Strategies to increase prediction accuracy in genomic selection of complex traits in Alfalfa (*Medicago sativa* L.)

Authors:

Cesar A. Medina[1], Harpreet Kaur[2], Ian Ray[2], *Long-Xi Yu[1]

[1] United States Department of Agriculture-Agricultural Research Service, Plant Germplasm Introduction and Testing Research, Prosser, WA, USA. [2] Department of Plant and Environmental Sciences, New Mexico State University, Las Cruces, NM, USA.

* Correspondence author: longxi.yu@usda.gov

Abstract

Agronomic traits such as biomass yield and abiotic stress tolerance are genetically complex and challenging to improve by conventional breeding strategies. Genomic selection (GS) is an alternative approach in which genome-wide markers are used to determine the genomic estimated breeding value (GEBV) of individuals in a population. In alfalfa, previous results indicated that low to moderate prediction accuracy values (<70%) were obtained in complex traits such as yield and abiotic stress resistance. There is a need to increase the prediction value in order to employ GS in breeding programs. In this paper we reviewed different statistic models and their applications in polyploid crops including alfalfa. Specifically, we used empirical data affiliated with alfalfa yield under salt stress to investigate approaches which use DNA marker importance values derived from machine learning models, and genome-wide association studies (GWAS) of marker-trait association scores based on different GWASpoly models, in weighted GBLUP analyses. This approach increased prediction accuracies from 50% to more than 80% for alfalfa yield under salt stress. This is the first report in alfalfa to use variable importance and GWAS-assisted approaches to increase the prediction accuracy of GS, thus helping to select superior alfalfa lines based on their GEBVs.

Key words: Genomic selection, WGBLUP, *Medicago sativa*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

1. Introduction

Alfalfa is an autotetraploid ($2n = 4x = 32$) perennial forage crop with a genome size of 800−1,000 Mb [1]. However, alfalfa breeding is complicated by its high heterozygosity, polysomic inheritance, and out-crossing nature, which hinder creation of inbred lines. Alfalfa breeding goals target improvement of forage yield, quality, and tolerance to biotic and abiotic stresses. This process requires selection of perennial plants that can maintain biomass productivity and quality over several years. Therefore, traits must be evaluated over multiple harvests each year for several years. Consequently, genetic gain is slower compared to annual crops. In addition, alfalfa breeding programs have largely focused on recurrent phenotypic selection (PS) in field environments to improve quantitative traits of interest. However, this approach is constrained by breeding population size, genotype × environment interactions, or low heritability of the trait, thus hindering the development of superior varieties.

One promising alternative to recurrent PS is indirect selection based on the use of molecular markers generated, for example, via genotyping by sequencing (GBS) [2]. Markers closely linked to quantitative trait loci (QTL) can then be used for marker-assisted selection (MAS) in breeding programs. Initially, QTLs are detected through genetic mapping or genome-wide association studies (GWAS), where marker-trait associations that exceed specific thresholds are declared statistically significant (Figure 1a). However, MAS is primarily effective for traits controlled by relatively few genes with large effects. For complex traits (e.g., stress tolerance or yield) in elite populations it can be difficult to clearly identify QTL with major effect because the trait is often controlled by multiple loci possessing small effect. To overcome this challenge, genomic selection (GS) is a promising alternative to determine the genetic potential or breeding value of an individual based on whole-genome markers (Figure 1a).

Genomic selection offers the potential to shorten alfalfa breeding and selection cycles. This method follows the infinitesimal model, which assumes that a quantitative trait is determined by an infinite number of unlinked and non-epistatic loci, each one with a very small effect, that satisfy normality and linearity [3]. This technique uses both parametric and non-parametric statistical models to determine associations of phenotypic trait values with genome-wide

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

molecular markers. This information is subsequently used to predict future breeding values (i.e., genomic-estimated breeding values, GEBVs) for each individual in a population based on their genome-wide marker profile/genotype [4]. Hence, rapid marker-based selection cycles can replace some time-intensive phenotypic selection cycles to accelerate genetic gain. In this paper, we review different GS models and their application to perennial polyploid crops. We also demonstrate the implementation of GS models on a real dataset of alfalfa to identify improved approaches to implement GS in alfalfa breeding programs.

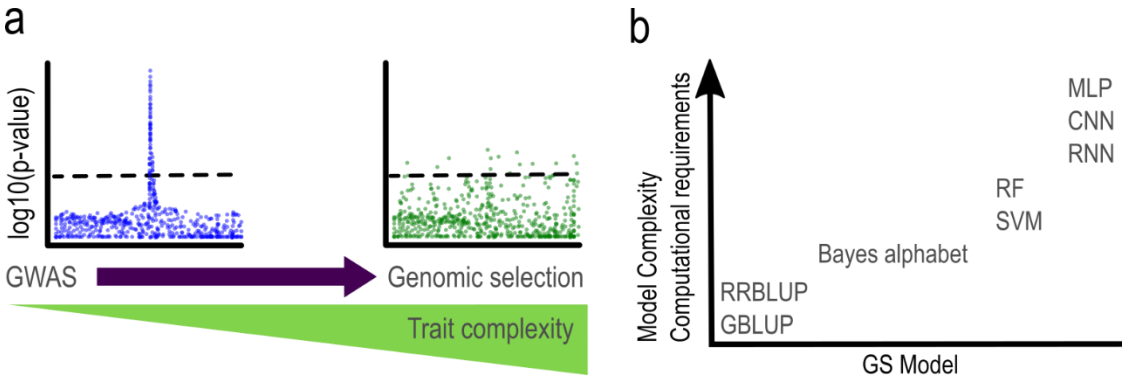

Figure 1. Indirect selection based on molecular markers. a) Generalized Manhattan plots illustrating comparison of GWAS effectiveness in simple (left) vs complex traits (right). Note: Bold dashed line indicates minimum threshold to select significant markers. A significant signal (i.e., QTL) was identified in the simple trait (left panel), while no defined QTL was identified for the complex trait. Therefore, genomic selection (GS) is more appropriate and practical for complex traits. b) Common parametric and non-parametric models used in GS and their computational requirements. GBLUP, genomic best linear unbiased prediction; RRBLUP, ridge-regression BLUP; RF, random forest; SVM, support vector machine; MLP, multilayer perceptron; CNN, convolutional neural network; RNN, recurrent neural network.

2. Statistical methods in GS

There is a wide repertoire of parametric and non-parametric models to obtain GEBVs which differ in complexity, accuracy, and computational requirements (Figure 1b). Some phenotypic traits are highly complex and more difficult to predict using their genetic information, therefore, accuracy in GS modeling is a cornerstone. Model accuracy metric is calculated as the Pearson's

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

correlation coefficient ($r_{GEBV:y}$) between GEBVs in a training population and observed phenotypes from testing population. Determining a GEBV can be solved as a regression:

$$y = \mathbf{X}\beta + e \text{ [1]}$$

where y is a vector ($n \times 1$) of phenotypic outcomes in n observations, $\mathbf{X}$ is a matrix ($n \times p$) with p number of markers or predictors in n observations, β is the vector ($p \times 1$) of marker effects and e is a vector of residual effects. In GS, however, molecular markers or predictors (p) are greater than observations (n), generating a large p small n problem ($p \gg n$). Therefore, estimation of marker effects via multiple regression by ordinary least squares is not possible. To resolve this issue, multiple methods have been developed to handle the high dimensionality of the genomic data. Shrinkage models such as best linear unbiased prediction using ridge-regression (RRBLUP) [5] or genomic best linear unbiased prediction (GBLUP) [6] are most popular models used in GS. Both of these models assume that the effect of all single nucleotide polymorphism (SNP) markers is normally distributed with equal variance [5]. Bayesian and least absolute shrinkage and selection operator (LASSO) models assume that some SNPs have large or moderate effects, and others have small or null effects [7]. Finally, machine learning (ML) models like random forest (RF), support vector machine (SVM) and deep learning (DL) algorithms do not assume linearity in the model. ML models can use nonlinear kernels to capture complex SNP-SNP interactions and nonlinear relationships.

2.1 Ridge-regression best linear unbiased prediction (RRBLUP)

The RRBLUP is a shrinkage method to obtains GEBVs by incorporating genomic information into BLUP using ridge regression (RR). This model has been widely implemented by development of the R package RRBLUP [5]. Prediction equations used by RRBLUP assume *a priori* that all loci explain equal amounts of the genetic variation. The core of the RRBLUP package is the function, `mixed.solve`, which solves any mixed model of the form:

$$y = \mathbf{X}\beta + \mathbf{Z}u + e \text{ [2]}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

where y is a vector $(n \times 1)$ of phenotypic outcomes in n observations, $\mathbf{X}$ is a matrix $(n \times p)$ with p number of markers or predictors in n observations, β is a vector $(p \times 1)$ of fixed effects, $u \sim N(0, \mathbf{K}\sigma_u^2)$ is a vector $(n \times 1)$ of random effects distributed normally with mean zero and variance σ_u^2 and $\mathbf{K}$ is a positive semidefinite matrix, $\mathbf{Z}$ is a design matrix $(n \times p)$ for the random effects, and $e \sim N(0, \mathbf{I}\sigma_e^2)$ is a vector $(n \times 1)$ of residual effects distributed normally with mean zero and variance σ_e^2 and $\mathbf{I}$ is the identity matrix.

2.2 Genomic best linear unbiased prediction (GBLUP)

GBLUP measure the relationship between individuals with the aid of marker data. The difference with RRBLUP is the use of marker-based relationship matrix named genomic relationship matrix (GRM) or **G** matrix [8]. **G** matrix defines the covariance between known relatives in a population based on DNA marker information. The mixed model for GBLUP analysis uses the following formula:

$$y = 1\mu + \mathbf{X}\beta + \mathbf{Z}g + e \quad [3]$$

where y, $\mathbf{X}$, β and e were defined in equation 2, μ is the overall mean, $g \sim N(0, \mathbf{G}\sigma_g^2)$ is a vector $(n \times 1)$ of random effects distributed normally with mean zero and variance σ_g^2 and $\mathbf{G}$ is the $\mathbf{G}$ matrix which can be obtained according to the approach of VanRaden [9]:

$$\boldsymbol{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\Sigma p_i(1 - p_i)} \quad [4]$$

where $\mathbf{Z}$ is an identity matrix for the markers and p_i are the observed minimum allele frequency (MAF) of all individuals genotyped. However, Yang et al. (2010) combined the information on all SNPs (*i*) coded as 0=AA, 1=BB, 2=AB according to alternative allele dosage to calculate the relationship between individuals *j* and *k* into a GRM ($\boldsymbol{G}_{ijk}$) using a weighting scheme based on allele frequencies:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

$$\boldsymbol{G}_{jk} = \frac{1}{N}\sum_{i}\boldsymbol{G}_{ijk} = \begin{cases} \frac{1}{N}\sum_{i}\frac{(w_{ij}-2p_i)(w_{ik}-2p_i)}{2p_i(1-p_i)}, j \neq k \\ 1+\frac{1}{N}\sum_{i}\frac{w_{ij}^2-(1+2p_i)w_{ij}+2p_i^2}{2p_i(1-p_i)}, j = k \end{cases} \quad [5]$$

where $\boldsymbol{G}_{jk}$ is the G matrix averaged over all SNP positions in the genome, N is the number of markers, w_{ij} is the element of $\boldsymbol{W}$ pertaining to marker i and individual j, w_{ik} is the element of $\boldsymbol{W}$ pertaining to marker i and individual k. The $\boldsymbol{G}_{ijk}$ or GRM produces the off-diagonal ($j \neq k$) and diagonal ($j = k$) elements [10]. Based on this approach, Slater et al. (2016) proposed a full autotetraploid model to obtain the $\boldsymbol{G}$ matrix:

$$\boldsymbol{G}_{jk} = \begin{cases} 1+\frac{1}{M}\sum_{i}\frac{(w_{ij}-p_i)(w_{ik}-p_i)}{p_i(1-p_i)}, j \neq k \\ 1+\frac{1}{M}\sum_{i}\frac{w_{ij}^2-2p_iw_{ij}+p_i^2}{p_i(1-p_i)}, j = k \end{cases} \quad [6]$$

where M is the number of markers × 5 and p_i is the frequency of each genotype. The genomic relationship matrices described are based on identity-by-state and simply measure the similarity of alleles between individuals [11].

Analysis results from RRBLUP and GBLUP can be similar; however, GBLUP is more computationally efficient than RRBLUP. GBLUP require a G matrix of dimensions $n \times n$ (where n is the number of individuals in the population) whereas RRBLUP requires a genotypic matrix $n \times m$ (where m is the number of markers) with high dimensionality. In summary, GBLUP does not provide marker effects but is more time/memory efficient than RRBLUP.

2.2.1 Weighted Genomic best linear unbiased prediction (WGBLUP)

The GBLUP method usually assumes that all SNPs explain the same fraction of genetic variance. However, traits are affected by different genetic architectures which are associated with SNPs that possess varying effects (e.g., major SNPs). To account for varying effects of different SNP

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

alleles, the weighted GBLUP (WGBLUP) method was developed to incorporate unequal weights for all SNPs [12]. The $\boldsymbol{G}^*$ matrix is constructed as follows:

$$\boldsymbol{G}^* = \frac{\mathbf{ZDZ}'}{2\Sigma p_i(1-p_i)} \quad [7]$$

where the asterisk symbol ($\boldsymbol{G}^*$) is used to differentiate weighted $\boldsymbol{G}$ matrix from regular $\boldsymbol{G}$ matrix, $\mathbf{Z}$ is an identity matrix for the markers, $\mathbf{D}$ is a diagonal matrix, where each element of the diagonal corresponds to SNP weights, and p_i are the observed minimum allele frequency (MAF) of all genotyped individuals. To obtain the $\mathbf{D}$ matrix, each element of this matrix is defined as:

$$\mathbf{D} = \begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{pmatrix}$$

where w are weights based on SNP effects from different methods. SNP weights can be obtained from Bayesian Regressions with good results [13]. Other methods to determine weights for SNPs include prioritization based on Wright's fixation scores (F_{st}) [14]. The F_{st} score measures the level of genetic differentiation between populations based on change in allele frequencies. When F_{st} scores were used to compute relative weights, prediction accuracy increased up to 5%. Ren et al. (2021) developed several methods to optimize WGBLUP by generation of different weighted G matrices. They noted that the choice of an optimal GBLUP matrix will depend on the number of loci controlling the trait. Results indicated that estimated marker-variance-weighted (EVW)-GBLUP was superior for traits controlled by loci of a large effect, and absolute value of the estimated marker-effect-weighted (AEW)-GBLUP was better for traits controlled by loci with moderate effect [15].

2.3 Bayesian models

Bayesian models applied to GS that do not assume normal distribution of marker effects. Instead, they assume that few markers will have large effects on the trait, allowing markers to have different effects and variances. Bayesian models impose stronger shrinkage towards zero on

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

small SNP effects and less shrinkage on relatively large SNP effects. The BGLR R package implements a large collection of Bayesian models [16]. The Bayesian models for continuous variables are represented by the equation:

$$y_i = 1\mu + \sum_{j=1}^{m} x_{ij}\,\beta_j + e_i \text{ [8]}$$

Where y_i is the vector of adjusted phenotypic observations $\{y_1, \ldots, y_n\}$, μ is the overall mean for the trait, m is the number of markers or SNPs, x_{ij} is the i^{th} genotype for j^{th} SNP, β_j is a vector for the effect of the j^{th} SNP, and e_i is a vector of residual effects with assumed normal distribution $e \sim N(0, I\sigma_e^2)$, where σ_e^2 is the residual variance and I is the identity matrix. Bayes A, B, Cπ and Bayesian LASSO (BL) are the most common models used. All models assume different prior distributions for SNP effects (Table 1).

Table 1. Different prior distributions for Bayesian models.

Model	Prior distribution‡	Ref
Bayes A	$\beta_j \sim t(df_\beta, S_\beta)$	[4]
Bayes B	$\beta_j = \begin{cases} 1/2\,\gamma\lambda \exp(-\lambda\lvert\beta_j\rvert) & for\ \beta_j \neq 0 \\ (1-\gamma) & for\ \beta_j = 0 \end{cases}$	[17]
Bayes Cπ	$\beta_j\lvert\pi, \sigma^2_{\beta_j} \begin{cases} \beta_j \sim 0 & with\ prob\ \pi \\ \beta_j \sim N\left(0, \sigma^2_{\beta_j}\right) & with\ prob\ (1-\pi) \end{cases}$	[18]
Bayesian LASSO	$\beta_j \sim DE(\lambda^2, \sigma_e^2)$	[19]

‡; t, scaled-t distribution; df_β, degree of freedom; S_β, scale parameters; γ, fraction of the SNPs that are in linkage disequilibrium (LD) with a quantitative trait locus (QTL); β_j, is the additive effect of the j^{th} SNP; π; proportion of markers with large effects; λ, parameter of exponential distribution; π, probability of the marker effect equal to zero; DE, double exponential.

2.4 Machine learning models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

Machine learning is a field that involves the application of computer algorithms and statistical models to interpret and predict large datasets. Algorithmic modeling is a rapidly developing discipline with strong potential to provide accurate and informative analyses or predictions using large and complex data sets [20]. These models are widely used to solve problems across different disciplines such as medicine, genomics, natural language processing, and stock market forecasting. Compared with classical statistical models, ML models have fewer assumptions about normality and distribution of data. One important remark is that ML models are being developed much faster than their interpretability, developing a new field to be explored. The most common problem with ML algorithms is data overfitting which results in models that poorly predict the behavior of future data. To avoid this problem, it is necessary to use a robust validation method, such as cross-validation, which provides an indication of performance on new data. SVM and RF are the most common ML models for classification and regression in GS [21,22].

2.4.1 Support Vector Machine (SVM)

SVM is a machine learning algorithm used in classification or regression problems [23]. The objective of SVM is to find the best hyperplane with the maximal margin in a $n-$dimensional space (genotypic matrix) with respect to a given collection of data (phenotypic values) and predict the correct classification/regression of unseen examples. Support vector regression (SVR) is an application of the SVM. In SVR, each n-dimensional input vector ($\mathbf{x}_i$) of p SNP markers is associated with a y_i as response variable (e.g., yield), where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Then, linear regression $f(\mathbf{x})$ is performed using the following equation [24]:

$$f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b \text{ [9]}$$

where, $\mathbf{w}$ is a vector of unknown weights (i.e., regression coefficients) and b is the bias. The training data is used to learn $\mathbf{w}$. The coefficients $\mathbf{w}$ and b are estimated by minimizing the following regularized loss function $R(C)$:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

$$R(C) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} L_\varepsilon\,(y_i - f(\mathbf{x}_i)) \text{ [10]}$$

where $\|\mathbf{w}\|^2 = \mathbf{w}'\mathbf{w}$, represents model complexity, C is a positive cost parameter specified by the user. C determines the trade-off between model complexity and training error, $y_i - f(\mathbf{x}_i)$ is the error associated with i^{th} training data point and $L\varepsilon$ is the empirical error measured by ε-intensive loss-function:

$$L_\varepsilon(y_i - f(\mathbf{x}_i)) = \begin{cases} 0\ if\ |y_i - f(\mathbf{x}_i)| < \varepsilon \\ |y_i - f(\mathbf{x}_i)| - \varepsilon \text{ otherwise} \end{cases} \text{ [11]}$$

where the loss function is zero ("insensitive") for any absolute error smaller than a predefined value ε. For an error value larger than ε, the loss function is the difference between the absolute error and ε. ε-SVR solutions is sparseness, with a fraction of errors are equal to zero and thereby vanishing in the final model $f(\mathbf{x})$ and only absolute errors > 0 are relevant and used as "support vectors". $f(\mathbf{x})$ can be assumed as linear or non-linear. For non-linear functions, the data can be mapped into a higher dimensionality space using a kernel space. In nonlinear SVR modeling different kernels can be used to increase the predictive power of the model. The kernel function provides a solution to classification/regression dataset by adding an additional dimension to the data. Different kernel functions can be selected to transform input data to feature space. Commonly used kernels in SVM include linear, polynomial, radial basis function (RBF) and sigmoidal kernels (Table 2). In high-dimensional data (i.e: microarrays or GS), the RBF kernel is preferred [24].

Table 2. Kernels used in support vector machine (SVM) model. Meta-parameters used for tuning include gamma (γ), degree of polynomial (d) and intercept (α).

Kernel	**Formula**‡
Linear	$K(x_i, y_j) = x_i^T y_j$
Polynomial	$K(x_i, y_j) = \gamma(x_i^T y_j + \alpha)^d$
Radial basis function	$K(x_i, y_j) = e^{-\gamma\|x_i - y_j\|^2}$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

Sigmoidal	$K(x_i, y_j) = \tan h\ (\gamma x_i^T y_j + \alpha)$

‡; x_i, y_j are two vectors in the n-dimensional space.

2.4.2 Random Forest (RF)

The RF method is a machine learning model for classification and regression problems based on the identification of an objective function and its optimization [25]. The objective function measures the distance between the RF output and desired scores to modify internal parameters to reduce this error. Random forests consist of numerous independent decision trees that are independently trained using a random subset of data. The final prediction is calculated as the average values over all the trees. RF model attempt to reduce the computational cost to train the model, capture complex interactions and reduce the over-fitting risk in the data [22]. In each decision tree multiple binary filters are applied creating bifurcations generating branch and a treelike structure. Every point where the samples are filtered is called a decision node. Optimization in RF consists of determining the best way to split samples at decision nodes based on the predictors. RF can optimize four different hyperparameters to increase the predictive power of the model: total number of observations (N), total number of predictors (M), subset of predictors chosen for determining a decision tree (*mtry*) and total number of decision trees to generate the RF (*ntree*). Subsequently, RF creates a series of filters based on the predictor variables. Gini impurity score (as described below) and mean squared error are used to select the best variables in decision nodes.

The RF approach can provide accurate predictions with complex genomic datasets. A very useful feature in the RF model is a function designated, variable importance metrics, which ranks each SNP's impact according to the trait. Two approaches to compute variable importance include mean prediction accuracy decrease when a variable/marker is removed, and the mean decrease in impurity (or Gini importance). Gini impurity measures how well a potential split is able to separate the samples of two classes at a particular node. Important limitations when using RF algorithms for GS involve slower processing time when a large number of trees are chosen for the model, or the number of SNP markers is too high. For RF analysis, we observed that a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

processing limit was reached when the genotypic matrix was composed of more than 10,000 markers.

2.4.3 Deep learning (DL)

Deep learning is a subfield of machine learning with great success in natural language processing, image recognition or virtual assistance [26]. The DL architecture uses several layers of nonlinear processing units called hidden layers. Hidden layers allow the network to capture higher order interactions from the data. This method uses artificial neural network architecture where the perceptron is the fundamental unit for comparison as a neuron in a biological neural network. The implementation of DL in genomic selection is recent and some studies have reported a modest increase in prediction accuracy in comparison with parametric and non-parametric models [27–30]. In theory, deep learning could perform better for traits with large epistatic effects and low narrow-sense heritability, a concept which is reinforced the high predictive ability of mixed models as prediction machinery [31]. Numerically encoded SNPs are the inputs to the first layer to produce a centered vector of phenotypes. The most common DL models used in GS are multilayer perceptron neural network (MLP), convolutional neural network (CNN) and recurrent neural networks (RNN).

The formal description of MLP is a feed-forward DL model composed of multiple perceptron ordered in hidden layers in a directed graph. In MLP each layer is fully connected with the next one by nonlinear activation functions, such as rectified linear unit activation function (ReLU), to minimize the mean square error. MLP is flexible because no assumption is made about the joint distribution of inputs and outputs. CNN is a special case of a neural network which uses convolution instead of a full matrix multiplication in the hidden layers. The convolution is a function that can be defined as an “integral transform” to reduce the number of hyperparameters to be estimated. CNN was proposed to accommodate situations where input variables are distributed along a space pattern resembling a SNP matrix. CNN seems to perform best in GS because it can detect patterns in the genotypic matrix discovering correlations between adjacent SNPs. In addition, CNN appears to perform better when epistatic components are important and the narrow-sense heritability is low [30]. It is important to note that DL depends on an adequate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

hyperparameter choice and high-performance computing with graphics processing units (GPUs) architecture, which can be challenging to implement in small breeding programs.

2.5 Other models

Klápště et al. (2020) presented a strategy to generate a marker-based relationship matrix that prioritized markers using Partial Least Squares (PLS). This approach downweighs noisy predictors but does not remove them from the model. The advantage of PLS is that it deals with multicollinearity and can handle several response variables at a time. The authors used PLS-CA (PLS-Canonical Analysis) for constructing marker-based relationship matrices with different numbers of markers. This strategy attempts to improve the accuracy of traits with low heritability by taking advantage of the genetic covariance common across all investigated traits. In order to perform the marker selection by PLS-CA, all individuals in the training population must be phenotyped for all traits that will be included in the analysis [32].

The incorporation of staking ensemble ML (SEML) in GS is a promising alternative to increase the predictive ability of ML models. The SEML method uses a meta-learning algorithm to determine how to best combine the predictions from two or more base ML models. Hence, SEML has the potential to generate predictions with better performance than any single model [33]. Liang et al. (2021) tested the prediction accuracy of the SEML approach using three ML models: SVM, kernel ridge regression (KRR) and elastic net (ENET) in Loblolly pine, beef, and dairy cattle. On average, there was an increase of 7.70% in prediction accuracy in nine traits tested. However, Bayes B demonstrated higher prediction accuracies for some traits including milk fat percentage or tree stiffness [34].

3. Genomic selection in polyploids

GS requires high quality genome-wide markers to determine *GEBV*s. Two types of high throughput genotyping methods can be employed: SNP arrays and GBS. There are SNPs arrays with different marker densities in potato [35] and wheat [36]. Alfalfa also has an array with 9,277 SNPs [37]. However, its use has not been widely adopted and GBS is currently the best

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

option to obtain genome-wide markers. During the genotyping process by GBS, different types of markers, such as single nucleotide polymorphisms (SNPs), insertion/deletions (indels), or short tandem repeats (STRs) can be obtained. Genome-wide markers can then be arrayed in a genotypic matrix of *m* samples and *n* markers. The genotypic matrix can be filtered to retain only biallelic SNPs, which are the most abundant and stable markers for identifying QTLs associated with traits of interest [38].

Allele dosage counts alternative allele frequency for each biallelic SNP. In diploid species the genotypic matrix is coded as $\{0,1,2\}$, reflecting if a given marker is present in the homozygous reference (AA), heterozygous (AB) or homozygous alternate (BB) allelic state. For biallelic SNPs in polyploid species with ploidy N, the biallelic dosage is $N + 1$ and the genotypic matrix is coded as $\{0, \dots, N\}$. Genotype calling in autotetraploids requires bioinformatics tools to distinguish among five possible genotypes (AAAA, AAAB, AABB, ABBB, BBBB) with biallelic SNPs coded as $\{0,1,2,3,4\}$. There are several R packages like polyRAD [39], superMASSA [40], FitTetra 2.0 [41] or Updog [42] to obtain allele dosage in numeric format from vcf format. Some of these R packages like Updog require users to specify genotype priors [42] to accurately calculate the allele dosage and distinguish between all possible genotypes. However, the most common option is to use high depth sequence reads (e.g. ~60×) which leads to 98.4% accuracy in genotypic calls [43] The effects of marker allele dosage on phenotype for genomic selection have been reported previously. Slater et el. (2016) described three different models for GS in autopolyploids: additive autotetraploid, pseudodiploid, and full autotetraploid. In the additive autotetraploid model, the allele dosage has an additive effect and 0,1,2,3,4 corresponds to AAAA, AAAB, AABB, ABBB, BBBB, respectively. In the pseudodiploid model all heterozygous genotypes (AAAB, AABB, ABBB) have the same effect of 1 on the genotypic variation, while the two homozygotes AAAA and BBBB have an effect of 0 and 2, respectively. Finally, the full autotetraploid model assumes that each genotype has its own effect with five possible effects per marker, assuming that the markers are fitted as random effects [11]. In addition, Rosyara et al. (2016) developed a software for genome-wide association studies in autopolyploids designated, GWASpoly. GWASpoly has different assumptions over allele dosages and conducts the hypothesis tests for each marker using six models (general, diploidized

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

general, diploidized additive, additive, simplex dominant and duplex dominant models) (Table 3).

Table 3. Coding effect assumptions of GWASpoly models according to allele dosage in biallelic SNPs.

Allele dosage¶	**AAAA**	**AAAB**	**AABB**	**ABBB**	**BBBB**
Numerical code	**0**	**1**	**2**	**3**	**4**
GWASpoly models	**Phenotypic effect**§				
Diplo-additive	0.00		0.50		1.00
Diplo-general‡	0.00		0.00 < x <1.00		1.00
Additive	0.00	0.25	0.50	0.75	1.00
1-dom-ref (`A>B` simplex)	1.00	1.00	1.00	1.00	0.00
2-dom-ref (`A>B` duplex)	1.00	1.00	1.00	0.00	0.00
1-dom-alt (`B>A` simplex)	0.00	1.00	1.00	1.00	1.00
2-dom-alt (`B>A` duplex)	0.00	0.00	1.00	1.00	1.00
General†	No restrictions				

¶, allele dosage A is coded as the reference allele and B is coded as the alternative allele; §, phenotypic effects are scaled from 0.00 to 1.00; ‡, for the diplo-general model all heterozygotes have the same effect (x), but x is not constrained to be halfway between the homozygous effects; †, the general model has no restrictions on the effects of the different dosage levels.

Amadeu et al. (2019) evaluated the inclusion of dominance effects for genomic prediction in autotetraploid crops. They reported that a full autotetraploid model, including additive and dominance effects jointly modeled into a unique general effect, increased the total genetic variance explained [44]. In potato, different covariance genomic marker- and pedigree-based matrices, designated G and A, respectively, were tested to identify additive and nonadditive genetic effects and to improve the accuracy in GS. A matrix (also known as numerator relationship matrix) was calculated from 13-generation pedigree. The A matrix was defined as a matrix containing kinship coefficients among all individuals in the population, multiplied by four. They reported that the G matrix was superior to the A matrix and adding allele dosage

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

information increased the prediction accuracy. Finally, the use of a pseudodiploid matrix reduced the prediction accuracy by 0.13, on average [45].

In the autotetraploid forage grass *Panicum maximum*, de C. Lara et al. (2019) compared the predictive ability of six GS models in six traits using tetraploid and pseudodiploid allele dosages and a minimum depth of 25 reads. Additionally, multiple harvests were modeled with a variance-covariance (VCOV) matrix for genotypes nested across harvests, treating the genotypes as a random factor. The incorporation of correlations among harvests provided a better fit for the traits analyzed. Including the tetraploid dosage also produced higher predictive accuracy compared with pseudodiploid dosages. In autotetraploids with highly mixed ploidy, such as sugarcane and sweet potato, the incorporation of allele dosage information increased model predictive abilities up to 140% in comparison to using diploidized markers [46].
Accuracy of different models showed little changes to ploidy or allele dosage information in in sugarcane on sweet potato. In sugarcane, the Brix trait possessed the highest mean predictive accuracy (0.24) using GBLUP model that included allele dosage. In sweet potato, prediction accuracies were moderate to high. For example, color saturation had the highest mean predictive ability (0.75) using a G model with allele dosage information.

In blueberry (*Vaccinium* spp.), several approaches have been evaluated to improve the genomic selection process because the conventional breeding pipeline takes up to 12 years [31,47,48]. Implementation of GS in early stages of the breeding program could shorten the cycle time to nine years and increase the expected genetic gain by 86%. De Bem Oliveira et al. (2019) compared diploid, tetraploid, and continuous allele dosages at the individual plant level for the application of genomic selection in potato and blueberry. In general, there was no difference among the models tested, but continuous genotypes resulted in a better predictive ability for some traits such as fruit firmness, fruit scar, and fruit diameter. Also, use of a marker-based relationship matrix generated better predictions than pedigree-based relationship matrix (A matrix). Ferrão et al. (2021) reported similar prediction accuracies of GBLUP for four traits using two genotype calling approaches (dosage and ratio) and two read-depth scenarios (6× and 60×). They also observed that combining allele dosage for low to mid sequencing depths (6×−12×) produced similar accuracies to that obtained by high read-depth (60×). The use of mid

 doi:10.20944/preprints202110.0305.v1

sequencing depths will allow modify economic resource allocation to increase the number of individuals genotyped.

Enormous progress has been made during the last few years in the application of GS approaches to polyploids. In alfalfa, GS has been tested in different traits using parametric and non-parametric models. Table 4 summarizes progress that has been made towards applying GS in multiple polyploids, including alfalfa. Although yield is the main trait in alfalfa breeding programs, other agronomic traits such as forage quality and plant regrowth have also been tested [49]. Also, use of an allele dosage genotype matrix has been reported to improve prediction accuracies of forage quality [50] and yield under salt stress [51]. The current challenge is to implement GS in breeding programs and to evaluate increases in GS-derived genetic gain in comparison with PS-derived materials.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

Table 4. Recent achievements in genomic selection (GS) in polyploid crops

Crop	Ploidy	Trait[§]	GS method	Acc[‡]	Notes	Author
Avena sativa	Allohexaploid	Seed lipid content	MK-BLUP	0.48	Use of additive marker effects of Bayesian models during the construction of G matrix	[52]
Brassica napus	Alloteteraploid	Seed yield	GBLUP	0.69	Several agronomic and seed quality traits were tested	[53]
Coffea arabica	Allotetraploid	Canopy diameter	GBLUP	0.40	18 agronomic traits were tested. Diploid dosage assumed	[54]
Eucalyptus nitens	Paleotetraploid	Wood density	MVGLUP	0.77	Marker selection in multivariate analysis. Requires uses multiple traits highly correlated	[32]
Medicago sativa	Autotetraploid	Yield	rrBLUP	0.66	Multi-environment trials over two generations. First report of GS in alfalfa.	[55]
Medicago sativa	Autotetraploid	Yield	SVM	0.35	Six GS models were tested. First report of machine learning models in alfalfa	[56]
Medicago sativa	Autotetraploid	Leaf crude protein	RRBLUP	0.40	Nine alfalfa forage quality traits were tested by five GS models	[50]
Medicago sativa	Autotetraploid	Fall plant height	Bayes B	0.65	15 quality traits and 10 agronomic traits were tested using three GS models	[49]
Medicago sativa	Autotetraploid	Yield under salt stress	SVM	0.50	Multi-environment trials with seven yield measurements. Eight GS models were tested	[51]
Panicum maximum	Autotetraploid	Organic matter	Bayes B-TD	0.39	Genomic selection using tetraploid dosage (GS-TD) vs diploid dosage (GS-DD)	[57]
Solanum tuberosum	Autopolyploid	Yield	GBLUP	0.55	Incorporation of additive and digenic dominant G covariance matrix	[45]
Solanum tuberosum	Autopolyploid	Tuber weight	RKHS	0.59	Four agronomic tuber traits were tested by eight GS models	[58]
Sugarcane	Octaploid and decaploid	Fiber	GBLUP	0.44	Inclusion of additive and non-additive genetic components for GS	[59]
Triticum aestivum	Allohexaploid	Grain yield	GBLUP	0.47	Multi-trait selection for grain yield and protein content	[60]

Triticum aestivum	Allohexaploid	Grain yield	GBLUP	0.53	GWAS markers as fixed effects in GS models.	[61]
Vaccinium corymbosum	Autotetraploid	Weight	GBLUP	0.49	Comparison of allele dosage with depth sequencing: 6×-60×)	[31]

§ For multiple traits, the trait with the highest predictive accuracy was selected; ‡, predictive accuracy measured as Pearson's correlation; MK-BLUP, multi-kernel trait-specific BLUP; MVGLUP, Multi-trait model GBLUP; MT-CV2, multi-trait genomic selection with cross-validation 2; SVM, support vector machine; Bayes B-TD, Bayes B with tetraploid allele dosage; RKHS, Reproducing Kernel Hilbert Space; † In multi-trait genomic selection (MT-GS) a secondary trait that is genetically correlated with the primary trait is incorporated in the prediction model, to predict the primary trait with higher accuracy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

4. Case study: Logan 2020 population

In this review we tested some models of GS using the dataset of alfalfa previously published [51]. Datasets were collected from a multi-parental population generated to select lines tolerant to salt stress. Forage yield under salt stress was measured over seven harvests in 265 individuals for two years. Each harvest was spatially corrected by a two-dimensional P-spline mixed model with Mr.Bean web application [62] using SpATS package [63]. Multiple best linear unbiased estimator (BLUE) values were adjusted in multi-environmental trials using Factor Analytic II covariance structure [64] with ASReml R software [65]. A genotypic matrix of 6,796 high quality GBS-derived SNP markers were obtained using NGSEP v4.0.0 software [66] with parameters previously reported in Medina et al. (2020). SNPs were coded from 0 to 4 according to allele types using Sommer R package [67].

A GS approach using regression analysis between phenotypes (y) and a genotypic matrix was transformed by eight different models ($f(X)$). These included: RRBLUP, Bayes A, B, C, Bayesian Lasso (BL), GBLUP using two G matrices (VanRaden [VR] [equation 4] and full-autotetraploid [FA] [equation 6]), RF and SVM (Figure 2a). All models were compared for execution time and Pearson's correlation using ten-fold cross-validation with the GROAN R package [68]. Execution time is an important factor to consider for GS modeling when computing power is limited. Consequently, system time (seconds) was measured for each model with cross-validation. The fastest models were GBLUP and RRBLUP with an average of 0.06 seconds, whereas SVM and RF required 10.57 and 12.99 seconds, respectively. More time was required for ML models when a grid search was used to estimate the best values for hyperparameters such as cost and sigma in SVM, or *mtry* in RF. However, time was reduced to 1.90 seconds with cross-validation in the SVM training model when hyperparameters were previously defined. Prediction accuracy was approximately 0.3 among the GS models RRBLUP, Bayes A, B, C, BL, RF, GBLUP-VR and GBLUP-FA. The SVM model possessed the highest accuracy (0.46), in agreement with previous reports [51,56].

Variable importance values or SNP weights were obtained using SVM and RF models with the Caret R package [69], or by retrieving −log10 p-values resulting from six models of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

GWASpoly R package (i.e., general, diploidized general, diploidized additive, additive, simplex dominant and duplex dominant models [Table 3]) [70]. SNP weights were used as input in a **D** diagonal matrix for the construction of a G^* matrix [equation 7] in the WGBLUP model (Figure 2b and 2c). Pearson's correlation among variable importance values of different models was measured to identify models with similar SNPs weights. Diploidized additive and diploidized general models had the highest Pearson's correlation (0.87), followed by additive and diploidized additive models (0.74). Variable importance values derived from RF had low correlations across all models tested (Figure 2c). Prediction accuracies for GBLUP with two G matrix and ten WGBLUP models were compared by measuring Pearson's correlation 10 times with ten-fold cross-validation. Incorporation of variable importance values in WGBLUP increased prediction accuracies. Pearson correlations ranged from a low of 0.32 in GBLUP-VR (no variable importance values) to 0.63 in WGBLUP-SVM, with the highest prediction accuracy (0.83) achieved when −log10 p-values from the additive GWASpoly model were used as a weight vector (Figure 2d). Thus, incorporation of a diagonal matrix **D** with variable importance values to the G matrix increased GS predictive ability almost three times without increasing computational time. This is the first report using SNPs weights to increase prediction accuracy of GS in alfalfa. Our results suggest that including SNP marker −log10 p-values derived from the additive GWASpoly model into a WGBLUP model may benefit prediction accuracy and selection for improvement of complex traits in alfalfa breeding programs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

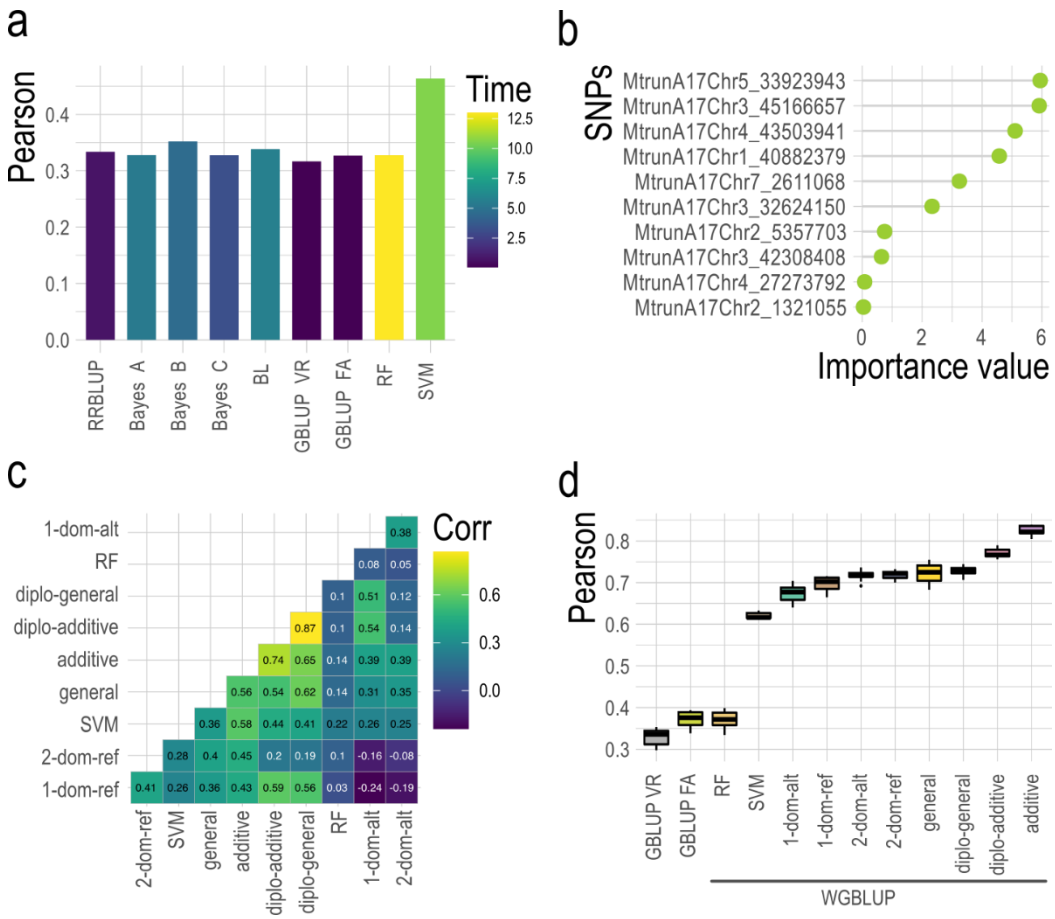

Figure 2. Optimization of GS models. a) GS model accuracy measured as Pearson's correlation after 10-fold cross-validation for biomass yield under salt stress. Computing time was measured as system time in seconds to run one cross-validation. b) Example of variable importance values derived from SVM for ten randomly chosen SNPs. c) Pearson's correlation for 6,796 SNPs weights obtained by variable importance (SVM, RF) or by −log10 p-values of different GWASpoly models. d) Accuracy of GBLUP (GBLUP VR and GBLUP FA) and WGBLUP models. Accuracy was measured ten times using Pearson's correlation with ten-fold cross-validation. SNP weights for WGBLUP were obtained from variable importance values (SVM, RF) or −log10 p-values of different GWASpoly models. RRBLUP, best linear unbiased prediction using ridge-regression; BL Bayes LASSO; GBLUP, genomic best linear unbiased prediction; VR, VanRaden G matrix; FA, full autotetraploid G matrix; RF, random forest; SVM, support vector machine; WGBLUP, weighted GBLUP; 1-dom-alt and 1-dom-ref, simplex dominant models; 2-dom-alt and 2-dom-ref, duplex dominant models; diplo-general, diploidized general; diplo-additive, diploidized additive.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2021 doi:10.20944/preprints202110.0305.v1

5. Conclusions

Genomic selection is a breeding strategy that predicts the genomic estimated breeding value (GEBV) of individuals in a population using genomic-wide genetic markers. A significant advantage of GS is the ability to select superior individuals in a population at very early stages in a breeding cycle based on their genotype, versus conducting lengthy/expensive phenotyping trials prior to each selection cycle. For instance, in a simulated wheat breeding program, selection based on GEBVs for grain yield tripled the genetic gain compared with PS [71]. In alfalfa breeding programs, GS can be implemented in elite large germplasm panels with genotyped individuals to decrease PS efforts, thus reducing overall selection cycle time and accelerating variety development. However, determining correct GEBVs relies on prediction accuracy which can vary according to a combination of phenotypic trait, genotypic information, and the statistical model. Consequently, GS research efforts have focused heavily on evaluating prediction accuracies of multiple parametric and non-parametric models to develop robust strategies which can be used in testing populations. Once robust modeling strategies are developed, GS has the potential to accumulate thousands of favorable alleles to develop climate-resilient crops with high yield potential. Additionally, as genotyping is required only once for a given population, multiple traits can be associated with the same genotypic matrix to determine GEBVs for each trait, thus making GS a valuable approach in multi-trait selection [72].

There is a need to explore new methodologies to improve molecular and bioinformatic tools for the application of GS in polyploid crops. The development of new approaches to obtain high quality genome-wide markers will help to resolve the genetic architecture of complex traits. For example, the PRINCESS platform uses long-read sequencing to detect SNPs, indels, or methylation sites with high accuracy [73]. In GWAS and GS, the number of individuals in a population is crucial to maximize statistical power. Therefore, researchers search for genotyping methods which optimize the balance between cost, sample size and the number of SNPs. In this regard, GBS is a relatively affordable genotyping methodology. But for polyploid crops, there is a need for high-coverage sequencing (i.e., read-depth) to accurately estimate allele dosage, which increases genotyping cost. Biallelic SNPs are commonly used in polyploid GS because they are

 doi:10.20944/preprints202110.0305.v1

the most abundant type and are easier to transform into a numerical format for developing a genotypic matrix. However, during construction of the variant call format file (vcf), ~20% of high-quality SNP markers are discarded because they are not biallelic. Additionally, indels or simple sequence repeats (SSR) could add important information to GS model to increase the prediction accuracy.

Parametric models such as RRBLUP assume that all SNPs have an effect on specific trait, but the actual effect of each SNP is very small (heavy shrinkage) [5]. Although multiple loci have an effect on a complex trait, they often have different weights. Thus, identifying trait-specific weights for SNP marker alleles should increase prediction accuracy in GS models such as WGBLUP. Such outcomes have been demonstrated in animal breeding research [74,75] but not in polyploid crops. In this regard, utilizing a variable importance approach based on −log10 p-values for the additive GWASpoly model for alfalfa yield under salt stress was the best strategy to generate a diagonal matrix **D**. For this complex trait, we present empirical evidence demonstrating that the WGBLUP model increased prediction accuracy by almost 3 times compared to RRBLUP, GBLUP, Bayesian or ML models. The WGBLUP approach is simple, does not require high performance computing, and can be applied to different crops to predict breeding values and accelerate selection cycles.

Data availability: The raw data of GBS were submitted to the NCBI Sequence Read Archive with bioproject ID: PRJNA611554 and biosample # SAMN14336867.

Acknowledgments: The authors would like to acknowledge to Ms. Martha Rivera for technical help. They would further like to thank USDA-NIFA for funding support (2015-70005-24071).

Funding Information: The authors would like to thank USDA-NIFA for funding support (2015-70005-24071).

Author Contributions: C.A.M.: Manuscript preparation and data analysis; H.K.: Manuscript preparation and data analysis; I.R.: Conceived and outlined this research and manuscript preparation; L.-X.Y.: Conceived and outlined this research and manuscript preparation. All authors have read and agreed to submit the manuscript.

Conflict of Interest and other Ethics Statements: The authors do not have conflict of interest

6. References

[1] F. Blondon, D. Marie, S. Brown, A. Kondorosi, Genome size and base composition in Medicago sativa and M. truncatula species, Genome. 37 (1994) 264–270. https://doi.org/10.1139/g94-037.

[2] R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, S.E. Mitchell, A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species, PLoS One. 6 (2011) e19379. https://doi.org/10.1371/journal.pone.0019379.

[3] M.G. Bulmer, The Effect of Selection on Genetic Variability, Am. Nat. 105 (1971) 201–211.

[4] T.H.E. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps., Genetics. 157 (2001) 1819–29.

[5] J.B. Endelman, Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP, Plant Genome. 4 (2011) 250–255. https://doi.org/10.3835/plantgenome2011.08.0024.

[6] B.J. Hayes, P.J. Bowman, A.C. Chamberlain, K. Verbyla, M.E. Goddard, Accuracy of genomic breeding values in multi-breed dairy cattle populations, Genet. Sel. Evol. 41 (2009) 51. https://doi.org/10.1186/1297-9686-41-51.

[7] P. Pérez, G. Campos, J. Crossa, D. Gianola, Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R, Plant Genome. 3 (2010) plantgenome2010.04.0005. https://doi.org/10.3835/plantgenome2010.04.0005.

[8] P.M. Vanraden, Genomic measures of relationship and inbreeding, Interbull Bull. 25 (2007) 33–33.

[9] P.M. VanRaden, Efficient methods to compute genomic predictions, J. Dairy Sci. 91 (2008) 4414–4423. https://doi.org/10.3168/jds.2007-0980.

[10] J. Yang, B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.E. Goddard, P.M. Visscher, Common SNPs explain a large proportion of the heritability for human height, Nat. Genet.

42 (2010) 565–569. https://doi.org/10.1038/ng.608.

[11] A.T. Slater, N.O.I. Cogan, J.W. Forster, B.J. Hayes, H.D. Daetwyler, Improving Genetic Gain with Genomic Selection in Autotetraploid Potato, Plant Genome. 9 (2016). https://doi.org/10.3835/plantgenome2016.02.0021.

[12] Z. Zhang, J. Liu, X. Ding, P. Bijma, D.-J. de Koning, Q. Zhang, Best Linear Unbiased Prediction of Genomic Breeding Values Using a Trait-Specific Marker-Derived Relationship Matrix, PLoS One. 5 (2010) e12648. https://doi.org/10.1371/journal.pone.0012648.

[13] A. Legarra, C. Robert-Granié, P. Croiseau, F. Guillaume, S. Fritz, Improved Lasso for genomic selection, Genet. Res. (Camb). 93 (2011) 77–87. https://doi.org/10.1017/S0016672310000534.

[14] L.-Y. Chang, S. Toghiani, E.H. Hay, S.E. Aggrey, R. Rekaya, A Weighted Genomic Relationship Matrix Based on Fixation Index (FST) Prioritized SNPs for Genomic Selection, Genes (Basel). 10 (2019) 922. https://doi.org/10.3390/genes10110922.

[15] D. Ren, L. An, B. Li, L. Qiao, W. Liu, Efficient weighting methods for genomic best linear-unbiased prediction (BLUP) adapted to the genetic architectures of quantitative traits, Heredity (Edinb). 126 (2021) 320–334. https://doi.org/10.1038/s41437-020-00372-y.

[16] P. Pérez, G. de los Campos, Genome-Wide Regression and Prediction with the BGLR Statistical Package, Genetics. 198 (2014) 483–495. https://doi.org/10.1534/genetics.114.164442.

[17] T.H. Meuwissen, T.R. Solberg, R. Shepherd, J.A. Woolliams, A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value, Genet. Sel. Evol. 41 (2009) 2. https://doi.org/10.1186/1297-9686-41-2.

[18] D. Habier, R.L. Fernando, K. Kizilkaya, D.J. Garrick, Extension of the bayesian alphabet for genomic selection., BMC Bioinformatics. 12 (2011) 186. https://doi.org/10.1186/1471-2105-12-186.

[19] G. de los Campos, H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, J.M. Cotes, Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree, Genetics. 182 (2009) 375–385. https://doi.org/10.1534/genetics.109.101501.

[20] L. Breiman, Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author), Stat. Sci. 16 (2001) 199–215. https://doi.org/10.1214/ss/1009213726.

[21] H. Drucker, C.J.C. Surges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, in: M.C. Mozer, M.I. Jordan, Petsche. T. (Eds.), Adv. Neural Inf. Process. Syst., MIT Press, Cambridge, Massachusetts, 1997: pp. 155–161.

[22] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324.

[23] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297. https://doi.org/10.1007/BF00994018.

[24] W. Liu, X. Meng, Q. Xu, D.R. Flower, T. Li, Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models, BMC Bioinformatics. 7 (2006) 182. https://doi.org/10.1186/1471-2105-7-182.

[25] S. Sun, Z. Cao, H. Zhu, J. Zhao, A Survey of Optimization Methods From a Machine Learning Perspective, IEEE Trans. Cybern. (2019). https://doi.org/10.1109/tcyb.2019.2950779.

[26] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444. https://doi.org/10.1038/nature14539.

[27] T. Pook, J. Freudenthal, A. Korte, H. Simianer, Using Local Convolutional Neural Networks for Genomic Prediction, Front. Genet. 11 (2020). https://doi.org/10.3389/fgene.2020.561497.

[28] K.S. Sandhu, D.N. Lozada, Z. Zhang, M.O. Pumphrey, A.H. Carter, Deep Learning for Predicting Complex Traits in Spring Wheat Breeding Program, Front. Plant Sci. 11 (2021). https://doi.org/10.3389/fpls.2020.613325.

[29] K. Sandhu, S.S. Patil, M. Pumphrey, A. Carter, Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program, Plant Genome. (2021) 1–17. https://doi.org/10.1002/tpg2.20119.

[30] L.M. Zingaretti, S.A. Gezan, L.F. V. Ferrão, L.F. Osorio, A. Monfort, P.R. Muñoz, V.M. Whitaker, M. Pérez-Enciso, Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species, Front. Plant Sci. 11 (2020) 1–14. https://doi.org/10.3389/fpls.2020.00025.

[31] L.F. V. Ferrão, R.R. Amadeu, J. Benevenuto, I. de Bem Oliveira, P.R. Munoz, Genomic

Selection in an Outcrossing Autotetraploid Fruit Crop: Lessons From Blueberry Breeding, Front. Plant Sci. 12 (2021). https://doi.org/10.3389/fpls.2021.676326.

[32] J. Klápště, H.S. Dungey, E.J. Telfer, M. Suontama, N.J. Graham, Y. Li, R. McKinley, Marker Selection in Multivariate Genomic Prediction Improves Accuracy of Low Heritability Traits, Front. Genet. 11 (2020) 1–15. https://doi.org/10.3389/fgene.2020.499094.

[33] G. Kyriakides, K.G. Margaritis, Hands-on ensemble learning with Python : build highly optimized ensemble machine learning models using scikit-learn and Keras LK, 1st ed., Birmingham: Packt Publishing Ltd, 2019. https://www.packtpub.com/product/hands-on-ensemble-learning-with-python/9781789612851.

[34] M. Liang, T. Chang, B. An, X. Duan, L. Du, X. Wang, J. Miao, L. Xu, X. Gao, L. Zhang, J. Li, H. Gao, A Stacking Ensemble Learning Framework for Genomic Prediction, Front. Genet. 12 (2021) 1–9. https://doi.org/10.3389/fgene.2021.600040.

[35] P.G. Vos, J.G.A.M.L. Uitdewilligen, R.E. Voorrips, R.G.F. Visser, H.J. van Eck, Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history, Theor. Appl. Genet. 128 (2015) 2387–2401. https://doi.org/10.1007/s00122-015-2593-y.

[36] M.O. Winfield, A.M. Allen, A.J. Burridge, G.L.A. Barker, H.R. Benbow, P.A. Wilkinson, J. Coghill, C. Waterfall, A. Davassi, G. Scopes, A. Pirani, T. Webster, F. Brew, C. Bloor, J. King, C. West, S. Griffiths, I. King, A.R. Bentley, K.J. Edwards, High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool, Plant Biotechnol. J. 14 (2016) 1195–1206. https://doi.org/10.1111/pbi.12485.

[37] X. Li, Y. Han, Y. Wei, A. Acharya, A.D. Farmer, J. Ho, M.J. Monteros, E.C. Brummer, Development of an Alfalfa SNP Array and Its Use to Evaluate Patterns of Population Structure and Linkage Disequilibrium, PLoS One. 9 (2014) e84329. https://doi.org/10.1371/journal.pone.0084329.

[38] J. Perkel, SNP genotyping: six technologies that keyed a revolution, Nat. Methods. 5 (2008) 447–453. https://doi.org/10.1038/nmeth0508-447.

[39] L. V. Clark, A.E. Lipka, E.J. Sacks, polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids, G3 Genes, Genomes, Genet. (2019) g3.200913.2018. https://doi.org/10.1534/g3.118.200913.

[40] G.S. Pereira, A.A.F. Garcia, G.R.A. Margarido, A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids, BMC Bioinformatics. 19 (2018) 398. https://doi.org/10.1186/s12859-018-2433-6.

[41] K. Zych, G. Gort, C.A. Maliepaard, R.C. Jansen, R.E. Voorrips, FitTetra 2.0 – improved genotype calling for tetraploids with multiple population and parental data support, BMC Bioinformatics. 20 (2019) 148. https://doi.org/10.1186/s12859-019-2703-y.

[42] D. Gerard, L.F.V. Ferrão, A.A.F. Garcia, M. Stephens, Genotyping Polyploids from Messy Sequencing Data, Genetics. 210 (2018) 789–807. https://doi.org/10.1534/genetics.118.301468.

[43] J.G.A.M.L. Uitdewilligen, A.-M.A. Wolters, B.B. D'hoop, T.J.A. Borm, R.G.F. Visser, H.J. van Eck, A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato, PLoS One. 8 (2013) e62355. https://doi.org/10.1371/journal.pone.0062355.

[44] R.R. Amadeu, L.F. V. Ferrão, I. de B. Oliveira, J. Benevenuto, J.B. Endelman, P.R. Munoz, Impact of Dominance Effects on Autotetraploid Genomic Prediction, Crop Sci. 60 (2019) 656–665. https://doi.org/10.2135/cropsci2019.02.0138.

[45] J.B. Endelman, C.A.S. Carley, P.C. Bethke, J.J. Coombs, M.E. Clough, W.L. Silva, W.S. De Jong, D.S. Douches, C.M. Frederick, K.G. Haynes, D.G. Holm, J.C. Miller, P.R. Muñoz, F.M. Navarro, R.G. Novy, J.P. Palta, G.A. Porter, K.T. Rak, V.R. Sathuvalli, A.L. Thompson, G.C. Yencho, Genetic Variance Partitioning and Genome-Wide Autotetraploid Potato, Genetics. 209: (2018) 77–87. https://doi.org/10.1534/genetics.118.300685/-/DC1.1.

[46] L.G. Batista, V.H. Mello, A.P. Souza, G.R.A. Margarido, Genomic prediction with allele dosage information in highly polyploid species, BioRxiv. (2021) 1–33.

[47] I. de Bem Oliveira, M.F.R. Resende, L.F. V. Ferrão, R.R. Amadeu, J.B. Endelman, M. Kirst, A.S.G. Coelho, P.R. Munoz, Genomic Prediction of Autotetraploids; Influence of Relationship Matrices, Allele Dosage, and Continuous Genotyping Calls in Phenotype Prediction, G3 Genes|Genomes|Genetics. 9 (2019) 1189–1198. https://doi.org/10.1534/g3.119.400059.

[48] I. de Bem Oliveira, R.R. Amadeu, L.F.V. Ferrão, P.R. Muñoz, Optimizing whole-genomic prediction for autotetraploid blueberry breeding, Heredity (Edinb). 125 (2020) 437–448.

https://doi.org/10.1038/s41437-020-00357-x.

[49] C. Jia, F. Zhao, X. Wang, J. Han, H. Zhao, G. Liu, Z. Wang, Genomic Prediction for 25 Agronomic and Quality Traits in Alfalfa (*Medicago sativa*), Front. Plant Sci. 9 (2018) 1220. https://doi.org/10.3389/fpls.2018.01220.

[50] E. Biazzi, N. Nazzicari, L. Pecetti, E.C. Brummer, A. Palmonari, A. Tava, P. Annicchiarico, Genome-Wide Association Mapping and Genomic Selection for Alfalfa (*Medicago sativa*) Forage Quality Traits, PLoS One. 12 (2017) e0169234. https://doi.org/10.1371/journal.pone.0169234.

[51] C.A. Medina, C. Hawkins, X.-P. Liu, M. Peel, L.-X. Yu, Genome-Wide Association and Prediction of Traits Related to Salt Tolerance in Autotetraploid Alfalfa (*Medicago sativa* L.), Int. J. Mol. Sci. 21 (2020) 3361. https://doi.org/10.3390/ijms21093361.

[52] M.T. Campbell, H. Hu, T.H. Yeats, L.J. Brzozowski, M. Caffe-Treml, L. Gutiérrez, K.P. Smith, M.E. Sorrells, M.A. Gore, J.-L. Jannink, Improving Genomic Prediction for Seed Quality Traits in Oat (*Avena sativa* L.) Using Trait-Specific Relationship Matrices, Front. Genet. 12 (2021) 1–12. https://doi.org/10.3389/fgene.2021.643733.

[53] M. Fikere, D.M. Barbulescu, M.M. Malmberg, P. Maharjan, P.A. Salisbury, S. Kant, J. Panozzo, S. Norton, G.C. Spangenberg, N.O.I. Cogan, H.D. Daetwyler, Genomic Prediction and Genetic Correlation of Agronomic, Blackleg Disease, and Seed Quality Traits in Canola (*Brassica napus* L.), Plants. 9 (2020) 719. https://doi.org/10.3390/plants9060719.

[54] T.V. Sousa, E.T. Caixeta, E.R. Alkimim, A.C.B. Oliveira, A.A. Pereira, N.S. Sakiyama, L. Zambolim, M.D.V. Resende, Early Selection Enabled by the Implementation of Genomic Selection in *Coffea arabica* Breeding, Front. Plant Sci. 9 (2019) 1–12. https://doi.org/10.3389/fpls.2018.01934.

[55] X. Li, Y. Wei, A. Acharya, J.L. Hansen, J.L. Crawford, D.R. Viands, R. Michaud, A. Claessens, E.C. Brummer, Genomic Prediction of Biomass Yield in Two Selection Cycles of a Tetraploid Alfalfa Breeding Population, Plant Genome. 8 (2015) 1–10. https://doi.org/10.3835/plantgenome2014.12.0090.

[56] P. Annicchiarico, N. Nazzicari, X. Li, Y. Wei, L. Pecetti, E.C. Brummer, Accuracy of genomic selection for alfalfa biomass yield in different reference populations, BMC Genomics. 16 (2015) 1020. https://doi.org/10.1186/s12864-015-2212-y.

[57] L.A. de C. Lara, M.F. Santos, L. Jank, L. Chiari, M. de M. Vilela, R.R. Amadeu, J.P.R. dos Santos, G. da S. Pereira, Z.-B. Zeng, A.A.F. Garcia, Genomic Selection with Allele Dosage in *Panicum maximum* Jacq., G3 Genes|Genomes|Genetics. 9 (2019) 2463–2475. https://doi.org/10.1534/g3.118.200986.

[58] S. Wilson, C. Zheng, C. Maliepaard, H.A. Mulder, R.G.F. Visser, A. van der Burgt, F. van Eeuwijk, Understanding the Effectiveness of Genomic Prediction in Tetraploid Potato, Front. Plant Sci. 12 (2021) 1–13. https://doi.org/10.3389/fpls.2021.672417.

[59] S. Yadav, X. Wei, P. Joyce, F. Atkin, E. Deomano, Y. Sun, L.T. Nguyen, E.M. Ross, T. Cavallaro, K.S. Aitken, B.J. Hayes, K.P. Voss-Fels, Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive genetic effects, Theor. Appl. Genet. 134 (2021) 2235–2252. https://doi.org/10.1007/s00122-021-03822-1.

[60] S. Michel, F. Löschenberger, C. Ametz, B. Pachler, E. Sparry, H. Bürstmayr, Simultaneous selection for grain yield and protein content in genomics-assisted wheat breeding, Theor. Appl. Genet. 132 (2019) 1745–1760. https://doi.org/10.1007/s00122-019-03312-5.

[61] D. Sehgal, U. Rosyara, S. Mondal, R. Singh, J. Poland, S. Dreisigacker, Incorporating Genome-Wide Association Mapping Results Into Genomic Prediction Models for Grain Yield and Yield Stability in CIMMYT Spring Bread Wheat, Front. Plant Sci. 11 (2020) 1–12. https://doi.org/10.3389/fpls.2020.00197.

[62] J.S. Aparicio Arce, Mr.Bean, (2018). https://apariciojohan.shinyapps.io/Mrbean/ (accessed March 18, 2020).

[63] M.X. Rodríguez-Álvarez, M.P. Boer, F.A. van Eeuwijk, P.H.C. Eilers, Spatial Models for Field Trials, (2016) 1–39. http://arxiv.org/abs/1607.08255.

[64] F. Isik, J. Holland, C. Maltecca, Multi Environmental Trials, in: F. Isik, J. Holland, C. Maltecca (Eds.), Genet. Data Anal. Plant Anim. Breed., Springer International Publishing, Cham, 2017: pp. 227–262. https://doi.org/10.1007/978-3-319-55177-7_8.

[65] D.G. Butler, B.R. Cullis, A.R. Gilmour, B.J. Gogel, R. Thompson, ASReml-R Reference Manual Version 4, ASReml-R Ref. Man. (2018) 176. http://www.homepages.ed.ac.uk/iwhite/asreml/uop.

[66] J. Duitama, J.C. Quintero, D.F. Cruz, C. Quintero, G. Hubmann, M.R. Foulquié-Moreno, K.J. Verstrepen, J.M. Thevelein, J. Tohme, An integrated framework for discovery and

genotyping of genomic variants from high-throughput sequencing experiments, Nucleic Acids Res. 42 (2014) e44–e44. https://doi.org/10.1093/nar/gkt1381.

[67] G. Covarrubias-Pazaran, Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer, PLoS One. 11 (2016) e0156744. https://doi.org/10.1371/journal.pone.0156744.

[68] R. Bernardo, J. Yu, Prospects for genomewide selection for quantitative traits in maize, Crop Sci. 47 (2007) 1082–1090. https://doi.org/10.2135/cropsci2006.11.0690.

[69] M. Kuhn, J. Contributions from Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R.C. Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, T. Hunt., caret: Classification and Regression Training, (2019). https://cran.r-project.org/package=caret (accessed March 20, 2020).

[70] U.R. Rosyara, W.S. De Jong, D.S. Douches, J.B. Endelman, Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato, Plant Genome. 9 (2016) 1–10. https://doi.org/10.3835/plantgenome2015.08.0073.

[71] B.B. Tessema, H. Liu, A.C. Sørensen, J.R. Andersen, J. Jensen, Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat, Front. Genet. 11 (2020) 1–12. https://doi.org/10.3389/fgene.2020.578123.

[72] S. Moeinizade, A. Kusmec, G. Hu, L. Wang, P.S. Schnable, Multi-trait Genomic Selection Methods for Crop Improvement, Genetics. 215 (2020) 931–945. https://doi.org/10.1534/genetics.120.303305.

[73] M. Mahmoud, H. Doddapaneni, W. Timp, F.J. Sedlazeck, PRINCESS: comprehensive detection of haplotype resolved SNVs, SVs, and methylation, Genome Biol. 22 (2021) 268. https://doi.org/10.1186/s13059-021-02486-w.

[74] B.I. Lopez, S.-H. Lee, J.-E. Park, D.-H. Shin, J.-D. Oh, S. de las Heras-Saldana, J. van der Werf, H.-H. Chai, W. Park, D. Lim, Correction: Weighted Genomic Best Linear Unbiased Prediction for Carcass Traits in Hanwoo Cattle. Genes 2019, 10, 1019, Genes (Basel). 11 (2020) 1013. https://doi.org/10.3390/genes11091013.

[75] X. Zhang, D. Lourenco, I. Aguilar, A. Legarra, I. Misztal, Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS, Front. Genet. 7 (2016) 1–14. https://doi.org/10.3389/fgene.2016.00151.