Journal of Plant Physiology 257 (2021) 153354

Contents lists available at ScienceDirect

Journal of Plant Physiology

journal homepage: www.elsevier.com/locate/jplph

Review article

Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data

Hao Tong [a,b,c], Zoran Nikoloski [a,b,c,*]

[a] Bioinformatics Group, Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany
[b] Bioinformatics and Mathematical Modeling Department, Centre for Plant Systems Biology and Biotechnology, Plovdiv, Bulgaria
[c] Systems Biology and Mathematical Modeling Group, Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

ARTICLE INFO

Keywords:
Genomic selection
Genomic prediction
Machine learning
Multiple traits
Multi-omics
G×E interaction

ABSTRACT

Highly efficient and accurate selection of elite genotypes can lead to dramatic shortening of the breeding cycle in major crops relevant for sustaining present demands for food, feed, and fuel. In contrast to classical approaches that emphasize the need for resource-intensive phenotyping at all stages of artificial selection, genomic selection dramatically reduces the need for phenotyping. Genomic selection relies on advances in machine learning and the availability of genotyping data to predict agronomically relevant phenotypic traits. Here we provide a systematic review of machine learning approaches applied for genomic selection of single and multiple traits in major crops in the past decade. We emphasize the need to gather data on intermediate phenotypes, e.g. metabolite, protein, and gene expression levels, along with developments of modeling techniques that can lead to further improvements of genomic selection. In addition, we provide a critical view of factors that affect genomic selection, with attention to transferability of models between different environments. Finally, we highlight the future aspects of integrating high-throughput molecular phenotypic data from omics technologies with biological networks for crop improvement.

1. Introduction

To fill the gap between increasing world population and food shortage, particularly in developing countries, the third agricultural (so-called green) revolution aims at increasing the world crop production by employing modern plant breeding techniques. Plant breeding denotes the process of improving desirable phenotypic traits in plants, e.g. yield and quality. It relies on artificial selection that is the process of controlling the genetic changes to benefit human demands (Gregory, 2009). As a result, the genotypes exhibiting improved phenotypic traits are selected by breeders, thus leading to accumulation of favorable alleles. Apart from relying on selection from naturally occurring genotypes, breeders create more diversity by mating phenotypically different genotypes for a given trait. Backcrossing is the most commonly applied procedure that aims to transmit resistant genes to a susceptible genotype. To this end, a parent with resistant genes is used as a donor, while that with susceptible background is the recurrent parent. The offspring are backcrossed with the recurrent parent, and the resulting offspring with a favorable phenotype are selected for further cycles of backcrossing. The conventional backcross breeding process relies on selecting genotypes that are visually close to the recurrent parent, with desired phenotypic traits. Thus, this procedure employs visual selection in manual and resource-intensive fashion that is prone to substantial errors.

Plant breeding has been revolutionized by usage of genetic markers, representing DNA sequences that identify genotypes. For instance, single nucleotide polymorphisms (SNPs) have become widely used low-cost and high-density genetic markers (Vignal et al., 2002). Marker-assisted selection (MAS) relies on genetic markers associated with target genes to accurately select genotypes of interest. As a result, few genotypes with desired phenotypic traits remain in every cycle of selection, leading to reduction in errors due to visual selection. Therefore, the selection process in MAS, particularly for qualitative traits, is considerably faster in recovering the recurrent parent genome compared to conventional breeding based on backcrossing. The key step in MAS is mapping of quantitative trait loci (QTL), regions of DNA that are associated with a particular trait. Therefore, MAS requires the resource-intensive development of biparental populations based on the parents with distant phenotypic traits. The main challenge in MAS breeding is that QTL mapping can only detect loci with major effects, although most

* Corresponding author at: University of Potsdam, Institute of Biochemistry and Biology, Karl-Liebknecht-Str. 24-25, 14476, Potsdam, Germany.
E-mail address: nikoloski@mpimp-golm.mpg.de (Z. Nikoloski).

https://doi.org/10.1016/j.jplph.2020.153354
Received 30 October 2020; Received in revised form 14 December 2020; Accepted 15 December 2020
Available online 29 December 2020

agronomically relevant traits are often complex and controlled by several minor effect loci (Desta and Ortiz, 2014).

To overcome this problem, an advanced breeding approach termed genomic selection (GS) has been proposed (Jannink et al., 2010). GS assumes that all QTL for a trait are in linkage disequilibrium with at least one marker, thus affording the use of data on genome-wide genetic markers which are increasingly available for agronomically relevant crops. To this end, GS forgoes the step of identifying the QTL for a trait. Most importantly, it allows the consideration of minor-effect QTL that cannot be detected by conventional approaches, thus leading to an increase in its prediction power in comparison to MAS. Meanwhile, unlike MAS, GS can be applied not only with biparental populations but also with populations that capture natural genetic variations.

2. General models for genomic selection

Breeding value is the genetic effect that can be stably passed to offspring and is equivalent to the additive effect for a trait of interest. Artificial selection relies on determining breeding values for the considered genotypes. Thus, the key question is that of quantifying the breeding value of a genotype for a phenotypic trait of interest. In animal breeding, the estimated breeding value (EBV) can be predicted based on a model that relates the phenotype over a population of genotypes with their pedigree information by using the best linear unbiased prediction (BLUP) (Nakaya and Isobe, 2012). However, this method is infeasible for populations without pedigree information or with complex population structure, which typically is the case in plant breeding. GS overcomes this problem by determining the genomic estimated breeding value (GEBV), which estimates the breeding value using genome-wide markers based on the application of various machine learning approaches (Meuwissen et al., 2001). To this end, the GS models are developed in a training population which is both genotyped and phenotyped (Fig. 1). The resulting GS model is used with another population, referred to as a breeding population (testing population) that is only genotyped, to predict phenotypic traits in unseen genotypes and for genotype selection in the breeding process. The described procedure underlying GS can be applied to predict phenotypic traits for inbred (Fig. 1A) as well as hybrid populations (Fig. 1B). In doing so, GS halves the time to complete the breeding cycle required for the development of new cultivars (e.g. six years for soybean (Matei et al., 2018), ten years for cassava (de Oliveira et al., 2012), ten years for wheat (Ahmar et al., 2020)).

Essentially, all GS approaches are statistical in nature as they rely on learning a mathematical relation between genotypic and phenotypic data that is later used for prediction. In this sense, GS approaches are part of machine learning that aims to develop and further analyze the performance of models by using a set of training data. The models are then used with unseen data to make predictions. Machine learning approaches come in different flavors, and can roughly be divided into unsupervised, supervised, and reinforcement learning. Unsupervised learning approaches aim to identify patterns in the provided data with no human supervision; the latter implies that relationships between responses and inputs are established in absence of input-response pairs. Typical representative of unsupervised approaches is ordinary least square regression. In contrast, supervised learning requires that input-response pairs (so-called labeled training data) are provided, and uses them in learning a model that maps inputs to responses. The classification of genotypes into good and bad performance (as labels) based on a set of molecular data is a prime example of supervised learning. Deep learning relies on usage of artificial neural networks to arrive at higher-level features extracted from the data on the inputs to predict the responses; it usually results in nonlinear models for the response in terms of the inputs, which although of good performance, may be difficult to interpret. Finally, reinforcement learning relies on feedback (i.e. reward) on the model performance from dynamic environment while continuously improving the model building by maximizing the rewards. Applications of reinforcement learning in biology is limited, with one recent prime example of DeepMind that combines deep and reinforcement learning to predict protein structure (Callaway, 2020).

The existing GS approaches can be categorized based on different criteria: (1) statistical/machine learning techniques; for instance, GS approaches can be roughly grouped into those relying on regression, classification, or deep learning models (Fig. 2), (2) number of traits predicted; GS models can be clearly divided into those modeling single or multiple traits, and (3) type of predictors used; based on this criterion, GS models can be trained only on genotypic data or can include intermediate phenotypes as well as variables that capture environmental factors. The goal of the review is not to dwell on the mathematical details of the underlying machine learning approaches (see the glossary of terms), but to point out the main advantages and disadvantages of the approaches together with the results of their applications in crop breeding in the last ten years.

2.1. Regression-based models for single trait genomic selection

Regression is the simplest statistical approach that can model the relationship between a single phenotypic trait and genotype specified by a set of genomic markers. In the case of GS, the number of predictors (i.e.

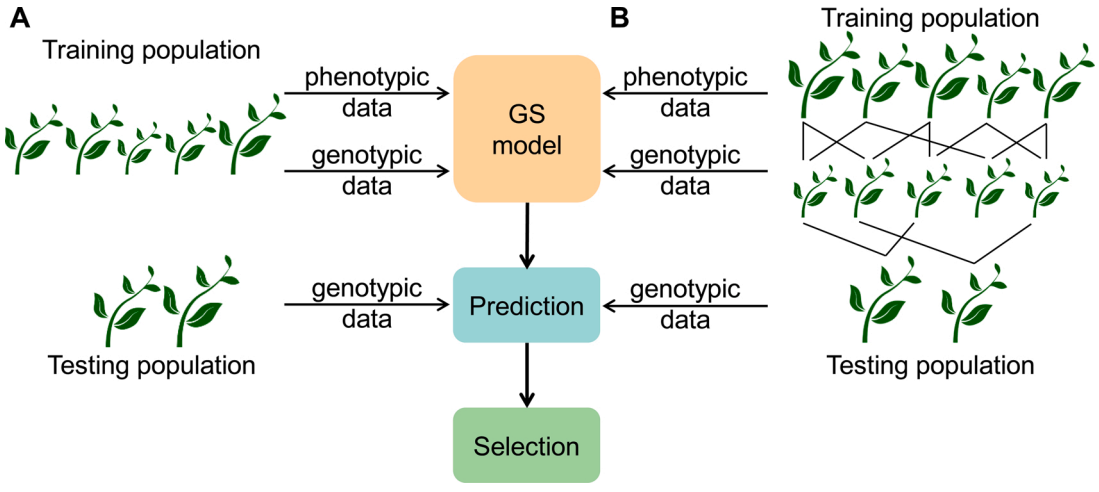

Fig. 1. The principle of genomic selection. Training population, consisting of (A) inbreeds and (B) hybrids are both genotyped and phenotyped. The resulting data are used to train GS models and then used on respective testing populations which are only genotyped. The predicted phenotypes are finally used in artificial selection.

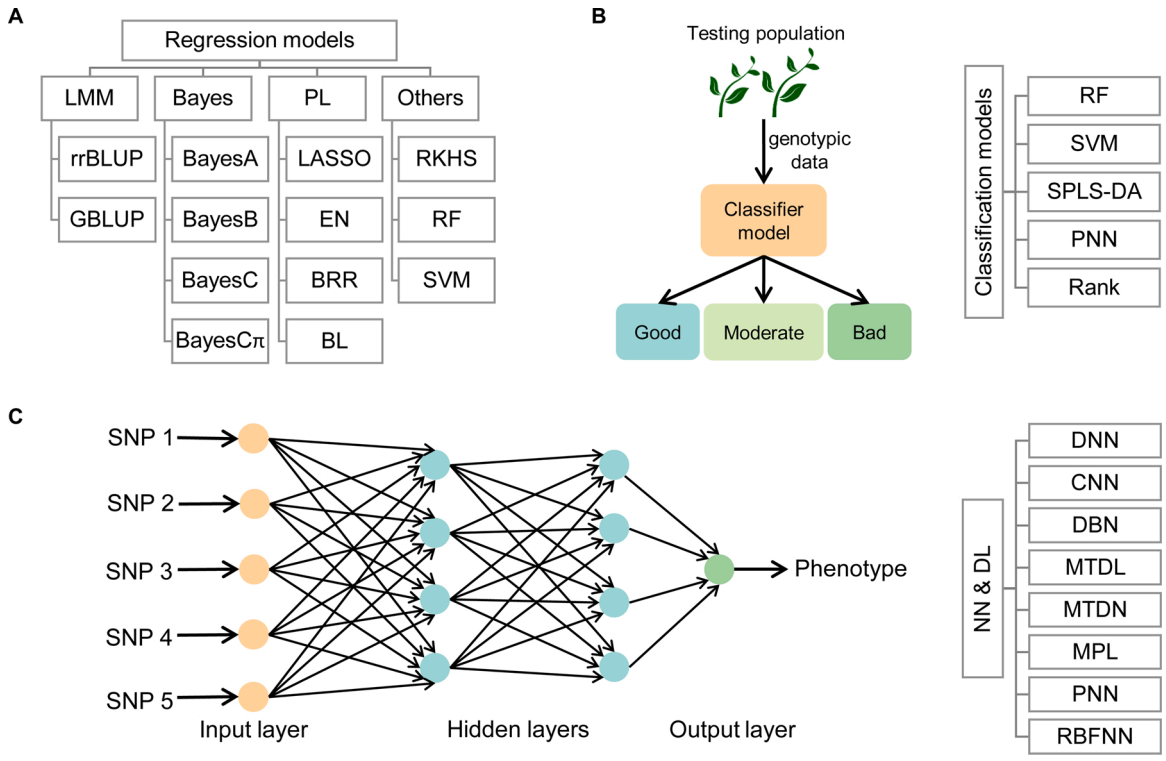

Fig. 2. Overview of statistical approaches for genomic selection. The genomic selection models can be classified into three groups based on statistical techniques used, namely (A) regression, (B) classification, and (C) deep learning. LMM: linear mixed model; rrBLUP: ridge regression best linear unbiased prediction; GBLUP: genomic best linear unbiased prediction; PL: penalized linear model; LASSO: least absolute shrinkage and selection operator; EN: elastic net; BRR: Bayesian ridge regression; BL: Bayesian LASSO; RKHS: reproducing kernels Hilbert spaces regression; RF: random forest; SVM: support vector regression; SPLS-DA: sparse partial least squares discriminant analysis; NN: neural network; DL: deep learning; DNN: deep neural network; CNN: convolutional neural network; DBN: deep belief network; MTDL: multi-trait deep learning; MTDN: multivariate poisson deep learning; MPL: multilayer perceptron; PNN: probabilistic neural network; RBFNN: radial basis function neural networks.

markers) is much larger than the number of observations (i.e. genotypes in the training population) for a phenotypic trait that is modeled. Therefore, ordinary least squares (OLS) regression cannot be used, as it can only be robustly applied in situations in which the number of observations is larger than the number of predictors. Such challenge is overcome by introducing regularization techniques, e.g. the classical ridge regression (Ogutu et al., 2012), which does not suffer from numerical instabilities and allow estimation of the regression coefficients. We note that in the classical ridge regression all predictors are treated as fixed effects. Plant breeders rely on ridge regression best linear unbiased prediction (rrBLUP) (Meuwissen et al., 2001), which inherits the advantages of ridge regression but is cast as a mixed linear model, given by

$$y = X\beta + Zu + e$$

where y denotes the phenotype as response, X is the fixed effect design matrix, β is the fixed effect, Z is a matrix of genetic markers, u is the marker effect, treated as a random effect, and e is the residual error. The variance of y is

$$V = ZZ^T\sigma_u^2 + I\sigma_e^2$$

where σ_u^2 is the marker effect variance and σ_e^2 is the residual error variance. In contrast to the classical ridge regression, the penalization factor is fixed to σ_e^2/σ_u^2, i.e. the ratio between the two random effect variance components. This approach can shrink all effects toward zero equally, corresponding to the assumption that all markers have a common variance. The equivalent model, termed genomic best linear unbiased prediction (GBLUP), estimates the kinship matrix from genomic markers to represent the pedigree information, and in turn estimates GEBV in a similar mixed linear model (Habier et al., 2013).

Another group of regularized regressions applied in GS includes the least absolute shrinkage and selection operator (LASSO) (Usai et al., 2009), which performs simultaneous feature selection and parameter estimations, and elastic net (Ogutu et al., 2012), which captures the trade-off between feature selection of LASSO and the shrinkage of the ridge regression. Bayesian model variants, such as Bayesian ridge regression (BRR) (Gianola et al., 2003) and Bayesian LASSO (BL) (de los Campos et al., 2009) have also been used in GS. Finally, other Bayesian regression models are also widely used in GS, including BayesA (Meuwissen et al., 2001), BayesB (Meuwissen et al., 2001), BayesC (Habier et al., 2011), and BayesCπ (Habier et al., 2011), corresponding to different assumptions about the maker effects and their distributions. The BayesA model assumes all SNPs with a non-zero effect and different variances. The BayesB model assumes that only percentage of SNPs are non-zero effect and the variances follow a mixture distribution. In contrast, BayesC assumes all SNPs to have the same variance, while in the BayesCπ model the prior probability of zero-effect SNPs is treated as unknown.

All the aforementioned regression approaches assume a linear relationship between genome-wide markers and phenotypic traits. Therefore, they capture additive effects of genome-wide markers and are suitable for prediction of phenotypic traits in inbreed populations. However, in hybrid populations, non-additive effects, i.e. dominance and epistasis, are highly relevant and must also be considered (Schnable and Springer, 2013). One way to address this issue is to consider extended GBLUP, in which pairwise marker interactions are considered (Su et al., 2012; Jiang and Reif, 2015; Martini et al., 2017). To this end, one can use different machine learning approaches, including reproducing kernels Hilbert spaces regression (RKHS) (Gianola and van Kaam, 2008; de los Campos et al., 2010), random forest (RF) (Holliday

et al., 2012), and support vector machine (SVM) regression (Long et al., 2011), shown to lead to better performance in different scenarios (Desta and Ortiz, 2014). The machine learning approaches and statistical models used in GS are summarized in Fig. 2A.

To compare the model performance of different GS models, one usually uses the measure of predictability. It is determined by k-fold cross-validation, whereby the population is partitioned into k (non-overlapping) groups, called folds, of genotypes. One fold corresponds to the part of the population that is treated as a cross-validation testing population. The remaining (k-1) folds are used as the cross-validation training population. Therefore, as explained above, the cross-validation training population is used to learn a model of interest, i.e. to determine its parameters. The cross-validation testing population is used to test the performance of the model as assessed by the predictability. The predictability is then given by the Pearson correlation coefficient between the predicted phenotypic value, from the model learned on the cross-validation training population, and the measured phenotypic value in the cross-validation testing population.

The aforementioned regression approaches are suitable for GS with single traits. Regression-based approaches for single trait GS has been applied in different species (e.g. alfalfa, barley, cassava, tomato, coffee, maize, pea, rice, sorghum, soybean, sugar cane, and wheat) to model biomass, yield, flowering, stress and disease resistance related traits in populations of varying sizes (Table 1). Notably, rrBLUP and GBLUP are the most widely used approaches, followed by Bayesian models (Table 1). Depending on the size of the training population, the type of markers used and their number, the predictability of a trait differs even for a single trait between studies (Table 1).

2.2. Classification-based models for genomic selection

The phenotypic traits of interests in crop breeding are not always measured on a continuous scale but are scored as categorical values. For instance, tomato fruit color can be scored as red, yellow or green, and later transformed into discrete numerical values. Another group of phenotypes can be scored on an ordinal scale with several points (e.g. disease resistance). For such phenotypic traits, classification-based models are more suitable in GS. Classification-based approaches can also be used in cases which continuous phenotypes are discretized into few values or groups (e.g. good, moderate, and bad performing genotypes) (Fig. 2B). In this context, the breeder is interested in identifying the genotypes with good focal phenotypes that can be used in the next breeding cycle.

Classification-based approaches to GS are sensitive to the way the values are discretized. One study suggested that the performance is improved if discretization into three rather than two groups is used, since the resulting groups may be expected to be more balanced (González-Camacho et al., 2016). Another aspect is the ratio of sample numbers between groups. One study predicting the grain yield in maize indicated that the ratio of 30%–70% of samples in the good and bad groups, respectively, in the training set had the best predictability in comparison to other scenarios (Qiu et al., 2016). In scenarios with highly

Table 1
Representatives of single-trait genomic selection studies in plants. The species, type of phenotypic trait, population size, and number of markers used are included. The summary indicates that the same trait can have different predictabilities within and between species, highlighting the importance of factors that affect the performance of GS models.

Species	Trait	Population size	Marker number	Predictability	Model*	Reference
Alfalfa	forage quality traits	154	11,450	0.01−0.4	rrBLUP, Bayesian models, SVM	Biazzi et al., 2017
Apple	fruit texture traits	537	8,294	0.01−0.81	rrBLUP	Roth et al., 2020
Avocado	fruit, seed traits	160	2,663	~0.025-~0.22	ridge regression	He et al., 2016
Banana	morphological, fruit traits	307	10,807	0.15−0.72	Bayesian models, RKHS	Nyine et al., 2018
Barley	yield, disease traits	1,317	4,056	0.07−0.49	rrBLUP, BL	Tsai et al., 2020
Brassica napus	yield, quanlity, other traits	475	24,403	0.29−0.81	rrBLUP	Jan et al., 2016
Cassava	yield, disease, other traits	411−899	155,871	0−0.68	GBLUP, Bayesian models, RKHS, RF	Wolfe et al., 2017
Tomato	quality, metabolite traits	163	5,995	0.05−0.81	rrBLUP	Duangjit et al., 2016
Arabidopsis	amino acid related traits	313	199,452	0.08−0.47	GBLUP	Turner-Hissong et al., 2020
Coffea arabica	yield, morphological, other traits	195	20,477	0.06−0.61	GBLUP	Sousa et al., 2019
Grape	yield, morphological, other traits	143	243	0−0.42	rrBLUP, BL	Viana et al., 2016
Kersting's groundnut	yield, morphological traits	281	493	0.02−0.79	rrBLUP	Akohoue et al., 2020
Maize	yield, morphological traits	291−435	8,271−37,803	0.36−0.77	GBLUP	Liu et al., 2018
Miscanthus	yield-related traits	568	46,177	0.39−0.65	rrBLUP	Clark et al., 2019
Oat	yield-related traits	194	545	0.12−0.87	rrBLUP	Mellers et al., 2020
Pea	yield, flowering, other traits	306	6,058	0.38−0.75	BL	Annicchiarico et al., 2019
Pearl millet	grain, flowering trait, plant height	357	32,463	~0.2-~0.8	rrBLUP	Liang et al., 2018
Perennial ryegrass	yield, quanlity, disease traits	1,918	1,447,122	0.27−0.68	GBLUP	Fè et al., 2016
Pine	wood density, plant height, other traits	1429	51,213	0.27−0.83	GBLUP, Bayesian models	Ukrainetz and Mansfield, 2020
Rice	yield-related traits	278	1,619	0.09−0.69	GBLUP, LASSO, SSVS	Xu et al., 2014
Rye	plant height, heading stage, disease	465	7,728	~0.72-~0.86	rrBLUP	Gaikpa et al., 2020
Sorghum	yield, quality, other traits	200	258,220	0.61−0.85	rrBLUP, Bayesian models	de Oliveira et al., 2018
Soybean	yield-related traits	324	4,947	0.49−0.83	Bayesian model	Matei et al., 2018
Strawberry	yield, fruit quanlity traits	1,628	17,479	~0.07-~0.59	GBLUP, Bayesian models, RKHS	Gezan et al., 2017
Sugar cane	yield, sugar content	467−1,146	47,531−57,675	~0.15-~0.47	GBLUP, Bayesian models, RKHS	Deomano et al., 2020
Switchgrass	yield, quality, other traits	483	19,342	0−0.88	GBLUP, PLS, SPLS, Bayesian model	Fiedler et al., 2018
Urochloa spp.	agronomical, nutritional traits	272	26,535	~0.11-~0.32	GBLUP	Matias et al., 2019
Wheat	yield, quality, other traits	5,520	3,075	~0.42-~0.71	rrBLUP, PLS, EN, RKHS, RF	Battenfield et al., 2016
Wheatgrass	biomass, agronomical traits	1,126	3,883	0.46−0.67	rrBLUP, Bayesian models, RKHS, RF	Zhang et al., 2016

* Stochastic search variable selection (SSVS), partial least squares regression (PLS), sparse partial least squares regression (SPLS).

unbalanced classes, well-established undersampling and upweighting techniques can be used (Anand et al., 2010) or creation of artificial samples from the minority class can be considered (Blagus and Lusa, 2013), although these have not yet been applied in the setting of GS.

Only few models reported using the classification-based models for GS in crop breeding (Fig. 2B). For instance, random forest (RF) classification and support vector machine (SVM) classification were applied in maize and wheat populations and demonstrated that the classification-based models can increase the predictability of grain yield, flowering time, and disease resistance traits compared to the regression-based models (Ornella et al., 2014). Moreover, a similar approach trying to predict the rank of genotypes, as results the RankSVM approach performed better than RKHS regression using the evaluation measures of normalized discounted cumulative gain (NDCG) (Blondel et al., 2015). Therefore, the classification-based GS approaches represent a valuable alternative to regression-based approaches of GS in plant breeding.

When comparing the predictability from cross-validations of several models, most studies use the correlation coefficient to assess the performance of prediction in the continuous phenotype scenarios. For the classification-based models, the correlation coefficient is not suitable as it is sensitive to outliers at the tails of the distribution—which are of high interest in GS (González-Camacho et al., 2018). To correctly evaluate the predictability of classification-based GS, one could simply use the percentage of cases correctly classified (PCCC) based on the counts (Montesinos-López et al., 2019a). Ornella et al. (2014) proposed other coefficients, i.e. Cohen's kappa coefficient (κ) and relative efficiency (RE) to assess the predictability in classification models. The kappa coefficient is defined as

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed fraction of correctly prediction and P_e is the expected correctly prediction. The RE coefficient is defined as

$$RE = \frac{\mu_\alpha^{'} - \mu}{\mu_\alpha - \mu}$$

where μ_α and $\mu_\alpha^{'}$ are the mean phenotypic value of the top α (α is usually 15 % or 30 %) genotypes based on the observation and prediction, respectively, and μ is the grand mean value (Qiu et al., 2016). In addition, González-Camacho et al. (2016) proposed to use the area under the receiver operating characteristic curve (AUC), in which receiver operating characteristic (ROC) curve is a plot of true positive rate against the false positive rate, as a coefficient to assess the predictability. Since there is no established statistic to measure the performance of classification-based GS models, presentation of multiple measures is of value to allow for objective assessment of different classification models for GS.

2.3. Deep learning models for genomic selection

Neural network and deep learning are algorithms that emulate the functionality of biological neural networks. The basic structure of neural network contains at least three layers representing the input, hidden, and output layers (Fig. 2C). In the context of GS, the input is the genotypic data, i.e. SNPs, and the output is the predicted phenotypic trait. The advantage of using deep learning to investigate the genomic selection is that deep learning approaches may capture complex higher-order interactions and achieve higher predictability (González-Camacho et al., 2012).

Many deep learning approaches have been applied in GS (Fig. 2C), resulting in the observation that the convolutional neural network (CNN) seems to perform best in predicting the phenotypes from genotypic data in crops (Pérez-Enciso and Zingaretti, 2019). Ma et al. (2018) developed the method DeepGS, based on CNN, and showed that the performance of DeepGS is complementary with other statistical models, e.g. rrBLUP, to predict grain-related traits in a wheat population. Another study predicted five traits in soybean using two parallel CNN streams, so-called dual-stream CNN, and showed improved performance in comparison to the single-stream CNN (Liu et al., 2019). Deep learning approaches are also applied to predict phenotypes in classification scenarios (González-Camacho, 2016; Montesinos-López et al., 2019a). The probabilistic neural network (PNN), which predict the probability of genotypes belonging to the good or bad groups, was tested in maize and wheat populations and showed better performance then multi-layer perceptron (MLP) models, which concluded that the PNN is an encouraging approach of genomic selection in crop breeding (González-Camacho, 2016).

3. Multi-trait models for genomic selection

During the scoring of phenotypic traits, the breeders measure several traits of interest at the same time, since crop yield, as a key trait of interest, is affected by other multiple factors and traits. In addition, crop breeders are interested in selecting elite genotypes that not only result in high yield, but also show high nutritional value and disease resistance. Therefore, artificial selection is based on multiple phenotypic traits instead of one focal phenotype. The GS models discussed above are suitable in the case of a single trait and are referred to as single-trait genomic selection (STGS) models. In this section, we review and discuss the multi-trait genomic selection (MTGS) models that consider the correlation structure of multiple traits.

The first multi-trait genomic selection model in plants was proposed by Jia and Jannink (2012) and was applied in a loblolly pine population with two disease resistant traits. The GBLUP and Bayesian models were applied to both STGS and MTGS scenarios, showing that the predictabilities are similar in the two scenarios. However, the analyses on synthetic data sets indicated that the predictability of phenotype with low-heritability can be improved by using MTGS model. In the last five years, several MTGS studies were reported in major crops, e.g. maize, rice, and wheat (Table 2). Most studies applied the multivariate linear mixed model, as an extension of the single-trait GBLUP (Lyra et al., 2017; Kristensen et al., 2019; Ortiz et al., 2019). At this point, it is worth noting that the MTGS models are often transformed to the STGS models by appending the traits to each other (referred to as vectorization). A similar model proposed by Wang et al. (2017) incorporated a multivariate relationship matrix which is calculated using the multiple traits. The applications in a rice hybrids population confirmed the conclusion that MTGS can improve the predictability for low-heritability traits. Another attempt relied on the multiple output regression utilizing the correlations between outputs to improve the predictability (He et al., 2016). Deep learning approaches have also been applied in the multi-trait scenario, for instance, a study in wheat population showed that the predictability of multi-trait deep learning is similar to that of the single-trait GBLUP model, but better when compared with the univariate deep learning model (Montesinos-López et al., 2019a, c).

Before the genomic era, the selection considering multiple traits was based on selection index (SI), which represents a weighted linear combination of multiple traits (Schulthess et al., 2016; Habyarimana et al., 2020). The weights assigned to multiple traits are related to their economic importance. The direct integration of GS and SI can be done by using the SI calculated from multiple traits as a phenotypic trait in STGS. One study predicting three biofuel-related traits in sorghum showed that the predictability of SI using the three traits is better than that based on a single trait (Habyarimana et al., 2020). In contrast, following the indirect integration one would first predict multiple traits based on the MTGS models discussed above, and would then calculate the SI as a final assessment for selection using the weighted linear combination. Based on a study in rye, it was suggested that the direct integration of GS and SI showed better predictability than the indirect integration in the scenario of balanced phenotypic data (Schulthess et al., 2016). Another extension is to use the two SIs as two phenotypes in MTGS and to attempt to

Table 2
Overview of the genomic selection studies using multi-trait models and the comparison with single-trait models. The table includes the number and type of traits modeled, along with the population size and number of markers used in the existing studies with multi-trait GS in eight different species. Values for the predictability of single- and multi-trait models are also provided, along with the GS models used, to facilitate comparison between the approaches.

Species	Trait	Population size	Marker number	Multi-trait number	Multi-trait predictability	Multi-trait model*	Single-trait predictability	Single-trait model*	Reference
Pine	disease resistance traits	769	4,755	2	~0.27-~0.31	GBLUP, Bayesian models	~0.26-~0.31	GBLUP, Bayesian models	Jia and Jannink, 2012
Rye	grain yield, protein content +ear weight	201	394	2 3	0.25–0.59 0.25–0.59	GBLUP	0.23–0.59	GBLUP	Schulthess et al., 2016
Avocado	fruit, seed traits	160	2,663	8 8	~0.05-~0.52 ~0.044-~0.55	multitask learning multiple output regression	~0.025-~0.22 ~0.045-~0.51	ridge regression	He et al., 2016
Maize	grain yield, plant height	738 452	146,670 52,700	2 2	0.42–0.56 0.71–0.81	GBLUP, GK	0.39–0.55 0.7–0.8	GBLUP, RKHS	Lyra et al., 2017
Rice	grain, morphological traits	575	3,299,150	2	0.47–0.88	MV-ADV	0.39–0.88	GBLUP	Wang et al., 2017
Sorghum	biomass yield, plant height, moisture	453	59,264	2	~0.38-~0.41	GBLUP	0.4	GBLUP	Fernandes et al., 2018
Wheat	baking quality traits	495	6,655	2–4	~0.23-~0.43	Bayesian multivariate model	~0.24-~0.43	BRR	Lado et al., 2018
Wheat	quality traits	1,152	11,058	2 or 4	0.5–0.65	GBLUP, Bayesian model	0.5–0.65	GBLUP, Bayesian model	Kristensen et al., 2019
Sorghum	yield, plant height, stay-green, flowering time	2,645	4,781	2–4	~0.35-~0.47	GBLUP	~0.36-~0.48	GBLUP	Ortiz et al., 2019
Wheat	grain yield, days to heading, plant height	270	14,163	3	~0.03-~0.13**	MTDL	~0.02-~0.14**	GBLUP	Montesinos-López et al., 2019c
Barley	yield, mating quanlity traits	980	6,482	4–8	~0.1-~0.3	Bayesian multivariate model	~0.1-~0.3	BRR	Bhatta et al., 2020
Sorghum	biofuel traits	369	61,976	3	0.59	Bayesian model with selection index	0.36–0.55	Bayesian model	Habyarimana et al., 2020

* Gaussian kernel regression (GK), regression with multivariate relationship matrix (MV-ADV). ** Report as the mean arctangent absolute percentage error.

combine the traits twice in the models, i.e. first by combining traits via SI, and then combining SI in the MTGS (Lyra et al., 2017). Nevertheless, it is worth pointing out that the weights for the traits in SI, as defined above, are fixed. However, the weight may not be optimal due to changes in each breeding cycle and also they do not account for nonlinear relationships. The algorithm of look-ahead selection (LAS), which maximizes one major trait while constraining other traits in the specified ranges, was applied in the GS with SI in maize (Moeinizade et al., 2020).

The predictability of multi-trait GS models are also determined in a cross-validation strategy, i.e. use the multiple traits in the training population with genotypic data to predict one or more focal trait in the testing population. Among the traits measured in the population, some traits are less expensive and labor-intensive to measure than others. The usage of these traits as secondary ones to predict the more complex and labor-intensive traits is referred to as trait-assisted GS (Fernandes et al., 2018), with the constraint that the secondary traits are correlated to the focal trait to be predicted. The cross-validation to predict focal trait with only genotypic data of testing population is referred to as CV1, while cross-validation by adding the phenotypic data of secondary traits in testing population to predict the focal trait is referred to as CV2 in the trait-assisted GS (Fig. 3A). Several studies in wheat (Lado et al., 2018; Kristensen et al., 2019), sorghum (Fernandes et al., 2018; Ortiz et al., 2019), and barley (Bhatta et al., 2020) reported that the predictability was considerably increased under the scenario of CV2 in the trait-assisted GS. One study in sorghum predicted biomass yield by using plant height as secondary trait in the testing population, demonstrating that trait-assisted GS leads to an increase in predictability by up to 50 % (Fernandes et al., 2018). It should be noted that the predictability assessment in CV2 scenario may be biased because of using the data of secondary traits in the testing population. Runcie and Cheng (2019) suggested three approaches to solve this issue, including the CV2* scenario which use the phenotypic data of the close relatives of the genotypes in the testing population.

When MTGS is applied, one critical question is how many traits should be involved in the models. A study showed that adding the third trait of ear weight to predict the grain yield in rye led to no improvement in comparison to the scenario when two traits were used in GS (Schulthess et al., 2016). However, another study in wheat showed that the predictability of GS that uses three traits increased up to 14 % compared to GS with two traits only (Lado et al., 2018). Looking at the combinations between two to four traits in sorghum, another study showed the best prediction is via three trait GS (Ortiz et al., 2019). The number of traits that should be included to observe the increase in GS model performance still lacks thorough treatment, as the correlation structure of the traits may have to be constrained to result in the increase of model performance. The reviewed studies indicate that inclusion of more traits may not be the best approach to follow, since it leads to computational complexity and convergence issues in solving the underlying mixed-model in the MTGS models (Schulthess et al., 2016). However, future efforts in multi-trait GS can be focused on overcoming these issues, while ensuring the increase in predictability of low-heritability traits.

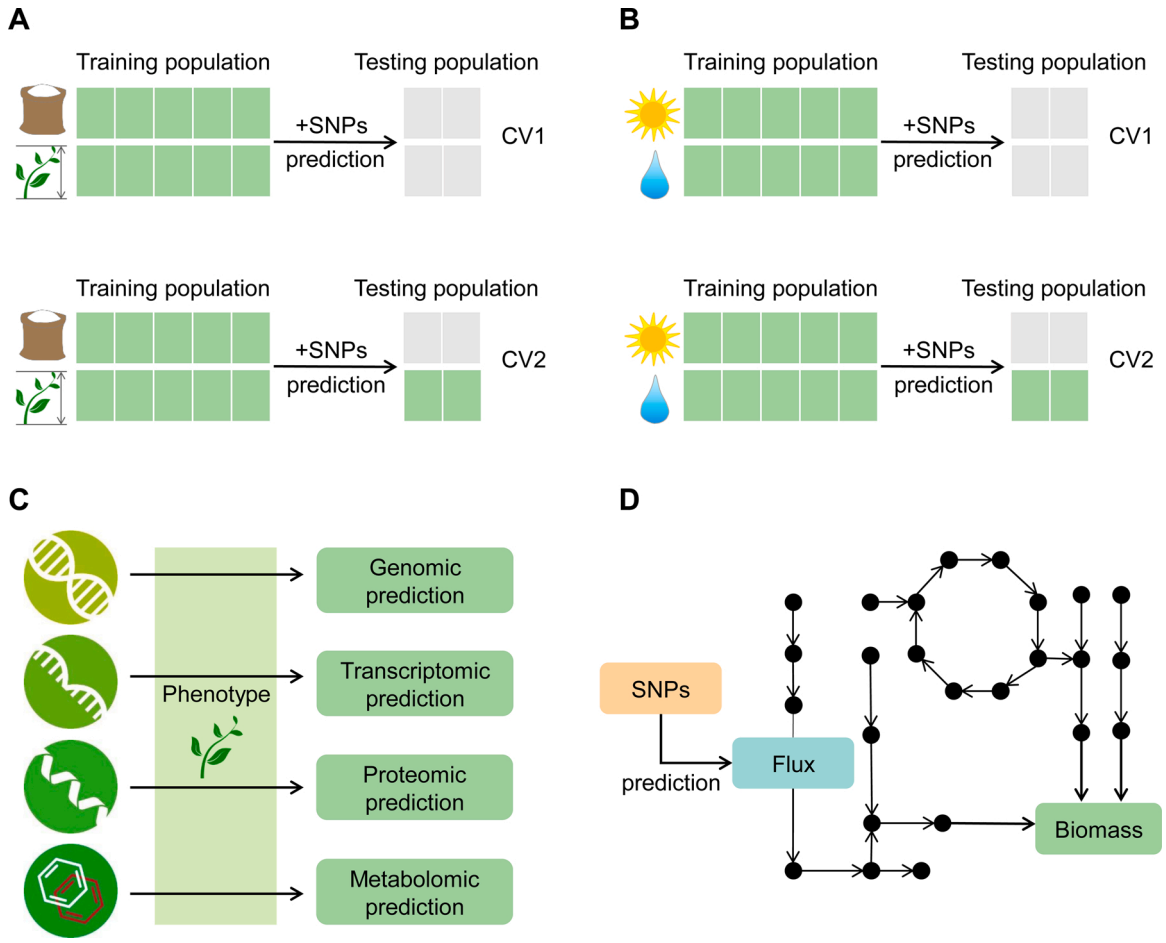

Fig. 3. Perspective directions of genomic selection based on the multiple trait model. The measured phenotypes used for prediction are marked in green, and the focal traits to be predicted are marked in gray in panels (**A**) and (**C**). The cross-validation to predict focal trait with only genotypic data of testing genotypes is referred to as CV1, while cross-validation using the phenotypic data of secondary traits (A) or environments (C) of testing genotypes to predict the focal trait of testing genotypes is referred to as CV2. Panel (**B**) illustrates multi-omics prediction, while Panel (**D**) shows the netGS approach in which the fluxes are treated as traits for genomic selection and are estimated by using constraint-based modelling with a genome-scale metabolic network.

4. Integration of intermediate phenotypes

GS predicts phenotypes based on genotypic data from various types of genomic markers. With the advent of high-throughput technologies, multiple omics data about the intermediate phenotypic traits, e.g. metabolite, protein, and gene expression levels, can also be used as predictors along with genetic markers in building models akin to GS. Hence, the resulting approach is termed multi-omics selection (MOS) (Fig. 3B).

There have been several studies conducted around the idea of MOS (Table 3). For instance, one study applied gene expression levels instead of genomic data to predict the grain yield in maize hybrids (Fu et al., 2012). By using the multiple linear regression, partial least squares regression (PLS), and SVM regression, the predictability can reach up to ~0.3 to ~0.98. Two studies reported the performance of MOS with metabolite levels in maize (de Abreu e Lima et al., 2018) and rice (Dan et al., 2020). By using 136 metabolites in maize showed the classification-based models can predict hybrid performance and discriminate good-performing hybrids. The larger metabolic profile data reported in rice showed that the predictability of yield heterosis using 3,746 metabolites was 0.24 to 0.61. The findings above indicate that MOS with intermediate phenotypic traits is a contending alternative in predictions of focal traits.

To further compare the GS predictability between genomic data and intermediate phenotypes, another study applied both genomic and metabolic data to predict seven biomass and bioenergy-related traits in maize (Riedelsheimer et al., 2012). The results demonstrated that the predictability with 130 metabolite levels is decreased, on average, by only 6.7 % in comparison to SNPs, although the number of metabolites is ~300-fold smaller than the number of SNPs used. However, a similar study in barley showed that by using the metabolite levels as predictors, the predictability decreased by 50.9 %, on average, for eight yield and flowering related traits (Gemmer et al., 2020). While combining both genomic and metabolic data in the GS model, the predictability does not increase in comparison to using only SNP data.

More comprehensive comparative studies utilized three omics data sets, i.e. genome, transcriptome, and metabolome, to predict the phenotypes and compare the predictability with one or two of these omics data sets. The study to predict four yield-related traits in hybrid rice showed that genomic data are the best predictors for high heritability traits (Xu et al., 2016). However, the predictability of yield almost doubled using the metabolite levels compared with using only SNPs. To expand the usage of the same data set focus on inbred parents, Hu et al. (2019) developed the multilayered least absolute shrinkage and selection operator (MLLASSO) model which learns three layers supervised by transcriptomic and metabolomic data. The MLLASSO model was shown to improve the predictability of yield by combining the three omics data sets comparing to the genomic data alone. A similar study in maize confirmed the conclusion, showing an 8.7 % increase when combining the three omics data sets (Guo et al., 2016). This study applied linear mixed model with the interaction terms between three omics data sets, and demonstrated that the consideration of interactions in the model

Table 3
Overview of the genomic selection studies based on intermediate phenotypes. The table provides a summary of predictabilities of agronomically relevant traits based on combination of different omics data sets on intermediate phenotypes. The population size, number of markers, and GS models used are also included along with the obtained predictability for the studied traits.

Omics	Species	Trait	Population size	Marker number	Predictability	Model*	Reference
Transcriptome	Maize	grain yield	98	10,810 transcripts	~0.3-~0.98	MLR, PLS, SVM, DB	Fu et al., 2012
Metabolome	Maize	biomass	328	136 metabolites	~0.29-~0.33**	PLS-DA, SVM, RF	de Abreu e Lima et al., 2018
Genome, Metabolome	Rice	yield	398	3,746 metabolites	0.24–0.61	PLS-DA	Dan et al., 2020
	Maize	biomass, bioenergy related traits	570	38,019 SNPs 130 metabolites	0.69–0.80 0.48–0.67	rrBLUP	Riedelsheimer et al., 2012
	Barley	yield, flowering related traits	1,307	33,005 SNPs 122 metabolites SNPs + metabolites	0.74–0.93 0.26–0.59 0.74–0.93	rrBLUP, Bayes model	Gemmer et al., 2020
	Rice	yield related traits	278	1,619 genetic bins 24,994 transcripts 683 leaf and 317 seed metabolites	~0.1-~0.7 ~0.1-~0.7 ~0.05-~0.6	GBLUP, LASSO, SSVS, PLS, SVM with radial basis and polynomial kernel function	Xu et al., 2016
	Rice	yield related traits	210	1,619 genetic bins genetic bins+24,994 transcripts genetic bins + transcripts+1,000 metabolites	0.16–0.76 0.25–0.74 0.23–0.72	MLLASSO	Hu et al., 2019
	Maize	yield, morphological, maturity traits	368	646,720 SNPs 28,850 transcripts 234 metabolites SNPs + transcripts SNPs + metabolites SNPs + transcripts + metabolites	0.27–0.6 0.18–0.57 0.04–0.58 0.26–0.61 0.25–0.61 0.24–0.63	GBLUP	Guo et al., 2016
	Maize	dry matter, quality traits	617	21,565 SNPs 1,323 transcripts 92 leaf and 283 root metabolites SNPs + transcripts	~0.2-~0.6 ~0.2-~0.75 ~0-~0.6 ~0.25-~0.75	LMM	Westhues et al., 2017
	Maize	dry matter traits	550	37,392 SNPs 300 mRNA transcripts 10,736 sRNA transcripts 148 root metabolites SNPs + mRNA transcripts SNPs + sRNA transcripts SNPs + metabolites mRNA + sRNA transcripts mRNA transcripts + metabolites	0.57–0.71 0.46–0.82 0.47–0.71 0.42–0.74 0.56–0.81 0.56–0.73 0.56–0.78 0.51–0.82 0.49–0.8	LMM	Schrag et al., 2018

* Multiple linear regression, forward selection (MLR), partial least squares regression (PLS), transcriptome-based distances (DB), partial least squares discriminant analysis (PLS-DA), stochastic search variable selection (SSVS), multilayered least absolute shrinkage and selection operator (MLLASSO). ** Report as the kappa coefficients.

does not improve the predictability compared to the one without interactions. Schrag et al. (2018) integrated data on sRNA and showed that the sRNA performed relatively poorly for two yield-related traits comparing to transcriptomic and genomic data. The GS improvement by integrating the intermediate phenotypic traits is likely due to the fact that intermediate phenotypes capture the epistasis and interactions between genomic markers, but also the effect of the environment. It is therefore expected that GS with intermediate phenotypes will further be explored in improving the GS predictability and in unraveling physiological mechanisms for agronomically important traits.

The levels of intermediate phenotypes capture the joint effect of the genotype and the environment, and in contrast to the invariant genomic data, change due to different environmental cues across sampled tissues and developmental stages; all these factors can affect the GS predictability. For instance, a study in rice using 683 metabolites from leaves and 317 metabolites from seeds showed that the predictability of yield-related traits is, on average, 0.28 from leaves and 0.22 from seeds (Xu et al., 2016). The comparison to a random selection scenario indicated that the difference in predictability between the two tissues is partly related to tissue-specific metabolic profiles, rather than the different number of metabolites measured. Another study in maize confirmed the finding that the predictability using root metabolites is, on average, better than the one using leaf metabolites (Westhues et al., 2017). Concerning the sampling time, a study in barley indicates that the sampling during a young and more homogeneous plant stage trends to result in better predictability (Gemmer et al., 2020).

One limitation of GS with intermediate phenotypes is the resource-intensive genome-scale measurements of metabolite, protein, and gene expression levels required in comparison to the decreasing costs of sequencing technology employed in generating genomic data. While the reviewed studies show that there is interest to measure and investigate these intermediate phenotypes, the scale of measured intermediate phenotypes is much smaller in comparison to the data obtained from genotyping technologies. Westhues et al. (2019) recently proposed an approach to impute the intermediate phenotypes based on the genomic data or pedigree information, i.e. to predict the missing value in incomplete data set of intermediate phenotypes and use them in GS. The study attempted a single-step framework to impute the transcriptomic data in maize inbred and hybrid population, demonstrating that by increasing the sample number of imputed transcription data, the predictability of GS models can be improved. To date, proteomics data have not yet been used as intermediate phenotypes in GS, largely due to the

smaller coverage in comparison to transcriptomic and metabolomic data as well as the technical difficulties involved in obtaining quantitative data.

5. Factors that affect the genomic selection performance

Several factors that affect the predictability of GS have already been discussed (Nakaya and Isobe, 2012). As noted above, the predictability of GS ultimately depends on the factors that affect the quality of the training models, but also the characteristics of the data sets used for training and testing. Therefore, the number of observations and markers used are expected to have an effect on predictability. For instance, the number of markers required in GS is determined by the LD decay in one species. The species with a shorter LD decay (e.g. maize) need more markers than others to ensure that considerable genetic variance, especially, from minor effect loci, is included. Simulation studies have suggested that a larger number of markers lead to improvement of the GS accuracy (Solberg et al., 2008). Similarly, increasing the number of samples in training population can achieve better GS accuracy (Heffner et al., 2011). The ultimate proof of the usefulness of GS models is their performance on breeding populations. Therefore, most importantly, the relationship between training and breeding population has a large effect on prediction. This is the case since GS prediction, in contrast to MAS, is based on genome-wide markers, rather than only markers that are tightly linked to target genes. GS in natural populations faces the same problems caused by population stratification as genome-wide association studies. A study suggested that combining several subpopulations as training population can achieve higher prediction accuracy than using single subpopulations (Technow et al., 2013).

6. Integration of environmental variables

In plant breeding schemes, breeders usually assess the genotype performance in multiple environment trials to check the robustness of the phenotype in various environments. In the context of GS, most studies used the averaged value across environments or pre-analyze phenotypes in linear mixed model (e.g. BLUP) to extract only genetic effect as the trait value. One study showed that usage of the average trait values of a phenotype over multiple environments to predict the performance, comparing to in a single environment, can achieve high predictability (Nyine et al., 2018). Therefore, incorporation of data from many environments could make the GS models more robust. However, the changes in the phenotype due to environmental alterations may be different from genotype to genotype, leading to the concept of genotype-by-environment (G × E) interaction (Kang, 1997). Analysis of G × E effect could help to select stable genotypes not only across environments, but also in a specific breeding environment.

Burgueño et al. (2011) extended the classical GS model to the scenario of multiple environments and showed improvement of predictability in comparison to a single environment model in wheat. The findings were further confirmed that the predictability of the model including G × E is higher than that of the within environment or across environment models neglecting G × E effects (Lopez-Cruz et al., 2015; Crossa et al., 2016). Thus, the GS model including environmental and G × E effect can increase the predictability of a trait. From a methodological point of view, GS in multiple environment is similar to modelling the multiple trait scenario, since the one trait in multiple environment can be seen as different traits in a multi-trait GS model. As a result, the multi-trait GS models discussed above can be directly applied to the multiple environment analysis discussed here. The cross validation to assess the performance of genotypes can be in two scenarios, like for multi-trait GS: evaluate the performance of new genotypes in the studied environments (CV1), or evaluate the performance with already measured phenotypes in some of the environments (CV2) to mimic the situation of field trials with missing phenotypic value in the new environment (Fig. 3C) (Burgueño et al., 2012).

Most studies incorporating environmental and G × E effects are extended from the GBLUP model by adding the term of environment and G × E as well as the pedigree and pedigree-by-environment effect. With the available genomic marker data, the classical definition of G × E effect can be extended to marker-by-environment (M × E) effects. The latter has been used in distinguishing markers whose effects vary across the environment, leading to the identification of environment-robust markers and environment-specific markers (Schulz-Streeck et al., 2013). In the assumption of the genetic value as the linear combinations of marker effects, the G × E effect is equivalent to the M × E effect. Under non-linearity assumption, one study in wheat showed that the nonlinear RKHS model with the G × E term led up to 68 % improvement of predictability in comparison to single environment GS (Cuevas et al., 2016). Moreover, applying the deep learning approach to investigate the G × E in maize indicated that the GS predictability was better without G × E than with G × E interaction (Montesinos-López, 2019c).

The G × E effect discussed above does not integrate the weather condition and soil data. A more comprehensive model to predict the yield of genotypes in the scenario of climate changes is crop growth model. Combining the GS with crop growth model and data on weather conditions could improve the predictability that one study in wheat reported the predictability of yield trait in new environments with observed weather condition data increased on average by 11.1 % (Heslot et al., 2014). To further extend the prediction in the new environment without in-season data, one study in barley tried to predict the yield in combination with the historical weather data (Gillberg et al., 2019). The results showed the G × E model performed better than the one without, and the soil type and daily rain are important factors to be considered in the interactions with genotypes. A recent study predicted the hybrid maize yield using the environmental data, i.e. intercepted radiation, soil water potential, and night temperature, with the predicted phenology of each genotype to identify the genotype-specific characteristics (Millet et al., 2019). The results showed the advantages of using phenotyping platform and sensor data to improve the GS performance in comparison to the models discussed above. Moreover, deep learning was applied to GS with weather data, by first predicting the weather information from the historical weather data using a neural network, and then applying the predictions as environment input data in a deep learning model (Khaki and Wang, 2019). The results showed that the predicted performance is better than the panelized model and other neural network models. The GS model considering the G × E still represented a challenge, and more investigations are required to untangle the effects of environment on different genotypes and reliably predict their plasticity.

7. Software used in genomic selection

The majority of GS studies implemented in the statistical environment R, and many R packages have been developed for different models. The commonly used models in GS based on linear mixed model and the Bayesian regression model is rrBLUP and Bayesian models, which are implemented in the R package rrBLUP (Endelman, 2011) and BGLR (Pérez and de los Campos, 2014), respectively. The linear mixed model based approaches can also be solved via the mixed model package, e.g. lme4 (Bates et al., 2015) and ASReml (Butler et al., 2017), while the penalized model is implemented in the R package glmnet (Friedman et al., 2015). For the multiple trait GS, the R package MTGS (Genomic Selection using Multiple Traits) (Budhlakoti et al., 2019) and BMTME (Bayesian multi-trait and multi-environment) (Montesinos-López et al., 2019b) can be used in this advanced scenario. The R package BGGE (Bayesian Genomic Genotype × Environment) provides an implementation of G × E GS model and other models relevant in crop breeding (Granato et al., 2018). The existing design software and R packages used in GS are summarized in Table 4.

Table 4
Overview of software used in genomic selection in breeding. The table lists the name of the software application/package along with the type of models it implements.

Software	Description	Reference
rrBLUP	ridge regression best linear unbiased prediction	Endelman, 2011
BLR	Bayesian linear regression	Pérez et al., 2010
BGLR	Bayesian generlized linear regression	Pérez and de los Campos, 2014
GS3	genomic selection, Gibbs sampling, Gauss Seidel	Legarra et al., 2016
solGS	web-based tool for genomic selection	Tecle et al., 2014
GVCBLUP	genomic prediction of additive and dominance effects	Wang et al., 2014
GenoMatrix	pedigree-based genomic prediction	Nazarian and Gezan, 2016
ShinyGPAS	interactive genomic prediction	Morota, 2017
Gselection	feature select and genomic prediction	Majumdar et al., 2019
GenomicLand	genome-wide association study and genomic prediction	Azevedo et al., 2019
SeqBreed	python tool for genomic prediction	Pérez-Enciso et al., 2020
GVCHAP	genomic prediction using haplotypes and SNPs	Prakapenka et al., 2020
BWGS	genomic selection in wheat breeding programme	Charmet et al., 2020
MTGS	genomic selection using multiple traits	Budhlakoti et al., 2019
BMTME	Bayesian multi-trait and multi-environment model	Montesinos-López et al., 2019b
BGGE	Bayesian genomic genotype × environment interaction	Granato et al., 2018

8. Conclusions – towards integration of genotypic data in mechanistic models

The last two decades have witnessed tremendous advances in metabolomics as the newest molecular profiling technology whose outcomes can be readily integrated with data from other technologies (Riekeberg and Powers, 2017). These developments have led to the ability to discover and annotate previously unknown metabolites and to document the underlying biochemical reactions, along with enzymes, in which they are involved (Thiele and Palsson, 2010). In addition, the gathered data on metabolite levels have been used to determine their contribution to endpoint phenotypes (Fernandez et al., 2016). Despite these developments, the utility of metabolic phenotypes, along with other intermediate phenotypes, to predict focal phenotypes are less explored in comparison to use genomic data in crop breeding (see section 4).

The most important characteristic of metabolic phenotypes is that the levels of metabolites are shaped by the interrelated biochemical reactions comprising genome-scale metabolic networks. Such genome-scale metabolic networks are already available for the major crops (Küken and Nikoloski, 2019) and can be combined with GS to improve predictability of growth and metabolic traits that can be readily simulated in these networks by using constraint-based approaches. To this end, a recent study combined GS with genome-scale metabolic network in a population of global *Arabidopsis thaliana* accessions showed that an agronomically important phenotype, namely plant growth, can be predicted from the flux distribution of each genotype using the properties of genome-scale metabolic model (Tong et al., 2020). A novel approach based on constraint-based modeling was proposed to estimate the flux distribution using the measured biomass value and metabolic profiles of each genotype (Fig. 3D). This is the first study that treats the flux as a polygenic phenotype in the classic quantitative genetic studies, which open a new scientific field for further investigations. The study showed that the GS predictability can be increased by 32.6 % compared to the state-of-the-art methods in the optimal nitrogen (N) conditions. Therefore, it paves the way for combining knowledge of metabolic mechanisms with outcomes of machine learning approaches underlying genomic selection. In addition, the constraint-based modeling used in this study provides the possible solution to simulate changes in the environment. To this end, it was demonstrated that the GS predictability for growth under low N condition using the data only from optimal N condition is 51.4 % larger in comparison to classical approaches.

We recognize that further exploration of the coupling between mechanistic models, which allow simulation of phenotypes under changing environments, and machine learning models, that can integrate data on genomic markers (excluded from the mechanistic models), holds the promise to address the pressing problem of model transferability between environments. On the other hand, although this approach opens the black box of machine learning, it is limited to species with assembled and well-annotated genomes, based on which metabolic networks can be generated. Therefore, another opportunity is to invest in refining approaches that use kmers rather than genomic markers as predictors (Voichek and Weigel, 2020). Such an approach necessitates the development of statistical framework that allows the incorporation of other genetic variants (e.g. indels) in GS models, thus rendering the approach applicable to a wider class of crops. These parallel developments will have to account for the fact that traits in crops, like in other organisms, are highly interdependent. Therefore, future developments would aim at predicting an entire intermediate phenotype, like transcriptome, proteome, and metabolome, following multi-trait GS and GS accounting for environmental effects, which can, in turn, be used for further improvements in predictability of this important approach.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

Z.N. and H.T. acknowledge the funding from the European Union's Horizon 2020 research and innovation program, project PlantaSYST SGA-CSA No. 739582.

References

Ahmar, S., Gill, R.A., Jung, K.H., Faheem, A., Qasim, M.U., Mubeen, M., Zhou, W., 2020. Conventional and molecular techniques from simple breeding to speed breeding in crop plants: recent advances and future outlook. Int. J. Mol. Sci. 21, 2590. https://doi.org/10.3390/ijms21072590.

Akohoue, F., Achigan-Dako, E.G., Sneller, C., Deynze, A.Van, Sibiya, J., 2020. Genetic diversity, SNP-Trait associations and genomic selection accuracy in a west African collection of Kersting's groundnut [Macrotyloma geocarpum(Harms) Maréchal & Baudet]. PLoS One 15, e0234769. https://doi.org/10.1371/journal.pone.0234769.

Anand, A., Pugalenthi, G., Fogel, G.B., Suganthan, P.N., 2010. An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids 39, 1385–1391. https://doi.org/10.1007/s00726-010-0595-2.

Annicchiarico, P., Nazzicari, N., Pecetti, L., Romani, M., Russi, L., 2019. Pea genomic selection for Italian environments. BMC Genomics 20, 603. https://doi.org/10.1186/s12864-019-5920-x.

Azevedo, C.F., Nascimento, M., Fontes, V.C., e Silva, F.F., de Resende, M.D.V., Cruz, C.D., 2019. GenomicLand: software for genome-wide association studies and genomic prediction. Acta Sci. Agron. 41, e45361 https://doi.org/10.4025/actasciagron.v41i1.45361.

Bates, D., Mächler, M., Bolker, B.M., Walker, S.C., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 https://doi.org/10.18637/jss.v067.i01.

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

Bhatta, M., Gutierrez, L., Cammarota, L., Cardozo, F., Germán, S., Gómez-Guerrero, B., Pardo, M.F., Lanaro, V., Sayas, M., Castro, A.J., 2020. Multi-trait genomic prediction model increased the predictive ability for agronomic and malting quality traits in barley (Hordeum vulgare L.). G3 Genes Genomes Genet. 10, 1113–1124. https://doi.org/10.1534/g3.119.400968.

Biazzi, E., Nazzicari, N., Pecetti, L., Brummer, E.C., Palmonari, A., Tava, A., Annicchiarico, P., 2017. Genome-wide association mapping and genomic selection for alfalfa (Medicago sativa) forage quality traits. PLoS One 12, e0169234. https://doi.org/10.1371/journal.pone.0169234.

Blagus, R., Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106. https://doi.org/10.1186/1471-2105-14-106.

Blondel, M., Onogi, A., Iwata, H., Ueda, N., 2015. A ranking approach to genomic selection. PLoS One 10, e0128570. https://doi.org/10.1371/journal.pone.0128570.
Budhlakoti, M., Mishra, D.C.M., Rai, A., 2019. R Package MTGS: Genomic Selection Using Multiple Traits. https://CRAN.R-project.org/package=MTGS.
Burgueño, J., Crossa, J., Cotes, J.M., Vicente, F.S., Das, B., 2011. Prediction assessment of linear mixed models for multienvironment trials. Crop Sci. 51, 944–954. https://doi.org/10.2135/cropsci2010.07.0403.
Burgueño, J., de los Campos, G., Weigel, K., Crossa, J., 2012. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 52, 707–719. https://doi.org/10.2135/cropsci2011.06.0299.
Butler, D.G., Cullis, B.R., Gilmour, A.R., Gogel, B.J., Thompson, R., 2017. ASReml-R Reference Manual Version 4. http://www.vsni.co.uk/.
Callaway, E., 2020. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature 588, 203–204. https://doi.org/10.1038/d41586-020-03348-4.
Charmet, G., Tran, L.G., Auzanneau, J., Rincent, R., Bouchet, S., 2020. BWGS: a R package for genomic selection and its application to a wheat breeding programme. PLoS One 15, e0222733. https://doi.org/10.1371/journal.pone.0222733.
Clark, L.V., Dwiyanti, M.S., Anzoua, K.G., Brummer, J.E., Ghimire, B.K., Głowacka, K., Hall, M., Heo, K., Jin, X., Lipka, A.E., Peng, J., Yamada, T., Yoo, J.H., Yu, C.Y., Zhao, H., Long, S.P., Sacks, E.J., 2019. Genome-wide association and genomic prediction for biomass yield in a genetically diverse Miscanthus sinensis germplasm panel phenotyped at five locations in Asia and North America. GCB Bioenergy 11, 988–1007. https://doi.org/10.1111/gcbb.12620.
Crossa, J., de los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., Pérez-Rodríguez, P., 2016. Extending the marker × Environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Sci. 56, 2193–2209. https://doi.org/10.2135/cropsci2015.04.0260.
Cuevas, J., Crossa, José, Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., de los Campos, G., Montesinos-López, O.A., Burgueño, Juan, Roo, Q., Crossa, J., Montesinos-López, O., Burgueño, J., 2016. Genomic prediction of genotypé environment interaction kernel regression models. Plant Genome 9. https://doi.org/10.3835/plantgenome2016.03.0024.
Dan, Z., Chen, Y., Zhao, W., Wang, Q., Huang, W., 2020. Metabolome-based prediction of yield heterosis contributes to the breeding of elite rice. Life Sci. Alliance 3, 1–10. https://doi.org/10.26508/lsa.201900551.
de Abreu e. Lima, F., Willmitzer, L., Nikoloski, Z., 2018. Classification-driven framework to predict maize hybrid field performance from metabolic profiles of young parental roots. PLoS One 13, e0196038. https://doi.org/10.1371/journal.pone.0196038.
de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182, 375–385. https://doi.org/10.1534/genetics.109.101501.
de los Campos, G., Gianola, D., Rosa, G.J.M., Weigel, K.A., Crossa, J., 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. (Camb) 92, 295–308. https://doi.org/10.1017/S0016672310000285.
de Oliveira, E.J., de Resende, M.D.V., da Silva Santos, V., Ferreira, C.F., Oliveira, G.A.F., da Silva, M.S., de Oliveira, L.A., Aguilar-Vildoso, C.I., 2012. Genome-wide selection in cassava. Euphytica 187, 263–276. https://doi.org/10.1007/s10681-012-0722-0.
de Oliveira, A.A., Pastina, M.M., de Souza, V.F., da Costa Parrella, R.A., Noda, R.W., Simeone, M.L.F., Schaffert, R.E., de Magalhães, J.V., Damasceno, C.M.B., Margarido, G.R.A., 2018. Genomic prediction applied to high-biomass sorghum for bioenergy production. Mol. Breed. 38, 49. https://doi.org/10.1007/s11032-018-0802-5.
Deomano, E., Jackson, P., Wei, X., Aitken, K., Kota, R., Pérez-Rodríguez, P., 2020. Genomic prediction of sugar content and cane yield in sugar cane clones in different stages of selection in a breeding program, with and without pedigree information. Mol. Breed. 40, 38. https://doi.org/10.1007/s11032-020-01120-0.
Desta, Z.A., Ortiz, R., 2014. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 19, 592–601. https://doi.org/10.1016/j.tplants.2014.05.006.
Duangjit, J., Causse, M., Sauvage, C., 2016. Efficiency of genomic selection for tomato fruit quality. Mol. Breed. 36, 1–16. https://doi.org/10.1007/s11032-016-0453-3.
Endelman, J.B., 2011. Ridge regression and other kernels for genomic selection with r package rrBLUP. Plant Genome 4, 250–255. https://doi.org/10.3835/plantgenome2011.08.0024.
Fè, D., Ashraf, B.H., Pedersen, M.G., Janss, L., Byrne, S., Roulund, N., Lenk, I., Didion, T., Asp, T., Jensen, C.S., Jensen, J., 2016. Accuracy of genomic prediction in a commercial perennial ryegrass breeding program. Plant Genome 9. https://doi.org/10.3835/plantgenome2015.11.0110.
Fernandes, S.B., Dias, K.O., Ferreira, F.D., Brown, J.P., 2018. Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theor. Appl. Genet. 131, 747–755. https://doi.org/10.1007/s00122-017-3033-y.
Fernandez, O., Urrutia, M., Bernillon, S., Giauffret, C., Tardieu, F., Le Gouis, J., Langlade, N., Charcosset, A., Moing, A., Gibon, Y., 2016. Fortune telling: metabolic markers of plant performance. Metabolomics 12, 158. https://doi.org/10.1007/s11306-016-1099-1.
Fiedler, J.D., Lanzatella, C., Edmé, S.J., Palmer, N.A., Sarath, G., Mitchell, R., Tobias, C. M., 2018. Genomic prediction accuracy for switchgrass traits related to bioenergy within differentiated populations. BMC Plant Biol. 18, 142. https://doi.org/10.1186/s12870-018-1360-z.
Friedman, J., Hastie, T., Tibshirani, R., 2015. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 https://doi.org/10.1201/b18401.
Fu, J., Falke, K.C., Thiemann, A., Schrag, T.A., Melchinger, A.E., Scholten, S., Frisch, M., 2012. Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. Theor. Appl. Genet. 124, 825–833. https://doi.org/10.1007/s00122-011-1747-9.
Gaikpa, D.S., Koch, S., Fromme, F.J., Siekmann, D., Würschum, T., Miedaner, T., 2020. Genome-wide association mapping and genomic prediction of Fusarium head blight resistance, heading stage and plant height in winter rye (Secale cereale). Plant Breed. 139, 508–520. https://doi.org/10.1111/pbr.12810.
Gemmer, M.R., Richter, C., Jiang, Y., Schmutzer, T., Raorane, M.L., Junker, B., Pillen, K., Maurer, A., 2020. Can metabolic prediction be an alternative to genomic prediction in barley? PLoS One 15, e0234052. https://doi.org/10.1371/journal.pone.0234052.
Gezan, S.A., Osorio, L.F., Verma, S., Whitaker, V.M., 2017. An experimental validation of genomic selection in octoploid strawberry. Hortic. Res. 4, 16070. https://doi.org/10.1038/hortres.2016.70.
Gianola, D., van Kaam, J.B.C.H.M., 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178, 2289–2303. https://doi.org/10.1534/genetics.107.084285.
Gianola, D., Perez-Enciso, M., Toro, M.A., 2003. On marker-assisted prediction of genetic value: beyond the ridge. Genetics 163, 347–365.
Gillberg, J., Marttinen, P., Mamitsuka, H., Kaski, S., Stegle, O., 2019. Modelling G3E with historical weather information improves genomic prediction in new environments. Bioinformatics 35, 4045–4052. https://doi.org/10.1093/bioinformatics/btz197.
González-Camacho, J.M., de los Campos, G., Pérez, P., Gianola, D., Cairns, J.E., Mahuku, G., Babu, R., Crossa, J., 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet. 125, 759–771. https://doi.org/10.1007/s00122-012-1868-9.
González-Camacho, J.M., Crossa, J., Pérez-Rodríguez, P., Ornella, L., Gianola, D., 2016. Genome-enabled prediction using probabilistic neural network classifiers. BMC Genomics 17, 208. https://doi.org/10.1186/s12864-016-2553-1.
González-Camacho, J.M., Ornella, L., Pérez-Rodríguez, P., Gianola, D., Dreisigacker, S., Crossa, J., 2018. Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. Plant Genome 11. https://doi.org/10.3835/plantgenome2017.11.0104.
Granato, I., Cuevas, J., Luna-Vázquez, F., Crossa, J., Montesinos-López, O., Burgueño, J., Fritsche-Neto, R., 2018. BGGE: a new package for genomic-enabled prediction incorporating genotype × environment interaction models. G3 Genes Genomes Genet. 8, 3039–3047. https://doi.org/10.1534/g3.118.200435.
Gregory, T.R., 2009. Artificial selection and domestication: modern lessons from Darwin's enduring analogy. Evol. Educ. Outreach 2, 5–27. https://doi.org/10.1007/s12052-008-0114-z.
Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., Wang, D., 2016. Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theor. Appl. Genet. 129, 2413–2427. https://doi.org/10.1007/s00122-016-2780-5.
Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12, 186. https://doi.org/10.1186/1471-2105-12-186.
Habier, D., Fernando, R.L., Garrick, D.J., 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194, 597–607. https://doi.org/10.1534/genetics.113.152207.
Habyarimana, E., Lopez-Cruz, M., Baloch, F.S., 2020. Genomic selection for optimum index with dry biomass yield, dry mass fraction of fresh material, and plant height in biomass sorghum. Genes (Basel) 11, 61. https://doi.org/10.3390/genes11010061.
He, D., Kuhn, D., Parida, L., 2016. Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. Bioinformatics 32, i37–i43. https://doi.org/10.1093/bioinformatics/btw249.
Heffner, E.L., Jannink, J.L., Sorrells, M.E., 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4, 65–75. https://doi.org/10.3835/plantgenome2010.12.0029.
Heslot, N., Akdemir, D., Sorrells, M.E., Jannink, J.L., 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor. Appl. Genet. 127, 463–480. https://doi.org/10.1007/s00122-013-2231-5.
Holliday, J.A., Wang, T., Aitken, S., 2012. Predicting adaptive phenotypes from multilocus genotypes in sitka spruce (picea sitchensis) using random forest. G3 Genes Genomes Genet. 2, 1085–1093. https://doi.org/10.1534/g3.112.002733.
Hu, X., Xie, W., Wu, C., Xu, S., 2019. A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol. J. 17, 2011–2020. https://doi.org/10.1111/pbi.13117.
Jan, H.U., Abbadi, A., Lücke, S., Nichols, R.A., Snowdon, R.J., 2016. Genomic prediction of testcross performance in canola (Brassica napus). PLoS One 11, e0147769. https://doi.org/10.1371/journal.pone.0147769.
Jannink, J.L., Lorenz, A.J., Iwata, H., 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9, 166–177. https://doi.org/10.1093/bfgp/elq001.
Jia, Y., Jannink, J.L., 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192, 1513–1522. https://doi.org/10.1534/genetics.112.144246.
Jiang, Y., Reif, J.C., 2015. Modeling epistasis in genomic selection. Genetics 201, 759–768. https://doi.org/10.1534/genetics.115.177907.
Kang, M.S., 1997. Using genotype-by-Environment interaction for crop cultivar development. Adv. Agron. 62, 199–252. https://doi.org/10.1016/S0065-2113(08)60569-6.
Khaki, S., Wang, L., 2019. Crop yield prediction using deep neural networks. Front. Plant Sci. 10, 621. https://doi.org/10.3389/fpls.2019.00621.

Kristensen, P.S., Jahoor, A., Andersen, J.R., 2019. Multi-trait and trait-assisted genomic prediction of winter wheat quality traits using advanced lines from four breeding cycles. Crop Breed. Genet. Genom. 1, e1900010 https://doi.org/10.20900/cbgg20190010.

Küken, A., Nikoloski, Z., 2019. Computational approaches to design and test plant synthetic metabolic pathways. Plant Physiol. 179, 894–906. https://doi.org/10.1104/pp.18.01273.

Lado, B., Vázquez, D., Quincke, M., Silva, P., Aguilar, I., Gutiérrez, L., 2018. Resource allocation optimization with multi-trait genomic prediction for bread wheat (Triticum aestivum L.) baking quality. Theor. Appl. Genet. 131, 2719–2731. https://doi.org/10.1007/s00122-018-3186-3.

Legarra, A., Ricard, A., Filangi, O., 2016. GS3: Genomic Selection-Gibbs Sampling-Gauss Seidel. https://github.com/alegarra/gs3.

Liang, Z., Gupta, S.K., Yeh, C.T., Zhang, Y., Ngu, D.W., Kumar, R., Patil, H.T., Mungra, K. D., Yadav, D.V., Rathore, A., Srivastava, R.K., Gupta, R., Yang, J., Varshney, R.K., Schnable, P.S., Schnable, J.C., 2018. Phenotypic data from inbred parents can improve genomic prediction in pearl millet hybrids. G3 Genes Genomes Genet. 8, 2513–2522. https://doi.org/10.1534/g3.118.200242.

Liu, X., Wang, Hongwu, Wang, Hui, Guo, Z., Xu, X., Liu, J., Wang, S., Li, W.X., Zou, C., Prasanna, B.M., Olsen, M.S., Huang, C., Xu, Y., 2018. Factors affecting genomic selection revealed by empirical evidence in maize. Crop J. 6, 341–352. https://doi.org/10.1016/j.cj.2018.03.005.

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., Xu, D., 2019. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. Front. Genet. 10, 1091. https://doi.org/10.3389/fgene.2019.01091.

Long, N., Gianola, D., Rosa, G.J.M., Weigel, K.A., 2011. Application of support vector regression to genome-assisted prediction of quantitative traits. Theor. Appl. Genet. 123, 1065–1074. https://doi.org/10.1007/s00122-011-1648-y.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.L., Singh, R.P., Autrique, E., de los Campos, G., 2015. Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3 Genes Genomes Genet. 5, 569–582. https://doi.org/10.1534/g3.114.016097.

Lyra, D.H., de Freitas Mendonça, L., Galli, G., Alves, F.C., Granato, Í.S.C., Fritsche-Neto, R., 2017. Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. Mol. Breed. 37, 80. https://doi.org/10.1007/s11032-017-0681-1.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., Ma, C., 2018. A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta 248, 1307–1318. https://doi.org/10.1007/s00425-018-2976-9.

Majumdar, S.G., Rai, A., Mishra, D.C., 2019. R Package GSelection: Genomic Selection. https://cran.r-project.org/web/packages/GSelection/index.html.

Martini, J.W.R., Gao, N., Cardoso, D.F., Wimmer, V., Erbe, M., Cantet, R.J.C., Simianer, H., 2017. Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). BMC Bioinformatics 18, 3. https://doi.org/10.1186/s12859-016-1439-1.

Matei, G., Woyann, L.G., Milioli, A.S., de Bem Oliveira, I., Zdziarski, A.D., Zanella, R., Coelho, A.S.G., Finatto, T., Benin, G., 2018. Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. Mol. Breed. 38, 117. https://doi.org/10.1007/s11032-018-0872-4.

Matias, F.I., Alves, F.C., Meireles, K.G.X., Barrios, S.C.L., do Valle, C.B., Endelman, J.B., Fritsche-Neto, R., 2019. On the accuracy of genomic prediction models considering multi-trait and allele dosage in Urochloa spp. interspecific tetraploid hybrids. Mol. Breed. 39, 100. https://doi.org/10.1007/s11032-019-1002-7.

Mellers, G., Mackay, I., Cowan, S., Griffiths, I., Martinez-Martin, P., Poland, J.A., Bekele, W., Tinker, N.A., Bentley, A.R., Howarth, C.J., 2020. Implementing within-cross genomic prediction to reduce oat breeding costs. Plant Genome 13, e20004. https://doi.org/10.1002/tpg2.20004.

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

Millet, E.J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., Charcosset, A., Welcker, C., Van Eeuwijk, F., Tardieu, F., 2019. Genomic prediction of maize yield across European environmental conditions. Nat. Genet. 51, 952–956. https://doi.org/10.1038/s41588-019-0414-y.

Moeinizade, S., Kusmec, A., Hu, G., Wang, L., Schnable, P.S., 2020. Multi-trait genomic selection methods for crop improvement. Genetics 215, 931–945. https://doi.org/10.1534/genetics.120.303305.

Montesinos-López, O.A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., Juliana, P., Singh, R., 2019a. New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. G3 Genes Genomes Genet. 9, 1545–1556. https://doi.org/10.1534/g3.119.300585.

Montesinos-López, O.A., Montesinos-López, A., Luna-Vázquez, F.J., Toledo, F.H., Pérez-Rodríguez, P., Lillemo, M., Crossa, J., 2019b. An R package for Bayesian analysis of multi-environment and multi-trait multi-environment data for genome-based prediction. G3 Genes Genomes Genet. 9, 1355–1369. https://doi.org/10.1534/g3.119.400126.

Montesinos-López, O.A., Montesinos-López, A., Tuberosa, R., Maccaferri, M., Sciara, G., Ammar, K., Crossa, J., 2019c. Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. Front. Plant Sci. 10, 1311. https://doi.org/10.3389/fpls.2019.01311.

Morota, G., 2017. ShinyGPAS: interactive genomic prediction accuracy simulator based on deterministic formulas. Genet. Sel. Evol. 49, 91. https://doi.org/10.1186/s12711-017-0368-4.

Nakaya, A., Isobe, S.N., 2012. Will genomic selection be a practical method for plant breeding? Ann. Bot. 110, 1303–1316. https://doi.org/10.1093/aob/mcs109.

Nazarian, A., Gezan, S.A., 2016. GenoMatrix: a software package for pedigree-based and genomic prediction analyses on complex traits. J. Hered. 107, 372–379. https://doi.org/10.1093/jhered/esw020.

Nyine, M., Uwimana, B., Blavet, N., Hřibová, E., Vanrespaille, H., Batte, M., Akech, V., Brown, A., Lorenzen, J., Swennen, R., Doležel, J., 2018. Genomic prediction in a multiploid crop: genotype by environment interaction and allele dosage effects on predictive ability in banana. Plant Genome 11. https://doi.org/10.3835/plantgenome2017.10.0090.

Ogutu, J.O., Schulz-Streeck, T., Piepho, H.P., 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc. 6, S10. https://doi.org/10.1186/1753-6561-6-S2-S10.

Ornella, L., Pérez, P., Tapia, E., González-Camacho, J., Burgueño, J., Zhang, X., Singh, S., Vicente, F., Bonnett, D., Dreisigacker, S., Singh, R., Long, N., Crossa, J., 2014. Genomic-enabled prediction with classification algorithms. Heredity (Edinb) 112, 616–626. https://doi.org/10.1038/hdy.2013.144.

Ortiz, R., Gowda, M., Zhao, Y., Velazco, J.G., Van Eeuwijk, F.A., Jordan, D.R., Mace, E.S., Hunt, C.H., Malosetti, M., 2019. Genomic prediction of grain yield and drought-adaptation capacity in Sorghum Is enhanced by multi-trait analysis. Front. Plant Sci. 10, 997. https://doi.org/10.3389/fpls.2019.00997.

Pérez, P., de los Campos, G., 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198, 483–495. https://doi.org/10.1534/genetics.114.164442.

Pérez, P., de los Campos, G., Crossa, J., Gianola, D., 2010. Genomic-enabled prediction based on molecular markers and pedigree using the bayesian linear regression package in r. Plant Genome 3, 106–116. https://doi.org/10.3835/plantgenome2010.04.0005.

Pérez-Enciso, M., Ramírez-Ayala, L.C., Zingaretti, L.M., 2020. SeqBreed: a python tool to evaluate genomic prediction in complex scenarios. Genet. Sel. Evol. 52, 7. https://doi.org/10.1186/s12711-020-0530-2.

Pérez-Enciso, M., Zingaretti, L.M., 2019. A guide for using deep learning for complex trait genomic prediction. Genes (Basel) 10, 553. https://doi.org/10.3390/genes10070553.

Prakapenka, D., Wang, C., Liang, Z., Bian, C., Tan, C., Da, Y., 2020. GVCHAP: a computing pipeline for genomic prediction and variance component estimation using haplotypes and SNP markers. Front. Genet. 11, 282. https://doi.org/10.3389/fgene.2020.00282.

Qiu, Z., Cheng, Q., Song, J., Tang, Y., Ma, C., 2016. Application of machine learning-based classification to genomic selection and performance improvement. ICIC 9771, 412–421. https://doi.org/10.1007/978-3-319-42291-6_41.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., Melchinger, A.E., 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. 44, 217–220. https://doi.org/10.1038/ng.1033.

Riekeberg, E., Powers, R., 2017. New frontiers in metabolomics: from measurement to insight. F1000Research 6, 1148. https://doi.org/10.12688/f1000research.11495.1.

Roth, M., Muranty, H., Di Guardo, M., Guerra, W., Patocchi, A., Costa, F., 2020. Genomic prediction of fruit texture and training population optimization towards the application of genomic selection in apple. Hortic. Res. 7, 148. https://doi.org/10.1038/s41438-020-00370-5.

Runcie, D., Cheng, H., 2019. Pitfalls and remedies for cross validation with multi-trait genomic prediction methods. G3 Genes Genomes Genet. 9, 3727–3741. https://doi.org/10.1534/g3.119.400598.

Schnable, P.S., Springer, N.M., 2013. Progress toward understanding heterosis in crop plants. Annu. Rev. Plant Biol. 64, 71–88. https://doi.org/10.1146/annurev-arplant-042110-103827.

Schrag, T.A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., Melchinger, A.E., 2018. Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. Genetics 208, 1373–1385. https://doi.org/10.1534/genetics.117.300374.

Schulthess, A.W., Wang, Y., Miedaner, T., Wilde, P., Reif, J.C., Zhao, Y., 2016. Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. Theor. Appl. Genet. 129, 273–287. https://doi.org/10.1007/s00122-015-2626-6.

Schulz-Streeck, T., Ogutu, J.O., Gordillo, A., Karaman, Z., Knaak, C., Piepho, H.P., 2013. Genomic selection allowing for marker-by-environment interaction. Plant Breed. 132, 532–538. https://doi.org/10.1111/pbr.12105.

Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Meuwissen, T.H.E., 2008. Genomic selection using different marker types and densities. J. Anim. Sci. 86, 2447–2454. https://doi.org/10.2527/jas.2007-0010.

Sousa, T.V., Caixeta, E.T., Alkimim, E.R., Oliveira, A.C.B., Pereira, A.A., Sakiyama, N.S., Zambolim, L., Resende, M.D.V., 2019. Early selection enabled by the implementation of genomic selection in coffea arabica breeding. Front. Plant Sci. 9, 1934. https://doi.org/10.3389/fpls.2018.01934.

Su, G., Christensen, O.F., Ostersen, T., Henryon, M., Lund, M.S., 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One 7, e45293. https://doi.org/10.1371/journal.pone.0045293.

Technow, F., Bürger, A., Melchinger, A.E., 2013. Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3 Genes Genomes Genet. 3, 197–203. https://doi.org/10.1534/g3.112.004630.

Tecle, I.Y., Edwards, J.D., Menda, N., Egesi, C., Rabbi, I.Y., Kulakow, P., Kawuki, R., Jannink, J.L., Mueller, L.A., 2014. solGS: a web-based tool for genomic selection. BMC Bioinformatics 15, 398. https://doi.org/10.1186/s12859-014-0398-7.

Thiele, I., Palsson, B.Ø., 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat. Protoc. 5, 93–121. https://doi.org/10.1038/nprot.2009.203.

Tong, H., Küken, A., Nikoloski, Z., 2020. Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth. Nat. Commun. 11, 2410. https://doi.org/10.1038/s41467-020-16279-5.

Tsai, H.Y., Janss, L.L., Andersen, J.R., Orabi, J., Jensen, J.D., Jahoor, A., Jensen, J., 2020. Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. Sci. Rep. 10, 1–15. https://doi.org/10.1038/s41598-020-60203-2.

Turner-Hissong, S.D., Bird, K.A., Lipka, A.E., King, E.G., Beissinger, T.M., Angelovici, R., 2020. Genomic prediction informed by biological processes expands our understanding of the genetic architecture underlying free amino acid traits in dry Arabidopsis seeds. G3 Genes Genomes Genet. 10, 4227–4239. https://doi.org/10.1534/g3.120.401240.

Ukrainetz, N.K., Mansfield, S.D., 2020. Assessing the sensitivities of genomic selection for growth and wood quality traits in lodgepole pine using Bayesian models. Tree Genet. Genomes 16, 14. https://doi.org/10.1007/s11295-019-1404-z.

Usai, M.G., Goddard, M.E., Hayes, B.J., 2009. LASSO with cross-validation for genomic selection. Genet. Res. (Camb) 91, 427–436. https://doi.org/10.1017/S0016672309990334.

Viana, A.P., de Resende, M.D.V., Riaz, S., Walker, M.A., 2016. Genome selection in fruit breeding: application to table grapes. Sci. Agric. 73, 142–149. https://doi.org/10.1590/0103-9016-2014-0323.

Vignal, A., Milan, D., SanCristobal, M., Eggen, A., 2002. A review on SNPs and other types of molecular markers. Genet. Sel. Evol. 34, 275–305. https://doi.org/10.1051/gse.

Voichek, Y., Weigel, D., 2020. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nat. Genet. 52, 534–540. https://doi.org/10.1038/s41588-020-0612-7.

Wang, C., Prakapenka, D., Wang, S., Pulugurta, S., Runesha, H.B., Da, Y., 2014. GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. BMC Bioinformatics 15, 270. https://doi.org/10.1186/1471-2105-15-270.

Wang, X., Li, L., Yang, Z., Zheng, X., Yu, S., Xu, C., Hu, Z., 2017. Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. Heredity (Edinb) 118, 302–310. https://doi.org/10.1038/hdy.2016.87.

Westhues, M., Schrag, T.A., Heuer, C., Thaller, G., Utz, H.F., Schipprack, W., Thiemann, A., Seifert, F., Ehret, A., Schlereth, A., Stitt, M., Nikoloski, Z., Willmitzer, L., Schön, C.C., Scholten, S., Melchinger, A.E., 2017. Omics-based hybrid prediction in maize. Theor. Appl. Genet. 130, 1927–1939. https://doi.org/10.1007/s00122-017-2934-0.

Westhues, M., Heuer, C., Thaller, G., Fernando, R., Melchinger, A.E., 2019. Efficient genetic value prediction using incomplete omics data. Theor. Appl. Genet. 132, 1211–1222. https://doi.org/10.1007/s00122-018-03273-1.

Wolfe, M.D., Carpio, D.P.D., Alabi, O., Ezenwaka, L.C., Ikeogu, U.N., Kayondo, I.S., Lozano, R., Okeke, U.G., Ozimati, A.A., Williams, E., Egesi, C., Kawuki, R.S., Kulakow, P., Rabbi, I.Y., Jannink, J.L., 2017. Prospects for genomic selection in cassava breeding. Plant Genome 10. https://doi.org/10.3835/plantgenome2017.03.0015.

Xu, S., Zhu, D., Zhang, Q., 2014. Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc. Natl. Acad. Sci. U. S. A. 111, 12456–12461. https://doi.org/10.1073/pnas.1413750111.

Xu, S., Xu, Y., Gong, L., Zhang, Q., 2016. Metabolomic prediction of yield in hybrid rice. Plant J. 88, 219–227. https://doi.org/10.1111/tpj.13242.

Zhang, X., Sallam, A., Gao, L., Kantarski, T., Poland, J., DeHaan, L.R., Wyse, D.L., Anderson, J.A., 2016. Establishment and optimization of genomic selection to accelerate the domestication and improvement of intermediate wheatgrass. Plant Genome 9. https://doi.org/10.3835/plantgenome2015.07.0059.

Glossary

Breeding value: Additive effect for a trait of interest that can be stably passed from parents to offspring.

Estimated breeding values: Breeding value estimated by a model relating the phenotype over a population of individuals with their pedigree information.

Genomic estimated breeding values: Breeding value estimated by using genome-wide markers based on application on various machine learning approaches.

Training population: Population of genotypes for which both genotypic and phenotypic data are available.

Breeding (testing) population: Population of genotypes for which only genotypic data are available.

Regression: Machine learning approach that aims to estimate the relationship between a response and predictors, provided data over different samples. Some approaches include: ridge regression and LASSO.

Classification: Machine learning approach that aims to estimate a model that maps inputs to responses, provided training data on input-response pairs. Some approaches include: support vector machine (SVM) and random forest (RF).

Regularization: Statistical technique used to prevent overfitting.

Fixed vs. random effect: Model parameters that are constant vs. random variables over the set of observations.

Predictability: For quantitative traits, measured by the Pearson correlation coefficient between predicted and observed phenotypes in cross-validation. For qualitative traits, measured by different coefficients (e.g. Cohen's kappa, AUC).

Deep learning: Machine learning approach that relies on artificial neural networks to learn a non-linear relationship between inputs and response.

Singe-trait model: Model that includes a single response to be predicted.

Multi-trait model: Model which includes multiple responses that are simultaneously predicted.