}^{t}\sigma_z^2} \times 100\right)$. The percentage of explained variability was computed for each model for each trait.

2.6. Modules

One of the principal objectives of this manuscript is to provide a template for implementing genomic prediction pipelines in an easy way. First, the different modules that are used for assembling the pipeline are presented. These modules work in a way that the outputs of these become the input of other modules in more advanced stages of the pipeline. The pipeline here presented considers all the different stages from the initial data sets until the prediction stage. Two examples of these pipelines are provided in the supplementary section.

The structure of the different modules is as follows: basically, all of these are comprised of three directories or folders "code", "input" and "output". The "code" folder contains the "mainCode.R" script which performs a specific routine depending on the module. The "input" folder contains a parameter file "parameters.R" where the inputs of the module are specified. During the implementation of the different modules, this is the only file that is modified according to the different conditions (parameters) to consider in the routines. The "output" folder is used for storing the results. Such that the outputs derived from these modules (routines) in previous stages will be used as the corresponding inputs in more advanced stages of the pipeline.

Initially, the phenotypic and genomic data are stored in a common repository/directory called "1.root.Data". Then the modules will refer to these data sets at the different stages of the pipeline. It is assumed that the genomic and phenotypic information (including missing values) are available for the same genotypes in both datasets. In our case, "Pheno.csv" and "SNPs.rda" correspond to the data sets containing the phenotypic and the molecular marker data. All of the modules used in our pipeline can be found in the 'modules' folder in the supplementary section. For assembling the pipeline at the different stages, these modules should be copied and renamed, and only the "parameter.R" file should be modified.

2.7. SplittingSNPs: Module for Applying Quality Control on Marker Data

The "SplitingSNPs" module performs the quality control based on missingness of the marker data by discarding molecular markers that exceed a given proportion of missing values (PMV). In the "parameters.R" file it should be specified the path (mm.file) to the marker data set "SNPs.rda", the highest proportion of missing values PMV (NaN.freq.i) to tolerate for including a marker SNP in the analysis. After applying the QC, the resulting set of maker SNPs ("X.csv") is stored in the "output" folder.

2.8. Gmatrix: Module for Constructing the Covariance Matrices Using Genomic and Environmental Factors

The "Gmatrix" module is used for constructing the relationship matrices between pairs of genotypes and between pairs of environments using the matrix of marker SNPs or the name (or ID) of the environments, respectively. In the "parameters.R" file, the path to the matrix of phenotypes (phenotype.file) and the path to the matrix of molecular markers (mm.file) should be specified for computing the kinship matrix using genomic data. If the value of "mm.file" is declared as "NULL" the covariance structure between genotypes or between environments is computed using the incidence matrices only. The value of the smallest allele frequency to tolerate for including markers in the analysis is specified with the "prop.MAF.j" option. The "colIDy" parameter indicates the column in the matrix of phenotypes that will be used to link the genotypes with marker data or phenotypes with environments. The outputs of this module will be store in the "output" folder and these are the resulting kinship matrix (G.rda), and its corresponding eigen value decomposition (EVD.rda) which are necessary to compute more elaborate model terms and for setting up the linear predictor, respectively.

2.9. Z: Module for Constructing Incidence Matrices for Genotypes and Environments

The "Z" module is used to include the main effects of genotypes or environments. In the "parameters.R" file in the "input" folder, the path to the matrix (phenotype.file) that contains the phenotypic information and the column (colIDy) that contains the information of the ID of the genotypes or environments should be specified. In the "output" folder the resulting incidence matrix "Z.rda" is stored as well as a graphical representation of the distribution of phenotypes across genotypes or environments (exp.des.pdf).

2.10. Imatrix: Module for Constructing the Interaction Matrix between Markers and Environments

The "Imatrix" module is used to compute the Hadamard product (or cell-by-cell product) between two covariance structures. The resulting matrix is needed for including the interaction between molecular markers and environments [25]. In the "parameters.R" file in the "input" folder, the path to the resulting matrices of the two factors (G1.file and G2.file) that will be considered in the interaction (G and E in our case) should be specified. The "output" folder will contain the resulting covariance matrix (G.rda) and its corresponding eigen value decomposition (EVD.rda).

2.11. Preparing.CV1.CV2: Module for Assigning Genotypes and Phenotypes to Folds

This module is used for assigning phenotypes/genotypes to training-testing sets for CV1 and CV2 cross-validation schemes. In the "parameters.R" file in the "input" folder, it should be specified the path to the matrix of phenotypes (phenotype.file), the number of folds to consider (folds) in the cross-validation, the column in the matrix of phenotypes (colIDy) that contains the names (IDs) of the genotypes, the type of cross-validation (either CV1 or CV2 or both). Also, it is possible to fix the seed value needed in the randomization and it varies between the replicates of the training-testing assignation. The resulting matrix (Y.csv) will be stored in the "output" folder. This matrix is identical to the initial matrix of phenotypes except that those column(s) containing the information of the fold assignation are added at the end of the matrix.

2.12. Preparing.CV0 and CV00 Module

Since one of the goals of this research is to provide a cross-validation scheme that preserves comparable sample sizes across different cross-validation schemes, the results from the cross-validation assignation CV1 and CV2 are used as inputs for the assignation of CV0 and CV00 schemes. For CV0 scheme, the resulting matrix from the previous module when conducting the CV2 assignation acts as input argument. In the "parameters.file" in the "input" folder, the path (phenotype.file) to the output file (Y.csv) derived from the "Preparing.CV2.CV1" module is specified, also the number of folds (it should be the same number of folds than in the previous output), the column that contains the phenotypic information (colPhen), and the column that contains the information of the folds (colFolds) for the CV2 scheme. The "output" folder will contain the resulting matrix of phenotypes. In this case, depending on the number of folds, the same number of extra columns are added masking as missing values those phenotypes belonging to the different folds at the different columns (one column for each fold). For the CV00 assignation, a similar procedure is performed but in this case, the column of the assignation of folds for CV1 scheme is used instead. Also, the resulting matrix of phenotypes is stored in the "output" folder.

2.13. Fitting.Models: Module for Performing the Predictions of the Missing Values and Compute the Variance Components

This module is used for fitting the models, perform the predictions of missing values and computing the variance components. In the "parameters.R" file in the "input" folder, the path to the matrix of phenotypes (phenotype.file) and the ID of the partitions (folds) to be predicted (e.g., ID of the folds [1, 2, 3, 4, and 5] or the ID of the environments [CV0 and CV00]) are specified. Also, since the BGLR [26,31] R package [40] was used for model fitting and it is based on the Bayesian framework, it is also necessary to specify the number of iterations (nIter) for the GIBBS sampler and the number of iterations to be used as burn-in (burnIn). Then, the linear predictor is built by providing the different models terms and their corresponding assumptions.

For this, a list is started "AB <- list()" to add the paths to the different model matrices that were created in previous stages. Such that the i^{th} element of the list corresponds to the i^{th} model term "AB[[i]] <-". Also, it is necessary to specify the type of the effects with "FIXED" for a fixed effect, "BRR" for a random main effect (RR-BLUP) or "RKHS" for the GBLUP model. In addition, it is necessary to provide the column numbers in the matrix of phenotypes that contain the ID (colVAR) of the genotypes, the phenotypic information (colPhen), the different training-testing partitions (colCV, folds or environments), and set/fix the seed for replicating exactly same the results "set.seed(i)" with the GIBBS sampler.

2.14. Pipeline

Each one of the different stages (2–6) of the pipeline is built using the modules stored in the repository folder (modules). For this, it is necessary to copy and rename these according to the different stages and only modify the "parameter.R" file in the "input" folder. The pipeline starts with the "1.root.Data" folder where the files with phenotypic (Phenos) and genomic information (SNPs) are stored. The next stage considers the implementation of quality control (QC) on the genomic data based on missing values and it corresponds to the "2.splitingSNPs" folder. Here, the path to the matrix of marker SNPs and the PMV should be provided; the resulting matrix "X.csv" is stored in the "output" folder.

The next stages consider the computation of the different model terms. The main and the interaction effects are stored in folders "3.Gmatrices" and "4.Imatrices", respectively.

In folder “3.Gmatrices”, the covariance matrices for “G” and “E” using the marker data and the ID of the environments, are computed. In addition, the incidence matrices that connects phenotypes with genotypes “ZL” and with environments “ZE” are also obtained. In the “4.Imatrices” folder, the outputs of the covariance matrices “G” and “E” are used to compute the interaction matrix “G×E”.

The 5th stage corresponds to the training-testing assignation, and it is divided in 3 sections. In the 5.1.CV2.CV1 section, for each replicate (1–10) the folds (5) are assigned for the cross-validations schemes CV1 (predicting new genotypes in observed environments) and CV2 (incomplete field trials). The resulting matrices are stored in the corresponding “output” folder. For the cross-validation CV0 and CV00, the configuration between them is very similar. Under CV0, in the folder “5.2.CV0” for each trait × replicate combination the training-testing assignation at each environment is performed by masking as missing values the corresponding observations derived from the CV2 scheme. While for CV00, the corresponding observations derived from CV1 scheme were considered.

The 6th stage corresponds to the prediction of missing values and it also comprises three sections. These correspond to the three different assignation schemes in the previous stage. Here, the “fitting.Models” module stored in the “modules” folder was implemented. For CV1 and CV2, for each trait × model × replicate combination, the prediction of the missing values is conducted, and the results are stored in the “output” folder according to the different folds (1–5). The model fitting for CV0 and CV00 schemes was performed for each trait × model × replicate × fold combination and the resulting predicted values are stored in the “output” folder.

Finally, the 7th stage considers the computation of the variance components. For this, the same module as in the previous state “fitting.Modules” was implemented for each trait × model combination by performing a full data analysis (i.e., no missing values were considered). Here, it is necessary to assign “−999” to the “folds” parameter in the “parameters.R” file. The resulting file “fm_full.R.Data” stored in the output folder contains the obtained variance components among other objects.

3. Results

Since one of the main objectives of this manuscript is to provide a template for implementing genomic prediction pipelines in an easy way, the obtained results are briefly described for both data sets. The supplementary materials section contains the full pipelines for data sets D1 and D2. The only difference between these two pipelines is that while for data set D1 the main effect of the environments was considered as fixed, for the second data set D2 it was treated as random. This was intended in this way to show the flexibility of the pipeline for considering different assumptions of the model terms.

3.1. D1: Sample of the USDA Collection

Percentage of variability explained by the different model terms.

Table 1 presents for each trait and model, the percentage of variability explained by each model term. For grain yield, with model M1 (E + L), the main effect of the environments (E) explains 55.7% of the total variability while the line effect (L) captures 25.2% and the residual term (R) 19.2%. The main effect of makers (G) introduced in M2 (E + L + G), captured 14.2% of the variability, and the residual variance (R) was increased to 25.5% compared with M1 (19.2%). The inclusion of the interaction between markers and environments (G×E) in M3, captured 9.0% of the total variability and the residual term (R) only 17.6%. Similar trends were observed for the remaining 8 traits. In general, for all traits, the model that includes the interaction between markers and environments (M3) returned the lowest residual variance. Also, as expected as the different model terms were added to the linear predictor, the variability explained by the environmental term (E) was reduced.

Table 1. Percentage of explained variability by each model term for all traits (9) using phenotypic and genomic data from the USDA soybean collection for 324 genotypes observed in 6 environments. Three models were considered, and these were constructed with the following components: E, main effect of environments; L, main effect of genotypes; G main effect of markers; and G×E for the interaction between genotypes and environments using marker data.

Trait	Model	E	L	G	G×E	R
Yield	M1: E + L	55.7	25.2			19.2
	M2: E + L + G	54.1	6.2	14.2		25.5
	M3: E + L + G + G×E	51.0	6.8	15.7	9.0	17.6
Height	M1: E + L	41.9	47.5			10.6
	M2: E + L + G	48.6	7.8	30.8		12.8
	M3: E + L + G + G×E	44.7	7.8	33.9	4.9	8.7
Lodging	M1: E + L	40.8	44.5			14.7
	M2: E + L + G	46.4	7.5	26.0		20.1
	M3: E + L + G + G×E	42.7	7.2	28.0	6.6	15.3
DaysToR8	M1: E + L	63.6	15.4			21.0
	M2: E + L + G	80.2	1.9	14.2		3.7
	M3: E + L + G + G×E	76.4	2.0	16.2	2.3	2.9
Oil	M1: E + L	39.6	42.5			17.9
	M2: E + L + G	49.5	6.2	20.3		24.0
	M3: E + L + G + G×E	46.6	7.0	21.9	10.8	13.7
Protein	M1: E + L	28.5	52.6			18.9
	M2: E + L + G	29.3	8.9	37.1		24.8
	M3: E + L + G + G×E	26.5	8.5	35.5	14.9	14.6
Seedweight	M1: E + L	36.0	55.8			8.2
	M2: E + L + G	49.4	6.5	29.2		14.9
	M3: E + L + G + G×E	46.1	6.6	32.8	5.4	9.1
Shaterly	M1: E + L	32.3	19.6			48.1
	M2: E + L + G	29.9	7.2	10.6		52.2
	M3: E + L + G + G×E	28.6	8.3	9.7	14.4	39.0
Stemtermscore	M1: E + L	40.2	48.0			11.8
	M2: E + L + G	49.0	6.1	28.8		16.1
	M3: E + L + G + G×E	45.3	6.0	31.6	5.5	11.6

Prediction Accuracy

Table 2 presents the mean (10 replicates) average correlation for 4 cross-validation schemes (CV2, CV1, CV0, and CV00) and 3 models (M1: E + L, M2: E + L + G, and M3: E + L + G + G×E). Under CV2, for grain yield, the models M1, M2, and M3 returned a mean average correlation of 0.576, 0.670, and 0.718, respectively. For CV1, the models M1-M3 returned a mean average correlation of −0.121, 0.635 and 0.662. While under CV0, the respective values for these three models were 0.114, 0.135 and 0.163; and for CV00, −0.010, 0.088 and 0.114. The predictive ability in CV2, CV1, CV0 and CV00 schemes was benefited when including the interaction between marker genotypes and environments with model M3. Similar trends were observed for the remaining traits.

3.2. D2: Sample of the SoyNAM

Percentage of variability explained by the different model terms.

Table 3 presents for each trait and model, the percentage of variability explained by each model term. For grain yield, under model M1 (E+L) the main effect of the environments (E) explained 68.0% of the total variability while the line effect (L) captured 7.8%, and the residual term (R) 24.2%. When the main effect of the markers (G) was included with M2, it captured 5% of the phenotypic variability and the residual term (R) 26.3%. The genotype by environment interaction (G×E) from M3 captured 7.2% of the variability and the residual term (R) addressed 20.4%. Also, for all traits the model that included the interaction term (M3) returned the lowest un-explained variability captured by the residual term (R). Similarly than with the previous data set (D1), as the different model terms were added the variability explained by the environmental term (E) was reduced.

Table 2. Average mean (10 replicates) correlation between predicted and observed values for four cross-validation scenarios, three models (M1: E + L, M2: E + L + G, and M3: E + L + G + G×E) and 9 traits from a sample of the USDA Soybean collection comprised of 324 genotypes observed in 6 environments (not all genotypes in all environments).

Trait	CV2			CV1			CV0			CV00		
	M1: E + L	M2: E + L + G	M3: E + L + G + G×E	M1: E + L	M2: E + L + G	M3: E + L + G + G×E	M1: E + L	M2: E + L + G	M3: E + L + G + G×E	M1: E + L	M2: E + L + G	M3: E + L + G + G×E
Yield	0.576	0.670	0.718	−0.121	0.635	0.662	0.114	0.135	0.163	−0.010	0.088	0.114
Height	0.835	0.835	0.862	−0.087	0.610	0.617	0.260	0.283	0.325	0.024	0.174	0.161
Lodging	0.771	0.808	0.812	−0.113	0.689	0.691	0.208	0.241	0.268	0.015	0.170	0.146
DysToR8	0.592	0.830	0.831	−0.090	0.674	0.676	0.048	0.126	0.142	0.007	0.095	0.102
Oil	0.788	0.832	0.834	−0.110	0.764	0.765	0.190	0.226	0.274	−0.009	0.186	0.188
Protein	0.726	0.757	0.766	−0.105	0.591	0.617	0.193	0.219	0.203	0.008	0.135	0.079
Seedweight	0.913	0.930	0.934	−0.089	0.860	0.861	0.329	0.356	0.464	0.005	0.311	0.353
Shaterly	0.674	0.690	0.649	−0.141	0.555	0.546	0.099	0.099	0.112	0.008	0.109	0.097
Stemtermscore	0.819	0.845	0.858	−0.070	0.710	0.720	0.287	0.315	0.351	0.020	0.233	0.193

Table 3. Percentage of explained variability by each model term for all traits (6) using phenotypic and genomic data from the SoyNAM experiment for 324 genotypes observed in 6 environments. Three models were considered, and these were constructed with the following components: E, main effect of environments; L, main effect of genotypes; G main effect of markers; and G×E for the interaction between genotypes and environments using marker data.

Trait	Model	E	L	G	G×E	R
Yield	M1: E + L	68.0	7.8			24.2
	M2: E + L + G	64.7	4.0	5.0		26.3
	M3: E + L + G + G×E	63.8	4.0	4.5	7.2	20.4
Moisture	M1: E + L	46.3	4.5			49.2
	M2: E + L + G	40.6	3.8	1.7		54.0
	M3: E + L + G + G×E	38.9	3.6	1.4	11.6	44.4
Protein	M1: E + L	42.9	27.5			29.5
	M2: E + L + G	37.4	9.6	20.4		32.6
	M3: E + L + G + G×E	35.3	10.6	20.2	8.5	25.5
Oil	M1: E + L	46.3	30.7			22.9
	M2: E + L + G	41.9	13.7	18.6		25.8
	M3: E + L + G + G×E	40.1	14.7	18.2	6.5	20.4
Fiber	M1: E + L	32.8	36.8			30.4
	M2: E + L + G	25.8	10.4	31.3		32.5
	M3: E + L + G + G×E	23.1	11.0	32.3	8.0	25.7
Size (seed)	M1: E + L	47.0	31.9			21.1
	M2: E + L + G	41.4	14.3	21.3		23.0
	M3: E + L + G + G×E	39.9	15.6	20.7	6.5	17.4

Prediction Accuracy

Table 4 presents the mean (10 replicates) average correlation for 4 cross-validation schemes (CV2, CV1, CV0, and CV00) and 3 models (M1: E + L, M2: E + L + G, and M3: E + L + G + G×E). Under CV2, for grain yield the models M1, M2 and M3 returned a mean average correlation of 0.342, 0.380 and 0.446, respectively. For CV1, the models M1–M3 returned a mean average correlation of −0.15, 0.296 and 0.373. While under CV0, the respective values for these three models were 0.197, 0.234 and 0.210; and for CV00, −0.014, 0.182 and 0.160. Also as expected, the predictive ability under the CV2 and CV1 schemes was slightly improved by including the interaction between marker genotypes and environments with model M3. However, for the remaining schemes (CV0 and CV00) the correlation between predicted and observed values was slightly reduced with the inclusion of the interaction effect (M3). Similar trends were observed for the remaining traits for all cross-validations schemes, showing only marginal improvements for oil content (0.647) under CV0.

Table 4. Average mean (10 replicates) correlation between predicted and observed values for four cross-validation scenarios, three models (M1: E + L, M2: E + L + G, and M3: E + L + G + G×E) and 6 traits from a sample of the SoyNAM experiment comprised of 324 genotypes observed in 6 environments (all genotypes in all environments).

Trait	**CV2**			**CV1**			**CV0**			**CV00**		
	M1: E + L	**M2: E + L + G**	**M3: E + L + G + G×E**	**M1: E + L**	**M2: E + L + G**	**M3: E + L + G + G×E**	**M1: E + L**	**M2: E + L + G**	**M3: E + L + G + G×E**	**M1: E + L**	**M2: E + L + G**	**M3: E + L + G + G×E**
Yield	0.342	0.380	0.446	−0.105	0.296	0.373	0.197	0.234	0.210	−0.014	0.182	0.160
Moisture	0.024	0.057	0.281	−0.099	0.076	0.305	0.033	0.031	0.025	0.003	−0.004	0.000
Protein	0.640	0.657	0.689	−0.086	0.474	0.515	0.580	0.614	0.597	−0.013	0.429	0.424
Oil	0.708	0.715	0.737	−0.058	0.469	0.495	0.631	0.631	0.647	−0.014	0.427	0.417
Fiber	0.686	0.698	0.717	−0.057	0.498	0.527	0.654	0.675	0.673	−0.020	0.466	0.463
Size	0.720	0.726	0.757	−0.072	0.432	0.469	0.644	0.673	0.665	−0.013	0.385	0.377

4. Discussion

4.1. Data Analysis

The main objective of this manuscript is to provide a template of a genomic prediction pipeline easy to implement, modify and adapt to other data sets. For this reason, the results derived from the implemented pipeline are briefly discussed focusing on the proposed implementation instead. These two data sets (D1: USDA soybean collection, and D2: SoyNAM) were already studied in several manuscripts [41,42]. The obtained results from our study (percentage of variability explained by the different model terms and prediction accuracy) are in line with the results of these studies.

In general, for both data sets (D1 and D2) and for all traits (15 = 9 + 6) the inclusion of the interaction term helped to reduce the percentage of variability explained by the environmental component. Also, it helped to decrease the non-explained variability addressed by the residual term. However, while the variability explained by the environmental term varied between 51% and 55.7% for grain yield in D1, for D2 it ranged between 63.8% and 68%, which represent around 10% of more variability captured by the environmental component in D2. A similar trend was observed for the residual term capturing a non-explained variability ranging between 19.2 and 17.6% for D1 and between 24.2% and 20.4% for D2. Thus, there was less unexplained variability in D1 which potentially contributed to deliver higher correlations between predicted and observed values for this data set compared with D2.

Regarding the predictive ability, across all models, traits and cross-validation schemes, different results were obtained in both data sets. Under CV2 and CV1 schemes, the correlation between predicted and observed values using molecular marker information (M2 and M3) was significantly higher for grain yield (0.670–0.718) for D1 than for D2 (0.380–0.446). While for protein and oil content the results were comparable in both data sets with these ranging between 0.757 and 0.832 for D1 and between 0.657 and 0.737 for D2. Under CV0 scheme, for D1 in 8 (except for protein content) out of the 9 traits, the M3 model outperformed the main effects models M2 while for D2 the M3 model was superior to M2 only for oil content. Regarding CV00, mixed results were observed; for D1 the M3 model slightly outperformed M2 in 4 (grain yield, days to maturity, oil content, and seed weight—100) out of the 9 traits while for D2 for all 6 traits M2 model outperformed M3. Perhaps the larger genetic diversity in D1 helped to increase the predictive ability of the genomic models (M2 and M3) with respect to D2 where all genotypes were observed in all environments.

4.2. Flexibility of the Pipeline

There are many implementations to conduct genomic prediction studies such as BGLR [26,31], rr-BLUP [30], asreml-R [27,28], sommer [29], BWGS [43], and bWGR [32]. However, to our knowledge, there are not available comprehensive examples of genomic prediction pipelines for conducting exhaustive studies while considering multiple factors. For example, providing plenty flexibility for selecting markers by applying quality control, different cross-validation schemes, model terms, model development (different types of effects and assumptions on these), etc. The pipeline here presented is based on a collection of modules that perform all the needed tasks that are required to conduct genomic prediction studies.

4.3. Potential Extensions of the Current Pipeline

The modules here described allow the construction of more elaborated pipelines in an easy way. For example, studies for finding the optimal quality control [44,45] can be performed by considering different combinations of percentage of missing values (PMV) and minor allele frequency (maf), different ways for composing training and testing sets [4,41], different features for sparse testing designs [46], study the distribution of the variance components by considering random sets of molecular markers [47], include weather data [38], leverage the information of correlated traits [48], and predict the performance of hybrid crosses using genomic inbred data [38] among other studies.

5. Conclusions

GS is a widely adopted method in plant and animal breeding programs, and conceptually its implementation is easy to follow. Initially, it requires a set of genomic and phenotypic data for model calibration for predicting the performance of candidate genotypes in target environments. However, in order to achieve the highest prediction accuracy between predicted and observed values, many factors should be assessed through cross-validation studies for a correct implementation of GS in real prediction problems.

For this reason, factors such as quality control on marker covariates, suitable cross-validation schemes mimicking real prediction scenarios, and the election of the prediction model among others should be evaluated. In most of the cases, the evaluation of these factors corresponds to minimal variations on the set of parameters to evaluate in the pipeline. Thus, there is no need to start from scratch the set of analyses when modifying the levels of the parameter(s) of interest(s). This allows the reuse of already developed code at different stages of the pipeline. Thus, it is easy to perform simple modifications in these codes to adapt them to particular cases.

In this study, we provide a set of modules that can be easily assembled to build complex prediction pipelines where the outputs of the early stages become the input of the more advanced ones. One feature of the proposed modules is that these can be used in a black box fashion where the specifics of the different analyses are controlled with a parameter file and there is no need to modify the main script (mainCode.R). With respect to the different cross-validation schemes, we provide a novel framework that allows similar sample sizes in calibration and prediction sets such that the results of the different prediction scenarios can be directly contrasted.

Finally, with respect to the obtained results in both data sets, we confirm again the advantages of considering the genotype-by-environment interaction in prediction models under the cross-validation schemes CV2 and CV1. Under the CV0 scheme, mixed results were observed for the first data set D1 while for D2 in most cases the main effects model was slightly superior. With CV00 scheme, no significant differences were observed for both data sets. We conclude, that using similar sample sizes in training sets the genotype-by-environment interaction can be leveraged when a portion of the data in the target environment has been observed via other genotypes while for the case of novel environments, there is a need of incorporating other sources of information such as soil and weather data to improve results.

Supplementary Materials: The Supplementary Material for this article can be found online at: https://uofnelincoln-my.sharepoint.com/:f:/g/personal/jhernandezjarquin2_unl_edu/EvvrOyt9ZhVCtHj1BwToUYIBMOtqT44QPTcDvVKS-RHLcQ?e=dBfDoL (accessed on 27 July 2021).

Author Contributions: R.P., Conceptualization, data collection, data analysis, elaboration of the first draft. M.G., Conceptualization, provided oversight of the study. D.J., Conceptualization, development of the modules, provided oversight of the study and participated in the conceptualization of the study. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by the Agriculture and Food Research Initiative Grant (NEB-21-176) from the USDA National Institute of Food and Agriculture, Plant Health and Production and Plant Products: Plant Breeding for Agricultural Production, A1211, Accession No. 1015252.

Data Availability Statement: The original contributions presented in the study are included in the article/Supplemental Material, further inquiries can be directed to the corresponding author.

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

References

1. Food and Agriculture Organization of the United Nations. *The Future of Food and Agriculture Trends and Challenges*; FAO: Rome, Italy, 2017; p. 180, ISSN 2522-722X.
2. Food and Agriculture Organization (FAO). *The Future of Food and Agriculture—Alternative Pathways to 2050*; Food and Agriculture Organization of the United Nations: Rome, Italy, 2018; p. 224.
3. Harris, J.; Spiegel, J. Food Systems Resilience: Concepts & Policy Approaches (Center for Agriculture and Food Systems). Available online: https://www.vermontlaw.edu/sites/default/files/2019-07/Food%20Systems%20Resilience_Concepts%20%26%20Policy%20Approaches.pdf) (accessed on 27 July 2021).
4. Widener, S.; Graef, G.; Lipka, A.E.; Jarquin, D. An Assessment of the Factors Influencing the Prediction Accuracy of Genomic Prediction Models across Multiple Environments. *Front. Genet.* **2021**, *12*, 689319. [CrossRef] [PubMed]
5. Bernardo, R. *Breeding for Quantitative Traits in Plants*; Stemma Press: Woodbury, MN, USA, 2002.
6. Breseghello, F.; Coelho, A. Traditional and Modern Plant Breeding Methods with Examples in Rice (Oryza sativa L.). *J. Agric. Food Chem.* **2013**, *61*, 8277–8286. [CrossRef] [PubMed]
7. Henderson, C.R. *Selection Index and Expected Genetic Advance. Statistical Genetics and Plant Breeding*; Hanson, W.D., Robinson, H.F., Eds.; National Academy of Sciences-National Research Council: Washington, DC, USA, 1963; pp. 141–163.
8. Henderson, C.R. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **1975**, *31*, 423. [CrossRef]
9. Henderson, C.R. *Applications of Linear Models in Animal Breeding*; University of Guelph: Guelph, ON, Canada, 1984.
10. Beaulieu, J.; Doerksen, T.K.; MacKay, J.; Rainville, A.; Bousquet, J. Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC Genom.* **2014**, *15*, 1048. [CrossRef]
11. de los Campos, G.; Hickey, J.M.; Pong-Wong, R.; Daetwyler, H.D.; Calus, M.P.L. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **2013**, *193*, 327–345. [CrossRef] [PubMed]
12. Fernando, R.L.; Grossman, M. Marker assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.* **1989**, *21*, 467. [CrossRef]
13. Soller, M.; Plotkin-Hazan, J. The use marker alleles for the introgression of linked quantitative alleles. *Theor. Appl. Genet.* **1977**, *51*, 133–137. [CrossRef]
14. Soller, M. The use of loci associated with quantitative effects in dairy cattle improvement. *Anim. Sci.* **1978**, *27*, 133–139. [CrossRef]
15. Bernardo, R. Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. *Crop Sci.* **2008**, *48*, 1649–1664. [CrossRef]
16. Bernardo, R. Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. *Crop Sci.* **1994**, *34*, 20–25. [CrossRef]
17. Meuwissen, T.H.E.; Hayes, B.; Goddard, M. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **2001**, *157*, 1819–1829. [CrossRef] [PubMed]
18. Malosetti, M.; Bustos-Korts, D.; Boer, M.P.; Van Eeuwijk, F.A. Predicting Responses in Multiple Environments: Issues in Relation to Genotype × Environment Interactions. *Crop Sci.* **2016**, *56*, 2210–2222. [CrossRef]
19. Crossa, J.; de Los Campos, G.; Pérez, P.; Gianola, D.; Burgueño, J.; Araus, J.L.; Makumbi, D.; Singh, R.P.; Dreisigacker, S.; Yan, J.; et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **2010**, *186*, 713–724. [CrossRef] [PubMed]
20. Heslot, N.; Yang, H.-P.; Sorrells, M.E.; Jannink, J.-L. Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Sci.* **2012**, *52*, 146–160. [CrossRef]
21. Piepho, H.P. Ridge Regression and Extensions for Genomewide Selection in Maize. *Crop Sci.* **2009**, *49*, 1165–1176. [CrossRef]
22. Crossa, J.; Pérez-Elizalde, S.; Jarquin, D.; Cotes, J.M.; Viele, K.; Liu, G.; Cornelius, P.L. Bayesian Estimation of the Additive Main Effects and Multiplicative Interaction Model. *Crop Sci.* **2011**, *51*, 1458–1469. [CrossRef]

23. Burgueño, J.; Campos, G.D.L.; Weigel, K.; Crossa, J. Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. *Crop Sci.* **2012**, *52*, 707–719. [CrossRef]
24. Schulz-Streeck, T.; O Ogutu, J.; Gordillo, A.; Karaman, Z.; Knaak, C.; Piepho, H.-P. Genomic selection allowing for marker-by-environment interaction. *Plant Breed.* **2013**, *132*, 532–538. [CrossRef]
25. Jarquín, D.; Crossa, J.; Lacaze, X.; Du Cheyron, P.; Daucourt, J.; Lorgeou, J.; Piraux, F.; Guerreiro, L.; Pérez-Rodríguez, P.; Calus, M.; et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* **2014**, *127*, 595–607. [CrossRef]
26. de los Campos, G.; Pérez-Rodríguez, P. *BGLR: Bayesian Generalized Linear Regression, R package Version 1(3)*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
27. Butler, D.; Cullis, B.; Gilmour, A.; Gogel, B.J. *ASReml-R Reference Manual, Version 3. Training and Development Series, No. QE02001*; Queensland Department of Primary Industries: Queensland, Australia, 2009.
28. Butler, D.G.; Cullis, B.R.; Gilmour, A.R.; Thompson, R. *ASReml-R Reference Manual, Version 4*; University of Wollongong: Wollongong, Australia, 2018. Available online: https://mmade.org/wp-content/uploads/2019/01/asremlRMfinal.pdf (accessed on 23 September 2011).
29. Covarrubias-Pazaran, G. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLoS ONE* **2016**, *11*, e0156744. [CrossRef]
30. Endelman, J.B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* **2011**, *4*. [CrossRef]
31. Pérez-Rodríguez, P.; de los Campos, G. Genome-wide regression and prediction with the BGLR statistical pack-age. *Genetics* **2014**, *198*, 483–495. [CrossRef] [PubMed]
32. Xavier, A.; Muir, W.M.; Rainey, K.M. bWGR: Bayesian whole-genome regression. *Bioinformatics* **2019**, *36*, 1957–1959. [CrossRef]
33. Bandillo, N.; Jarquin, D.; Song, Q.; Nelson, R.L.; Cregan, P.; Specht, J.; Lorenz, A. A Population Structure and Ge-Nome-Wide Association Analysis on the USDA Soybean Germplasm Collection. *Plant Genome* **2015**, *8*, 2015. [CrossRef]
34. Diers, B.W.; Specht, J.; Rainey, K.M.; Cregan, P.; Song, Q.; Ramasubramanian, V.; Graef, G.; Nelson, R.; Schapaugh, W.; Wang, D.; et al. Genetic architecture of soybean yield and agro-nomic traits. *G3 Genes Genomes Genet.* **2018**, *8*, 3367–3375.
35. Xavier, A.; Jarquin, D.; Howard, R.; Ramasubramanian, V.; Specht, J.E.; Graef, G.L.; Beavis, W.D.; Diers, B.W.; Song, Q.; Cregan, P.B.; et al. Genome-Wide Analysis of Grain Yield Stability and Environmental Interactions in a Multiparental Soybean Population. *G3 Genes Genomes Genet.* **2018**, *8*, 519–529. [CrossRef]
36. Habier, D.; Fernando, R.L.; Dekkers, J.C.M. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics* **2007**, *177*, 2389–2397. [CrossRef]
37. VanRaden, P.M. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* **2008**, *91*, 4414–4423. [CrossRef] [PubMed]
38. Jarquin, D.; de Leon, N.; Romay, C.; Bohn, M.; Buckler, E.S.; Ciampitti, I.; Edwards, J.; Ertl, D.; Flint-Garcia, S.; Gore, M.A.; et al. Utility of Climatic Information via Combining Ability Models to Improve Genomic Prediction for Yield within the Genomes to Fields Maize Project. *Front. Genet.* **2021**, *11*, 1819. [CrossRef]
39. Tiezzi, F.; de Los Campos, G.; Gaddis, K.P.; Maltecca, C. Genotype by environment (climate) interaction improves genomic prediction for production traits in us holstein cattle. *J. Dairy Sci.* **2017**, *100*, 2042–2056. [CrossRef] [PubMed]
40. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019; Available online: https://www.R-project.org/ (accessed on 27 September 2021).
41. Jarquin, D.; Specht, J.; Lorenz, A. Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions. *G3 Genes Genomes Genet.* **2016**, *6*, 2329–2341. [CrossRef]
42. Persa, R.; Hiroyoshi, I.; Jarquin, D. Use of family structure information in interaction with environments for leveraging genomic prediction models. *Crop J.* **2020**, *8*, 843–854. [CrossRef]
43. Charmet, G.; Tran, L.-G.; Auzanneau, J.; Rincent, R.; Bouchet, S. BWGS: A R package for genomic selection and its application to a wheat breeding programme. *PLoS ONE* **2020**, *15*, e0222733. [CrossRef]
44. Jarquin, 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]
45. Jarquín, D.; Howard, R.; Graef, G.; Lorenz, A. Response Surface Analysis of Genomic Prediction Accuracy Values Using Quality Control Covariates in Soybean. *Evol. Bioinform.* **2019**, *15*, 1176934319831307. [CrossRef] [PubMed]
46. Jarquin, D.; Howard, R.; Crossa, J.; Beyene, Y.; Gowda, M.; Martini, J.W.R.; Pazaran, G.C.; Burgueño, J.; Pacheco, A.; Grondona, M.; et al. Genomic Prediction Enhanced Sparse Testing for Multi-environment Trials. *G3 Genes Genomes Genet.* **2020**, *10*, 2725–2739. [CrossRef]
47. Gage, J.L.; Jarquin, D.; Romay, C.; Lorenz, A.; Buckler, E.S.; Kaeppler, S.; Alkhalifah, N.; Bohn, M.; Campbell, D.; Edwards, J.; et al. The effect of artificial selection on phenotypic plasticity in maize. *Nat. Commun.* **2017**, *8*, 1–11. [CrossRef]
48. Jarquin, D.; Howard, R.; Xavier, A.; Das Choudhury, S. Increasing Predictive Ability by Modeling Interactions between Environments, Genotype and Canopy Coverage Image Data for Soybeans. *Agronomy* **2018**, *8*, 51. [CrossRef]