CellPress
Partner Journal

Molecular Plant
Review article

Genomic selection in plant breeding: Key factors shaping two decades of progress

Admas Alemu[1,*], Johanna Åstrand[1,2], Osval A. Montesinos-López[3], Julio Isidro y Sánchez[4], Javier Fernández-Gónzalez[4], Wuletaw Tadesse[5], Ramesh R. Vetukuri[1], Anders S. Carlsson[1], Alf Ceplitis[2], José Crossa[6], Rodomiro Ortiz[1,*] and Aakash Chawade[1]

[1]Department of Plant Breeding, Swedish University of Agricultural Sciences, Alnarp, Sweden

[2]Lantmännen Lantbruk, Svalöv, Sweden

[3]Facultad de Telemática, University de Colima, Colima, Colima 28040, Mexico

[4]Centro de Biotecnología y Genómica de Plantas (CBGP, UPM-INIA), Universidad Politécnica de Madrid (UPM) – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Campus de Montegancedo-UPM, 28223 Madrid, Spain

[5]International Center for Agricultural Research in the Dry Areas (ICARDA), Rabat, Morocco

[6]International Maize and Wheat Improvement Center (CIMMYT), Km 45, Carretera México-Veracruz, Texcoco, México 52640, Mexico

***Correspondence: Admas Alemu (admas.alemu.abebe@slu.se), Rodomiro Ortiz (rodomiro.ortiz@slu.se)**

https://doi.org/10.1016/j.molp.2024.03.007

ABSTRACT

Genomic selection, the application of genomic prediction (GP) models to select candidate individuals, has significantly advanced in the past two decades, effectively accelerating genetic gains in plant breeding. This article provides a holistic overview of key factors that have influenced GP in plant breeding during this period. We delved into the pivotal roles of training population size and genetic diversity, and their relationship with the breeding population, in determining GP accuracy. Special emphasis was placed on optimizing training population size. We explored its benefits and the associated diminishing returns beyond an optimum size. This was done while considering the balance between resource allocation and maximizing prediction accuracy through current optimization algorithms. The density and distribution of single-nucleotide polymorphisms, level of linkage disequilibrium, genetic complexity, trait heritability, statistical machine-learning methods, and non-additive effects are the other vital factors. Using wheat, maize, and potato as examples, we summarize the effect of these factors on the accuracy of GP for various traits. The search for high accuracy in GP—theoretically reaching one when using the Pearson's correlation as a metric—is an active research area as yet far from optimal for various traits. We hypothesize that with ultra-high sizes of genotypic and phenotypic datasets, effective training population optimization methods and support from other omics approaches (transcriptomics, metabolomics and proteomics) coupled with deep-learning algorithms could overcome the boundaries of current limitations to achieve the highest possible prediction accuracy, making genomic selection an effective tool in plant breeding.

Key words: genomic selection, genetic gain, genomic prediction optimization, deep learning, training population optimization

Alemu A., Åstrand J., Montesinos-López O.A., Isidro y Sánchez J., Fernández-Gónzalez J., Tadesse W., Vetukuri R.R., Carlsson A.S., Ceplitis A., Crossa J., Ortiz R., and Chawade A. (2024). Genomic selection in plant breeding: Key factors shaping two decades of progress. Mol. Plant. **17**, 552–578.

INTRODUCTION

Global population growth is likely to continue at a similar or faster pace in the coming decades. Demand for food is expected to increase by the same amount to feed the population while crop productivity has been curtailed by various biotic and abiotic stresses exacerbated by anthropogenic climate change. Plant breeding is fundamental to developing new cultivars with higher yield, improved quality, and tolerance or resistance to several abiotic and biotic stresses. For example, wheat production at the global level has increased from 200 million tons in 1961 to 775 million tons in 2023 (FAO, 2023) with no significant change in total area

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

of wheat production (220 million hectares). This is principally due to the development and deployment of semi-dwarf high-yielding and input-responsive new wheat cultivars (Borlaug, 2002) with resistance and tolerance to major biotic and abiotic stresses, respectively, along with improved agronomic management, mechanization, favorable policies, and infrastructures across the entire wheat value chain (Tadesse et al., 2019).

Genetic enhancement of crops has long relied on conventional cross-breeding methods whereby breeding and selection of genotypes are solely based on pedigree and phenotypic performance. Rigorous evaluation of parents for different traits, targeted crossing, generation advancement using the summer and winter shuttle breeding schemes to shorten the breeding cycle, key location evaluation of elite germplasms, and effective database management have played significant roles in developing improved crop cultivars. However, the expeditious emergence of DNA-sequencing technology has allowed breeders to gain comprehensive genomic information on crops, which is very valuable for selection. The development of several DNA-marker-based genotyping systems significantly increased the number of DNA markers available to plant breeders (Crossa et al., 2017). This breakthrough allowed plant breeders to select plant performance based on their genetic marker composition rather than solely on their phenotypic performance, which is prone to several limitations in selection efficiency.

The application of genomic tools in the breeding practice of plants, generally termed genomic-assisted breeding, has progressed through various stages in the last four decades (Varshney et al., 2021). It started with linkage-based mapping of quantitative trait loci (QTLs) (Soller and Plotkin-Hazan, 1977) where, with a limited number of DNA markers, those segregating with a particular trait were identified as linked to a QTL and used for marker-assisted selection (MAS). The method required a set of segregating individuals developed from biparental crosses, a time-consuming procedure, with a narrow allelic variation and poor resolution that leads to low impact in practical plant-breeding programs (Bernardo, 2008). The genome-wide association study (GWAS) approach became a popular and powerful method for identifying markers closely linked to QTLs of target traits (Zhu et al., 2008; Tibbs Cortes et al., 2021). However, the practical implementation of the method via MAS has been constrained to limited numbers of major QTLs while numerous small-effect QTLs in complex traits have remained unknown and unutilized (Jannink et al., 2010).

Genomic selection (GS), when developed GP models are applied in practical selection, has emerged as a powerful tool in plant breeding, particularly after the advancement of readily available genome-wide single-nucleotide polymorphisms (SNPs). Besides early contributors (Lande and Thompson, 1990; Bernardo, 1994; Nejati-Javaremi et al., 1997; Haley and Visscher, 1998; Whittaker et al., 2000), GS was first elaborated two decades ago by Meuwissen et al. (2001). In this groundbreaking study, the authors paved the way to a new avenue in plant breeding, suggesting that prediction of genetic values from marker profiles could extensively increase genetic gain in plant and animal breeding, particularly if combined with reproductive techniques to shorten the generation interval. The conventional MAS approaches tend to focus solely on a limited set of markers linked with well-investigated major QTLs excluding the vast majority of minor-effect QTLs. In contrast to these methods, GP employs large number of genome-wide SNPs to quantify the comprehensive genetic merit of individual plants encompassing most contributing QTLs of a target trait (Bernardo and Yu, 2007; Heffner et al., 2009). The continued rapid advancement of next-generation sequencing technology to produce dense genome-wide SNP markers, coupled with its substantial cost reduction for genotyping in several crops, makes GS a must-implement method in most breeding programs. Empirical research has shown the advantage of GS for accelerating the genetic gains per unit of time over pedigree-based selection. GS has emerged with huge potential to reduce the cost per breeding cycle, increase selection intensity and accuracy, and significantly reduce the time required to develop a cultivar compared to phenotypic-based selection (Crossa et al., 2010, 2017; Edwards et al., 2019).

Developing statistical machine-learning models and training population optimization are the two main thematic areas actively explored in plant GP research. This is because of their potential to improve the prediction accuracy while the current achievement is far from optimal. This review begins with a simplified explanation of GP followed by an exploration of the up-to-date widely applied cross-validation (CV) methods in plant breeding. After a comprehensive overview, details of the key factors affecting GP accuracy identified over the last two decades are elaborated. Moreover, empirical research results are analyzed using wheat, maize, and potato as examples of self-pollinating, cross-pollinating, and clonally propagated crops, respectively, to illustrate the impact of the identified factors on the accuracy of GP in various traits. Finally yet importantly, the implementation of GS is highlighted in a showcase example from ongoing empirical studies from public and private breeding programs. In summary, valuable suggestions are forwarded to support the successful implementation of GS in plant-breeding programs.

GENOMIC PREDICTION

GP is the most recent data-driven method that has been widely accepted and used as a valuable tool to accelerate genetic gain in plant-breeding programs (Desta and Ortiz, 2014; Bassi et al., 2016; Xu et al., 2020). GP employs advanced statistical machine-learning models to select individuals within a breeding population based on breeding values estimated from genome-wide markers. This selection process relies on data from a training population, encompassing both phenotypic and genotypic information (Figure 1A). After a rigorous training procedure, these models generate predictions of breeding or phenotypic values for traits of a target population consisting only of genotypic data. However, the performance of prediction models should be first evaluated through CV before applying selection (see the next section for details of CV methods). This step in GP is critical in order to evaluate the performance of prediction models and compare different sets of statistical machine-learning models with various scenarios, such as incorporating multiple traits, known major genes and marker-trait associations (QTLs), genotype × environment (G×E) interaction, and other omics data such as transcriptomics, metabolomics, and proteomics (Figure 1A).

Comparisons among GP methods are evaluated through their prediction accuracy, which is directly linked to the breeder's

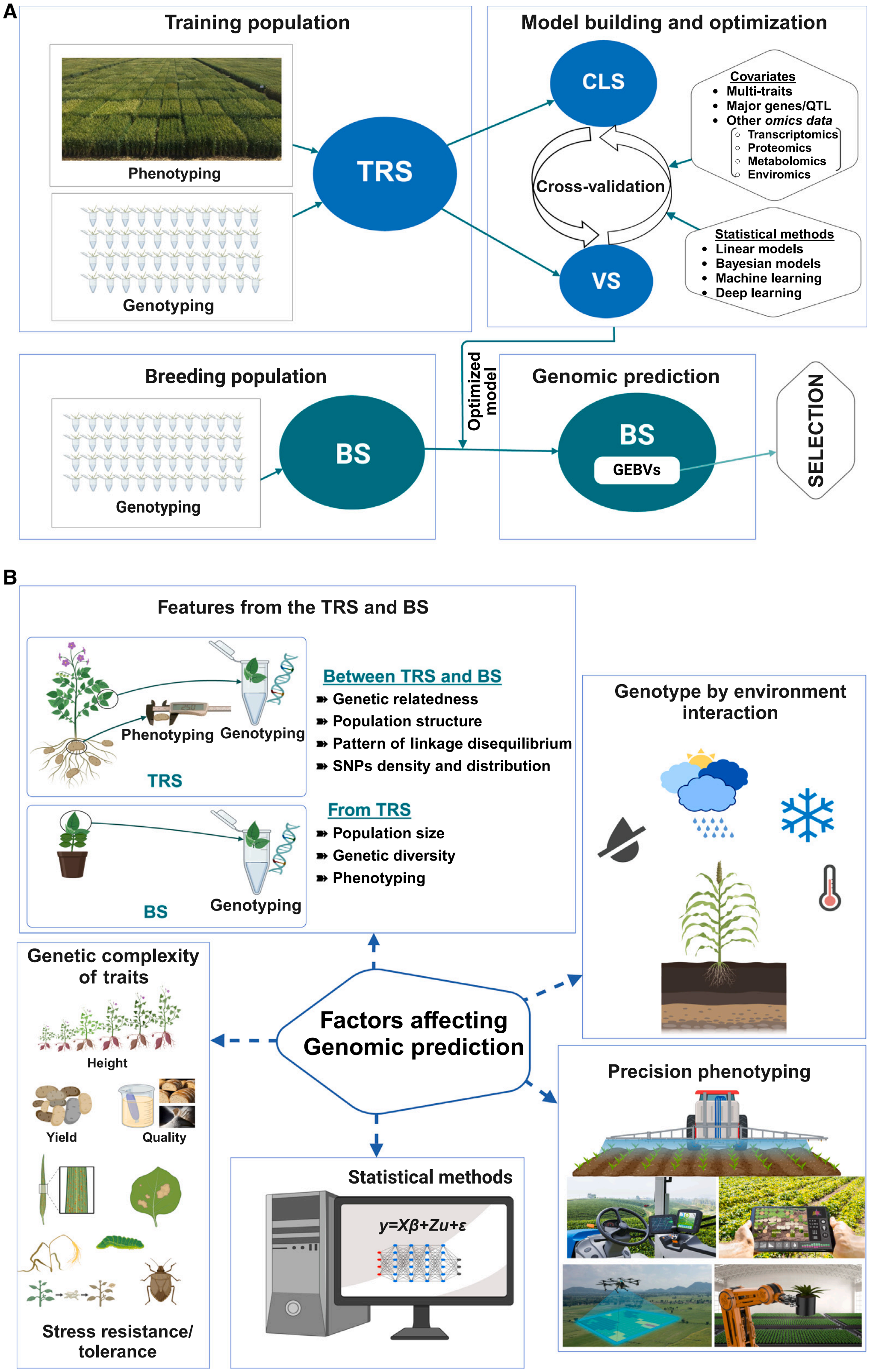

Figure 1. Schematic overview of GP model building and optimization, and major factors affecting genomic prediction.
(A) In genomic prediction, phenotypic and genotypic data as well as other covariates can be applied to develop and optimize various machine-learning methods splitting the optimized training population into calibration and validation sets and estimating the prediction accuracy through cross-validation.

(legend continued on next page)

equation (Akdemir and Isidro-Sánchez, 2019). Various factors can affect GP, and the accuracy score varies significantly across experiments for a single trait. For instance, the prediction accuracy of a single trait in wheat, maize, and potato hugely varied across different experimental research due to the different setups in training population composition, applied statistical machine-learning models, and other factors (Supplemental Tables 1–3). The GP accuracy (r_{MG} [correlation between marker predicted value with true predicted genetic value]) is measured as the Pearson's correlation between genomic estimated breeding value (GEBV) and true breeding value (Combs and Bernardo, 2013; Isidro et al., 2015), which gives an estimate of selection accuracy (Merrick et al., 2022). Selection accuracy is directly related to selection response (R), also known as genetic gain, and in the breeders' equation is calculated as $R = \ ir\sigma_A/t$, where i and r are the selection intensity and accuracy, respectively; while σ_A is the square root of the additive genetic variance and t is the cycle time (Falconer and Mackay, 1996).

GP considers the breeding values of parental average and deviation of Mendelian sampling to define GEBVs of an offspring, which allows the method to be used for: (1) rapid selection cycle with short breeding interval at early generations via prediction of the additive effects (i.e., GS at the F_2 level of a biparental cross); and (2) selection of lines at late stages of selection by predicting the genotypic values of individuals, with both additive and non-additive effects determining the final commercial value of the lines (Crossa et al., 2014; Dreisigacker et al., 2023).

Numerous factors affect GP and can significantly reduce its accuracy (Figure 1B). Consequently, unless adequately addressed, they can hinder the effective utilization of GP in plant-breeding programs. The population size, genetic diversity, and genetic relatedness with the breeding population are key features to target during training population optimization. Factors such as the level of linkage disequilibrium between QTLs and markers (in both the training and breeding [testing] population), genetic complexity and heritability of target traits, quality/precision phenotyping, statistical machine-learning models, G×E interaction, and other non-additive factors are the other major features that further complicate GP in plant breeding.

CROSS-VALIDATION METHODS

CV is a fundamental technique in statistical machine-learning methods that aids model evaluation, hyperparameter tuning, and ensuring robust model performance. It plays a crucial role in building models that can make accurate predictions on new, unseen data while avoiding overfitting and data-specific biases. GP models should initially be evaluated using CV methods before applying for the selection of candidate individuals in the breeding population. CV simulates the model's prediction performance by dividing the training population (training set; TRS) into calibration and validation sets.

Different GP CV methods are utilized depending on various determining scenarios (Figure 2). The K-fold CV is one of the most widely applied methods, where the entire dataset is divided into an equal number of folds. In the 5-fold CV method, for example, the TRS dataset is randomly grouped into 5-folds and prediction models are trained using the 4-folds as a calibration set while the remaining fold is used as a validation set. The accuracy could be measured after either averaging multiple runs from each fold or averaging runs comprising all folds. Leave-one-out CV (LOOCV) is the other method in which a single genotype is excluded from the calibration set and used as the validation set in each single iteration. An equal number of CV iterations are required with the number of samples or genotypes in this method. Hence, LOOCV is computationally intensive and only suitable for few genotypes (samples), while the 5-fold CV method is ideal for large datasets (Cheng et al., 2017). The other CV scenario has arisen in the case of multi-environment GP analysis (Crossa et al., 2017). Cross-validation 1 (CV1) is a scenario in which the GEBVs of newly developed lines or varieties are predicted in tested environments, thus being CV1 appropriate for predicting untested lines in tested environments. CV2, also known as sparse testing, is a method for genotypes tested in some environments and predicted in other tested environments. For this reason, CV2 is a reasonable option for predicting tested lines in tested environments. The other scenarios are CV0, which arises from the prediction of tested genotypes in an untested (unobserved) environment, while CV00 is used for predicting GEBVs of untested genotypes in unobserved environment (Figure 2).

TRAINING POPULATION

A TRS is used to establish the statistical relationship between genetic markers and phenotypic data for target traits to predict the phenotypic performance of individuals from their genotypic profile. In GP, the TRS should first be optimized to enhance the prediction accuracy and efficiency in breeding programs (see "training population optimization"). The optimized TRS can be of two types during the GP model optimization and application in the practical selection scenario. The first type is the parcel of the optimized TRS (calibration set) used to train the prediction models and estimate the GEBVs of the remaining individuals within the TRS (validation set) via CVs (Figure 1A). The second type is the overall optimized TRS applied to train the optimized GP models in a

The optimized model with the highest possible prediction accuracy is identified and applied to predict GEBV of the breeding population followed by selection of individuals based on their genetic merit for target traits.

(B) Various factors affect genomic prediction accuracy in plant-breeding programs. These factors arise from diverse sources at different stages during the analysis. Population size and genetic diversity of the training population, genetic relationship (kinship) and population structure of the training population with the breeding population, and quality of the phenotypic data applied in the statistical machine-learning models are features connected with training population and should be optimized during TRS development. Other factors including density and distribution of genetic markers across chromosomes, level of linkage disequilibrium between QTL alleles and marker alleles, genetic complexity and heritability of target traits, applied statistical methods, and non-additive genetic factors such as genotype-by-environment (G×E) interactions hugely affect the final output of the GP accuracy. TRS, training population; BS, breeding population/set; CLS, calibration set; VS, validation set; GEBV, genomic estimated breeding value. All figures are created with BioRender (https://biorender.com/).

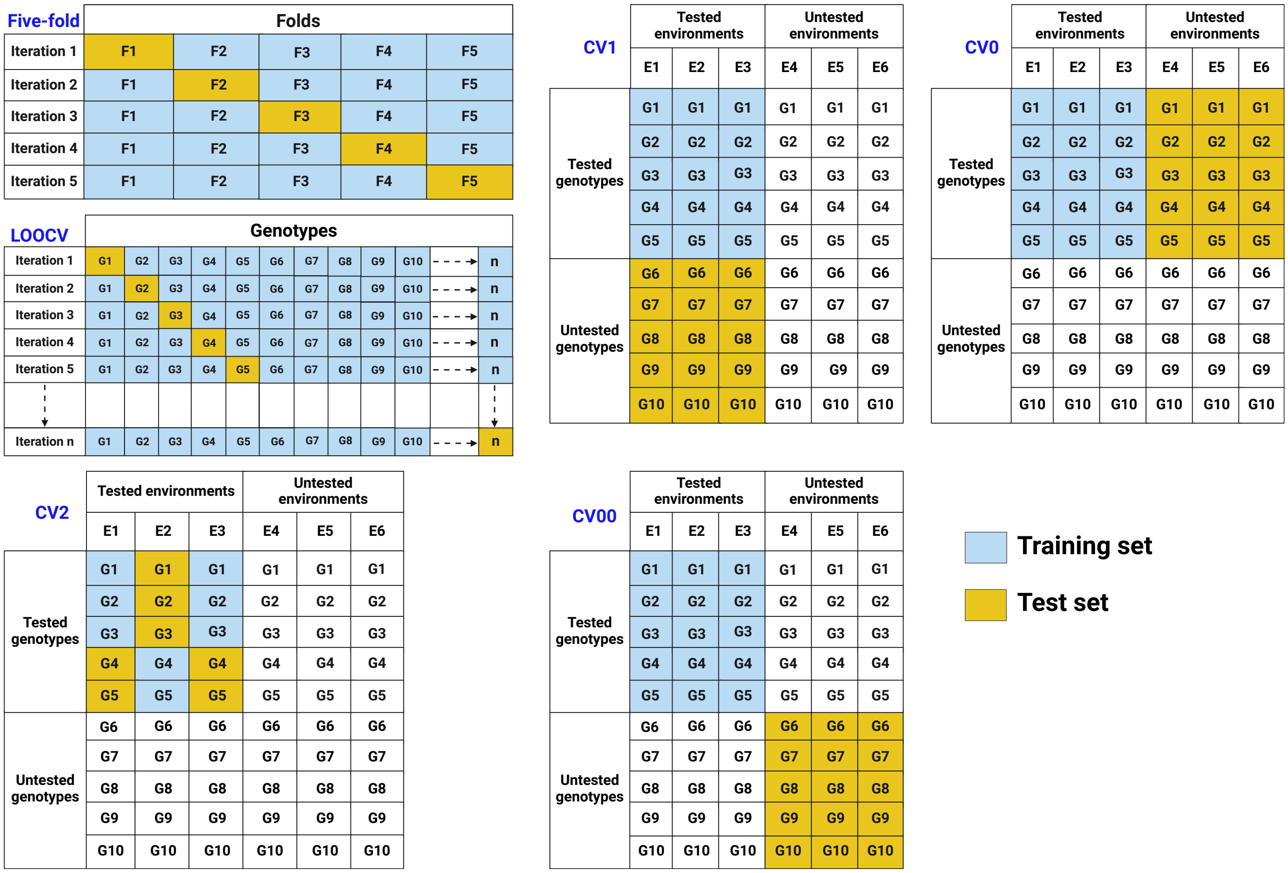

Figure 2. Genomic prediction cross-validation methods in plant breeding.

With the 5-fold cross-validation method, the complete population is initially allocated at random to 5-folds (F5). The 4-folds are then used as a calibration set in order to develop the GP model while retaining the remaining one as a validation set. A single genotype is excluded from the calibration set in the LOOCV, and its GEBV is predicted in every iteration. In multi-environment GP, a newly developed untested genotype can be predicted in tested environments (CV1), a genotype tested in some environments but untested in others (also known as sparse testing [CV2]), tested genotype predicted in an untested environment (CV0), and an untested genotype in an untested environment (CV00).

practical breeding scenario to estimate the GEBVs of individuals in the breeding population/set (BS), which are ready for selection. Features of the TRS including the population size, genetic diversity and genetic relatedness with the BS, population structure, level of linkage disequilibrium (LD) related to the BS, and the quality of phenotypic and genotypic data significantly affect the GP accuracy (Pszczola et al., 2012; Crossa et al., 2014; Hickey et al., 2014; Zhang et al., 2017a; Edwards et al., 2019).

Sample size of the training population

The ultimate goal of plant breeders is to achieve highly accurate but inexpensive estimates of genetic value (Lorenz and Nice, 2017). In GP, increasing the TRS size could inflict both positive and negative consequences for successful implementation in plant breeding (Merrick et al., 2022). The size of the TRS affects the accuracy of GP models (Goddard, 2009; Daetwyler et al., 2010; Combs and Bernardo, 2013; Bassi et al., 2016) and often correlates positively with the increase in size (Lorenzana and Bernardo, 2009; Zhong et al., 2009; Albrecht et al., 2011; Bentley et al., 2014; Isidro et al., 2015). However, research has shown a plateau in prediction-accuracy increment after reaching an optimum TRS size (Arruda et al., 2015; Sverrisdóttir et al., 2018; Fernández-González et al., 2023). Increasing the size of TRS demands greater effort and higher costs required for phenotyping as the genotyping cost has been significantly reduced. In addition, increasing the TRS could adversely affect the quality of collected phenotypic data, leading to reduced prediction accuracy. The TRS optimization encircles balancing to achieve the highest possible r_{MG} with minimum resource allocation through selective phenotyping (Figure 3A) (Lorenz and Nice, 2017; Akdemir and Isidro-Sánchez, 2019). Research has been conducted to identify an optimized TRS size and demonstrate the effects of numerous determining factors, such as the genetic kinship and population structure with the BS, LD extent, heritability, and genetic architecture of target traits (Isidro et al., 2015; Akdemir and Isidro-Sánchez, 2019; Sarinelli et al., 2019). Broadly, to achieve a higher r_{MG}, the size of TRS should increase when the genetic kinship with the BS decreases. Likewise, accuracy is often low for less-heritable traits, which is directly related to the complexity of the genetic architecture with several contributing small-effect QTLs and when LD between markers and QTLs is low (Habier et al., 2007; Daetwyler et al., 2010; Clark et al., 2012; Combs and Bernardo, 2013; Wientjes et al., 2013; Isidro et al., 2015). New optimization methods with the capability to automatically

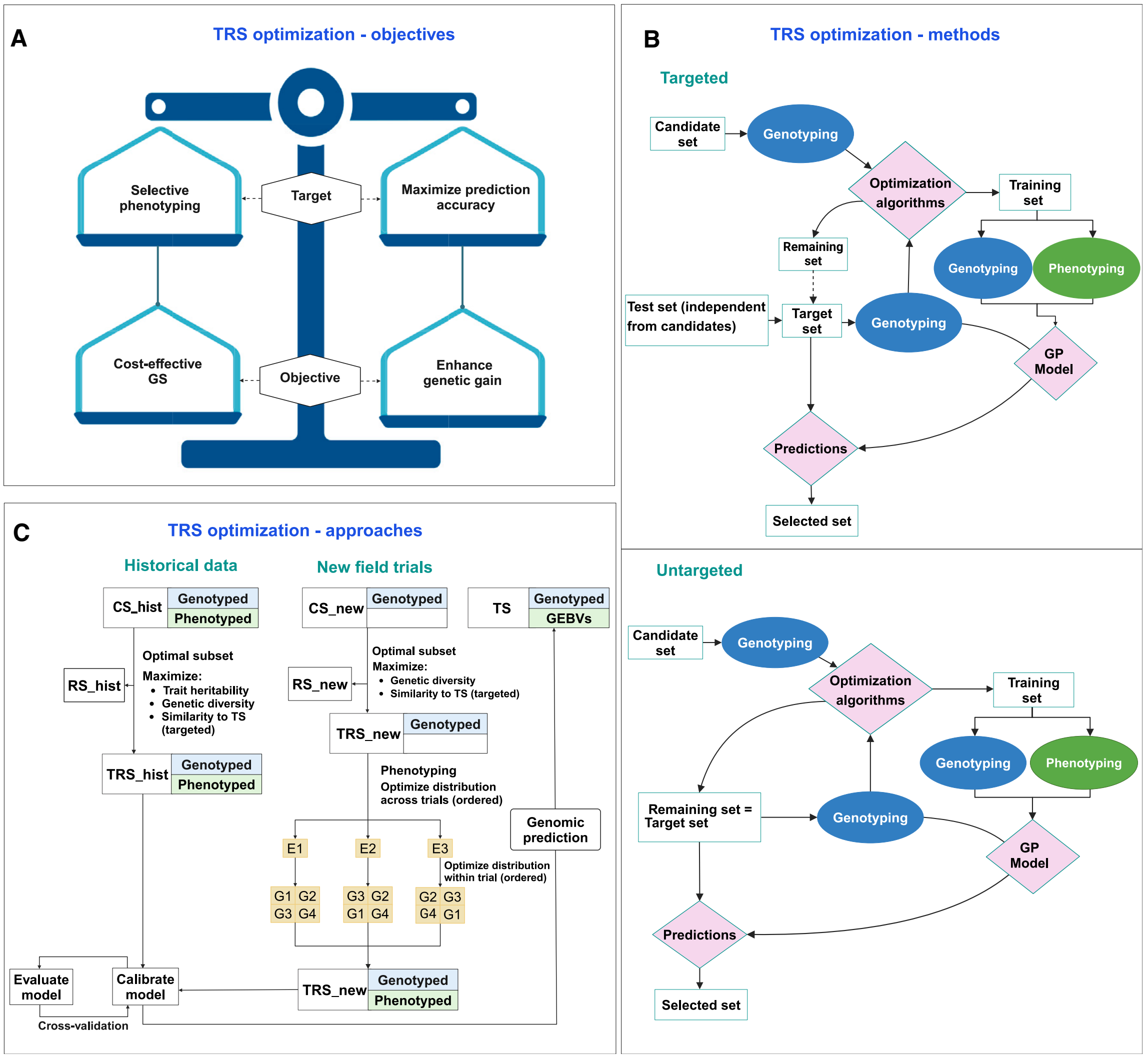

Figure 3. Training population optimization.

(A) The objective of training population optimization is to maximize the genetic gain of plant breeding by enhancing the GP accuracy while minimizing the phenotyping costs by reducing the size of training population.

(B and C) (B) The targeted and untargeted training set optimization methods and **(C)** optimization approaches with historical and new field trials data. The aim of TRS optimization is to find a subset of a CS to be used as an optimal TRS to make predictions on a target population of interest. In targeted optimization, there must be a test set containing genotypes different to those in the CS, which is common when working with historical data. The target population includes this independent test set, but it may contain the genotypes in the RS as well if predicting their genotypic values is of interest. The genotypic information of the target population can then be used as an input on the optimization algorithm that allows the abstention of a training set specifically tailored for it. Conversely, in the untargeted TRS optimization scenario, there is no independent test set, which is common for selective phenotyping of new field trials. In this scenario, the target population comprises all genotypes in the CS whose genotypic values are of special interest. The target population would often be equal to the RS, but it can also be the entirety of the CS. The TRS data may be of two types: historical data and data from new field trials. When both data sources are available, their subsequent TRS can be combined to maximize model performance. It is worth noting that the steps highlighted as targeted mandate the availability of genotypic information from the test set. There may also be instances of population overlap within the process should the GEBVs of the remaining set form a prediction target; for example, RS inherently forms a part of the test set. TRS, training set; CS, candidate set; RS, remaining set; TS, test/target set; E, environment; G, genotype.

find the optimal TRS size have been recently developed (Fernández-González et al., 2023, 2024; Wu et al., 2023). More details are available in Supplemental File 1.

Population structure and genetic relationship with breeding population

One of the pitfalls of GP in a practical breeding scenario is the inability to develop a dependent and effective TRS in the long term without targeting any specific BS. Because of this, breeding programs have to update and optimize the TRS at every single stage where selection is assisted with GP models (see "training population optimization"). This is because the genetic kinship, population structure, and the extent of LD between the training and breeding populations play a huge role in the accuracy. Hence, developing a TRS targeting the candidates for selection is the most critical step in GP (Akdemir et al., 2015; Lorenz and Smith, 2015; Akdemir and Isidro-Sánchez, 2019). Adding genetically unrelated individuals in the training population adversely affects the GP models, as has been shown with a reduction in r_{MG} (Habier et al., 2010; Clark et al., 2012; Lorenz and Smith, 2015; Alemu et al., 2023). For instance, Riedelsheimer et al. (2013) reported a huge decline (42%) in prediction accuracy when the training and breeding population was changed from within full-sib double haploid (DH) maize lines to between half-sib DH lines.

A specific population having distinct allele frequency from others due to founder effects and selection processes creates population structure (Isidro et al., 2015; Norman et al., 2018). This allele frequency difference often bring association between phenotypic performances with markers, irrespective of their true linkage to the causative QTL, which causes bias on r_{MG} unless properly accounted in the GP statistical machine-learning models (Windhausen et al., 2012; Wray et al., 2013; Albrecht et al., 2014; Guo et al., 2014). In GP, population structure can arise within the TRS or between the TRS and BS, and both affect prediction models. Research indicates an adverse impact of population structure on r_{MG} in both self- and cross-pollinated crops (Windhausen et al., 2012; Riedelsheimer et al., 2013; Hickey et al., 2014; Isidro et al., 2015; Würschum et al., 2017; Werner et al., 2020). However, de Los Campos et al. (2015) argued that natural and artificial breeding populations always have different degrees of stratification due to differences in allele frequency and LD patterns that act as a modifier effect rather than a confounding effect. Daetwyler et al. (2012) mentioned that the key is accounting for spurious population structure, such as that originating from admixtures, but without affecting relatedness between individuals. Nevertheless, several research studies indicated a significant reduction in GP accuracy when population structure was accounted in the statistical analysis (Guo et al., 2014; Norman et al., 2018; Werner et al., 2020; Callister et al., 2022). Different strategies have been proposed to account population structure in GP. Admixing individuals from different groups during TRS optimization and phenotyping is one option to connect the different populations (Esfandyari et al., 2015; Rio et al., 2019). Accounting population structure by exploiting the mean performances of subpopulations defined through breeding origin, pedigree, or molecular markers is the other developed method (Albrecht et al., 2011; Windhausen et al., 2012; Guo et al., 2014). Another approach is incorporating principal components and admixture coefficients derived from a genomic relationship matrix as covariates in GP mixed models as fixed effects (Daetwyler et al., 2012; Crossa et al., 2016b; Edriss et al., 2017). However, this method has limitations, such as inability to account markers' effect difference across subpopulations (Lehermeier et al., 2015) and "double counting" of population structure (Janss et al., 2012). Different approaches have been proposed to overcome this problem, such as genomic best linear unbiased prediction (G-BLUP) re-parameterization and modeling genetic covariances between individuals from different groups by adapting multi-trait models (Janss et al., 2012; Guo et al., 2014; Lehermeier et al., 2015).

Genetic diversity

Genetic diversity of the TRS is the other major contributing factor in GP (Habier et al., 2007; Lorenzana and Bernardo, 2009; Norman et al., 2018; Berro et al., 2019). Including individuals with diverse genetic backgrounds helps to capture the full spectrum of genetic variants influencing the target traits. This diversity ensures that the predictive models can accurately capture the genetic effects and make reliable predictions across a wide range of genetic backgrounds. The TRS should encompass a broad range of allelic variation for the traits of interest to capture maximum possible contributing QTLs (Norman et al., 2018). However, it has to be developed targeting the BS, since increasing the diversity with individuals genetically distant from the BS negatively affect the GP model accuracy (Crossa et al., 2014; Akdemir and Isidro-Sánchez, 2019; Berro et al., 2019).

TRAINING POPULATION OPTIMIZATION

The GP efficiency in practical breeding scenarios is highly dependent on the r_{MG} of the genetic merit of candidate individuals. Extensive research supports the notion that configuring the optimal TRS is critical to determine the prediction accuracy (Lorenzana and Bernardo, 2009; Riedelsheimer et al., 2012; Isidro et al., 2015; Akdemir and Isidro-Sánchez, 2019; Berro et al., 2019; Ou and Liao, 2019; Isidro y Sánchez and Akdemir, 2021; Fernández-González et al., 2023). An inadequately constructed TRS substantially diminishes prediction accuracies, while optimized TRS significantly improves accuracy (see Isidro y Sánchez and Akdemir, 2021). The TRS optimization aims to maximize the accuracy of the predictions made on a test or target set (TS) while minimizing the TRS size to reduce phenotyping costs (Figure 3A) (Crossa et al., 2017).

The TRS optimization is key in plant-breeding programs for three main reasons. First, as predictions rely on markers or line effects determined by the TRS, there is a need to carefully curate the TRS to enhance the efficiency and efficacy of GS. Second, the substantial costs of phenotyping have driven the search for innovative alternatives to reduce expenditure (Isidro y Sánchez and Akdemir, 2021). Breeding programs can allocate resources more efficiently by focusing on a smaller yet representative TRS. This not only reduces phenotyping expenditure but also enhances the quality of data applied in the GP models. This allows breeding programs to invest in advanced tools for intricate traits or increase the number of measurements for specific traits, an approach termed sparse or selective phenotyping. Third, the conventional TRS methods that rely on random sampling do not always lead to improved predictive capability due to an under-representation or over-representation of critical genetic information. Thus, optimization serves to streamline the sparse phenotyping process, aiming to curtail phenotyping expenses while preserving or enhancing prediction models' accuracy.

There are two key aspects in TRS optimization: (1) TRS is a dynamic population that must be updated through the breeding cycle (Lorenz and Smith, 2015; Pszczola and Calus, 2016; Akdemir and Isidro-Sánchez, 2019); and (2) the test set needs to be taken into account when building the TRS (Akdemir and Isidro-Sánchez, 2019; Isidro y Sánchez and Akdemir, 2021; Fernández-González et al., 2023).

Here, we review the types of populations available in breeding programs and their role during TRS optimization, the applied methodologies, and the broader implications on GP accuracy and efficiency. We offer the perspectives of TRS optimization in the context of the broader breeding landscape. We do not delve into the exhaustive details of every algorithm or methods and associated pros and cons that can be found elsewhere, such as Isidro y Sánchez and Akdemir (2021). Nevertheless, a summary of the key developed algorithms for TRS optimization can be found in Supplemental Table 4.

Breeding population types involved in optimization

In GS-assisted breeding, the classification and utilization of different breeding population sets are crucial in streamlining the prediction process and maximizing the efficiency of the breeding pipeline. Each set plays a distinct role, and its composition can significantly influence the accuracy and effectiveness of GP. The summary of breeding population sets and their respective purposes and interrelations can be summarized as follows.

(1) Candidate set (CS): collection of genotypes available to breeders. Optimization aims to identify an optimal CS subset to be used as the TRS (Figure 3B).
(2) Remaining set (RS): includes genotypes from the CS not selected for the TRS. When accompanied by phenotypic data, RS enhances the evaluation of model performance.
(3) Training or calibration set (TRS): basis for the GP equation, containing both genotypic and phenotypic data. The goal is to maximize accuracy on the TS with minimal phenotypic and genotypic information.
(4) Test or target set (TS): a set of genotypes to be predicted. It holds only the genotypic information required to predict their GEBVs. However, genotypic information may or may not be available in time for the TRS optimization step.

Optimization scenarios

The TRS is often constructed with new field trials datasets. However, it can be supplemented with old historical data, and optimization can be performed on both data sources (Figure 3C).

(1) (Historical data: utilizing a CS that encompasses comprehensive historical data with both genotyped and phenotyped information can enrich the TRS in terms of size and diversity, a key advantage in GS (Pszczola et al., 2012; Rincent et al., 2012; Isidro y Sánchez and Akdemir, 2021; Fernández-González et al., 2024). Increasing sample size improves the potential to capture the majority of many allele effects and enhance the robustness and accuracy of GP models (Akdemir and Isidro-Sánchez, 2019; Isidro y Sánchez and Akdemir, 2021; Fernández-González et al., 2024). Such inclusion could however diminish the TRS's resemblance to the TS and may adversely affect the prediction accuracy (Lorenz and Smith, 2015), prompting the need for optimization.
(2) New field trials: sparse testing is suggested in cases where the CS provides only genotypic data with limited field trials preventing complete phenotyping (Crespo-Herrera et al., 2021; Montesinos-López et al., 2023a; Melchinger et al., 2023). In this scenario, an optimal experimental design could be designed as follows: (1) determine the subset from CS to undergo field-testing, thereby forming the TRS (TRS optimization); (2) for multi-environment trials, ascertain the ideal TRS genotype distribution across locations; and (3) define the most effective genotype distribution within the field (which genotype in which plots). Steps 2 and 3 represent ordered optimization focusing on the strategic optimal spatial arrangement of genotypes.

TRS optimization is categorized as either targeted or untargeted depending on the availability of genotypic information from the TS (Figure 3B). Targeted optimization takes advantage of TS genotypic information to construct the TRS and often outperforms the untargeted approaches (Akdemir and Isidro-Sánchez, 2019; Isidro y Sánchez and Akdemir, 2021; Fernández-González et al., 2023). Even without TS genotype information, a detailed pedigree linking the CS with TS remains feasible in targeted optimization. However, research in this area is lacking. Unordered optimization focuses on selecting a CS subset, while ordered optimization emphasizes the spatial genotype distribution in the field. The latter may utilize data related to blocking structures, spatial influences, and environmental variables (Akdemir et al., 2021).

Training population optimization algorithms

Several design criteria have been proposed for selecting and optimizing the TRS in GP. The classical standard random or stratified sampling method is commonly applied because of its simplicity. Nevertheless, GP accuracy enhancement has been achieved using other optimization criteria, which can be classified as parametric, non-parametric, and multiple design criteria. Many of the established criteria mostly serve as evaluation metrics for the TRS, and appropriate heuristic is imperative to maximize or minimize it. Numerous R program packages have been developed and provide suitable heuristics often based on genetic algorithms. For instance, the STPGA (Akdemir, 2017), TSDFGS (Ou and Liao, 2019), and odw (Butler et al., 2013) are developed but are limited to built-in criteria. In contrast, TrainSel (Akdemir et al., 2021) supports both built-in and user-defined criteria.

Parametric design criteria

Parametric design criteria assume that the researcher predetermines a model prior to data collection. These criteria typically rely on a scalar function tied to the model's information matrix. In practice, it is usually derived from the prediction error variance-covariance matrix (PEV) for the additive genotypic effects in linear mixed models. The A, D, and E criteria (Laloë, 1993), the coefficient of determination criterion (CD_mean), and the prediction error variance criterion (PEV_mean) are examples (Laloë, 1993). Parametric criteria are a powerful approach but are computationally intensive. Attempts have been made to resolve this problem, including updating the PEV matrix in each iteration instead of calculating *de novo* (Butler et al., 2013) and applying principal component analysis to reduce dimensionality such as in PEV_mean$^{\text{ridge}}$ and CD_mean$^{\text{ridge}}$ methods (Akdemir et al., 2015; Heslot and Feoktistov, 2020). An in-depth discussion on computational efficiency of algorithms is available in Supplemental File 1, Note 2.

Sparse selection index is a recently proposed prediction model with a built-in optimization process (Lopez-Cruz and de los Campos, 2021; Lopez-Cruz et al., 2021; Lopez-Cruz et al., 2022). Here, a selection index that specifies the TS genotypic values as a linear combination of the CS ones is defined. The regression coefficients of the linear combination are subjected to a lasso regularization (L1) penalty to enforce sparsity that is equivalent to the selection of a subset of the genotypes as a TRS. This is conceptually similar to the bandwidth parameter in the reproducing kernel Hilbert space (De Los Campos et al., 2009; Lopez-Cruz et al., 2021) but takes it one-step further. This method is suitable for historical data optimization because it makes a specific TRS for each TS individual and the phenotypic information of the CS should be available for parameter tuning.

Non-parametric design criteria

The methods of this type do not assume any predefined underlying models but often revolve around metrics of distance or similarity with the intention of uniformly distributing the TRS throughout the design landscape, a method known as space-filling design. Such designs particularly facilitate the selection of a condensed set of candidates and minimize the computational complexity associated with the optimization of parametric design criteria. Various metrics assist in evaluating the distribution of points within this design space. For instance, the partition around medoids approach centers on identifying a series of central entities, termed medoids, within clusters based on a specific distance measure (Guo et al., 2019). In general, methods for representative subset selection in data mining could be used for training set design, which opens up many possibilities. Numerous metrics have been developed to minimize genetic relationships within the TRS (i.e., maximizing diversity) and/or maximizing its relationship to the TS, for example, the maximin and minimax (Johnson et al., 1990), Avg_GRM (Atanda et al., 2021a), OPT_MIN (Lemeunier et al., 2022), Avg_GRM self, and Avg_GRM_MinMax (Fernández-González et al., 2023). Latin hypercube sampling (Helton and Davis, 2003) involves segmenting the design space into equal cubes. The objective is to ensure that each cube hosts a single sample point, further aiming to comprehensively explore the range of each scalar input in alignment with a given probability distribution. Tails and Tails_GEBVs select genotypes with extreme phenotype or GEBVs for the TRS and discard the rest (Neyhart et al., 2017; Fernández-González et al., 2024). Adversarial selection tries to ensure that the TRS and TS are indistinguishable by a binary classifier; i.e., their similarity is maximized (Montesinos-López and Montesinos-López, 2023).

Multi-objective criteria

This method attempts to handle the choice issue by combining the different criteria into one with some type of averaging methods such as the Pareto front approach (Akdemir et al., 2015; Isidro y Sánchez and Akdemir, 2021). It is adept at evaluating multiple criteria and defining a suite of non-dominated designs. The method has been effectively applied to optimize the integration of historical data balancing the TRS diversity, its association with the TS, and trial heritability with an extensive empirical dataset from an industrial breeding program (Isidro y Sánchez and Akdemir, 2021; Fernández-González et al., 2024).

Summary of the key algorithms

A detailed comparison and breakdown of TRS optimization methods and algorithms can be found in Supplemental Table 4. However, the large number of available methods makes selecting a single effective method challenging. Recent research on TRS optimization offers invaluable insights into selection of appropriate methods. Fernández-González et al. (2023) conducted an extensive comparison of these optimization methods across various datasets and genetic architectures. In light of their findings and those of other researchers, we provided a small summary focusing on the key algorithms for each field within TRS optimization. Furthermore, an in-depth, systematic example on the implementation of CDmean and Avg_GRM_self is provided in Supplemental File 1, Note 1 and an R-script with example of implementing two algorithms on real TRS optimization scenario (Supplemental File 2). Nevertheless, it is important to note that there is no single best algorithm in all aspects, and some of the methods in Supplemental Table 4 may be more suitable for niche applications.

Here, we forward our recommendation of general-purpose, effective methods that constitute a good first option for new optimization projects.

(1) TRS size optimization: tailored mainly for situations incorporating historical data, two main algorithms emerge:
   ○Target accuracy methods: aimed at predicting GP accuracy and identifying the minimal TRS size without substantial accuracy loss (Fernández-González et al., 2023; Wu et al., 2023). In our experience Avg_GRM_self is the best option due to its fast computational time, essential in this application. Importantly, budgetary constraints play a crucial role, but typically, including 50%–85% of the entire candidate population maintains an accuracy decrease below 5%.
   ○Best solution methods: these seek the optimal TRS size by identifying local maxima or inflection points. For example, Avg_GRM_MinMax (Fernández-González et al., 2023) and Min_GRM (Fernández-González et al., 2024).
(2) Optimizing TRS composition: an extensively researched area. Findings suggest that targeted optimization usually surpasses untargeted methods, with CDmean being highly efficient, albeit computationally intensive. Maintaining TRS diversity is especially important in the presence of a strong population structure. Therefore, it is advisable to apply CDmean for smaller datasets, while the fast Avg_GRM_self (untargeted) or Avg_GRM_MinMax (targeted) are suitable for larger datasets.
(3) Simultaneous size and composition optimization: beneficial when utilizing historical data and the training set size is not determined by available field resources, although it may reduce the versatility of the algorithm for optimization of new field trials due to the potential difficulty of matching optimal TRS size to actual field resources. MaxCD (Guo et al., 2019) was originally described for TRS design in hybrid breeding, but we believe that its ability of optimizing TRS size could be useful for optimization in historical data. The latter role could be filled by other methods such as adversarial selection (Montesinos-Lopez et al., 2023a, 2023b) or multi-objective optimization (Akdemir et al., 2021; Fernández-González et al., 2024), which are not specific to hybrids.
(4) Spatial distribution/ordered optimization: pertinent for new field trials, this optimization is computationally demanding, especially when incorporating environmental or spatial data. Two notable R package algorithms, “odw” (Butler

et al., 2013) and “TrainSel” (Akdemir et al., 2021), have been developed for this purpose. Parametric criteria such as A-opt and CDmean are the best-suited approaches for this application

DENSITY AND DISTRIBUTION OF MARKERS AND LINKAGE DISEQUILIBRIUM

Increasing the density of SNP markers distributed across chromosomes helps to accurately capture most contributing QTL ultimately leading to an increased r_{MG}. The number of SNP markers required to develop an optimum GP depends on the genome size, extent of LD, and complexity of the trait under investigation. A study targeting a complex trait controlled by several QTLs (e.g., yield) in a crop with a large genome size and low LD relatively requires a highly dense SNP marker distributed across chromosomes. On the contrary, a highly heritable trait controlled by fewer genes and with high LD could need a relatively low SNP marker density to reach the maximum possible r_{MG}. In addition, LD in outcrossing crops, such as maize, decays rapidly compared to self-crossing crops (e.g., rice) (Flint-Garcia et al., 2003; Kaler et al., 2022), requiring highly dense SNP marker distribution to achieve the optimum r_{MG}. In general, the optimum density and distribution of SNP markers relies on the most contributing QTL of a target trait being under LD with DNA markers included in prediction models (Hayes and Goddard, 2001; Kaler et al., 2022). The pattern of LD of populations particularly helps to develop GP models with cost-effective, low-density SNP markers (Bolormaa et al., 2015; Wu et al., 2016; Silva et al., 2018; Ballesta et al., 2020).

Optimizing the marker density in GP could be beneficial, as most SNPs in large marker datasets are phenotypically neutral and contain only a relatively small proportion of SNPs relevant for a specific trait (Bermingham et al., 2015; Al Kalaldeh et al., 2019; Weber et al., 2023). Selecting optimal subsets of markers for specific traits has been a promising approach to increase the accuracy in GP (Bermingham et al., 2015; van den Berg et al., 2016; Filho et al., 2019; Alemu et al., 2023; Weber et al., 2023). One method for marker subsetting is selection based on previous association mapping studies. A beneficial GP accuracy improvement was observed when significant markers identified through GWAS were fitted as fixed effects (Kim et al., 2022; Anilkumar et al., 2023; Chen et al., 2023), only the top 100–10 000 markers with highest significance were used as predictors (Bermingham et al., 2015; Filho et al., 2019), or markers surrounding the significant markers were included (van den Berg et al., 2016; Filho et al., 2019). Another optimization approach is applying haplotype blocks based on marker LD in prediction models (Alemu et al., 2023; Weber et al., 2023). Predictions based on haplotype blocks, rather than single SNP markers, could efficiently capture local epistasis and better account for LD to QTLs leading to improved GP accuracy (Weber et al., 2023). Selection of marker panels can highly influence trait associations, and several research articles have demonstrated the impact of marker density on the GP accuracy (Zhang et al., 2017a, 2019; Liu et al., 2018; Norman et al., 2018).

GENETIC ARCHITECTURE AND HERITABILITY OF TRAITS

Most crop traits of economic importance, such as yield, are multigenic and have a complex genetic architecture involving several QTLs or genes with varying levels of phenotypic effects. One of the key advantages of GP over the conventional MAS methods is its ability to efficiently evaluate genotypes for such genetically complex multigenic traits by considering the high numbers of small-effect QTLs. Generally, genetic complexity and heritability (h^2) are directly related to the number of QTLs and their interaction to control a trait. A trait controlled by small numbers of large-effect QTLs usually has higher heritability than those with several genes with different levels of genotypic effect. GP is affected by the complexity of traits, genetic architecture, and heritability. A trait with low h^2 should be compensated by increasing the TRS size (N) to achieve an optimum GP accuracy, since Nh^2 determines the power of GP models (Bernardo, 2016). Furthermore, machine-learning models that account for epistatic interactions have the potential to improve the prediction accuracy when epistatic interaction largely or partially contributes to the true genetic architecture of a trait (De Los Campos et al., 2010; Wang et al., 2012; Morgante et al., 2018). Several empirical investigations and simulation research have demonstrated that the r_{MG} generally increases as the number of QTLs decreases and trait heritability increases (Hayes et al., 2009; Lorenzana and Bernardo, 2009; Zhong et al., 2009; Jannink et al., 2010; Combs and Bernardo, 2013; Zhang et al., 2017a; Jung et al., 2020).

PRECISION PHENOTYPING

The phenotypic data recorded from the TRS is required to connect the genomic profile with the phenotype, enabling GP models to evaluate and provide weights to individual SNP markers. These markers are then used to assess individuals in the BS solely from their genomic profile and assist selection and decision making in breeding programs. High-density SNP markers combined with precision phenotyping evaluated in suitable statistical machine-learning models could link the genome with the phenome of crops, leading to GP models with high prediction accuracy. Efficiency limitations of the conventional plant phenotyping methods have been considered as the bottleneck to successfully connecting the bridge between genotype with phenotype information (Araus and Cairns, 2014; Araus et al., 2018). Hence, advanced technologies for high-throughput phenotyping (HTP) and high-throughput field phenotyping (HTFP) methods have attracted tremendous attention recently for their potential to provide comprehensive and precise phenotypic data for primary as well as secondary traits in several crops (Cabrera-Bosquet et al., 2012; Araus and Cairns, 2014; Zhang et al., 2017b; Araus et al., 2018; Moreira et al., 2020). The HTP and HTFP can be referred to collectively as the high-throughput phenotyping platform (HTPP). The HTPP allows researchers to screen massive numbers of individual plants at a very low cost. HTPP aims to produce high-density phenotypes on very large numbers of individuals or breeding lines across time and space at low cost using remote or proximal sensing. This can increase both the accuracy and intensity of selection and, therefore, the selection response while decreasing phenotyping costs. The main idea of HTPP is to use predictor traits related to grain yield, disease resistance, or end-use quality that could be advantageous in early-generation testing of lines (Rutkoski et al., 2016). Previous research has shown the potential of HTPP methods in the GP accuracy of several traits (Crain et al., 2018; Juliana et al., 2019a; Galán et al., 2020; Wang et al., 2023b).

INTEGRATING OTHER OMICS DATA

GP relies on estimating the phenotypic performance of individuals from their genomic profile. The genomic profile, however, must be transcribed to RNA (tRNA, sRNA, mRNA) and then translated to protein before being expressed as a phenotype (Cobb, 2017). These results come from transcriptomics and proteomics research, respectively. The integration of this intermediate phenotype data (i.e., transcriptomics, proteomics, and metabolomics) with genomic data has demonstrated the potential to improve GP accuracy (Hu et al., 2019; Li et al., 2019; Haile et al., 2020; Martini et al., 2022; Wang et al., 2023a, 2023b). Multi-omics-based GP has been implemented successfully, improving prediction accuracy in several crops such as maize (Guo et al., 2016; Zenke-Philippi et al., 2016; Westhues et al., 2017; Xu et al., 2017; Schrag et al., 2018), wheat (Zhao et al., 2015), oats (Hu et al., 2021), barley (Wu et al., 2022), rice (Hu et al., 2019; Wang et al., 2019), and rapeseed (Knoch et al., 2021).

Schrag et al. (2018) reported combining messenger RNA (mRNA) with pedigree and genomic datasets, resulting in beneficial improvements in r_{MG} to estimate the breeding values of agronomic traits in untested maize hybrids. Hu et al. (2019) outlined an r_{MG} improvement in four yield and yield-related traits of untested rice RILs through a multi-layered least absolute shrinkage and selection operator model integrating transcriptome and metabolome along with genomic profiles in a single model. Incorporating both transcriptomic and metabolomic profiles to the genomic datasets has also improved the r_{MG} of several agronomic and seed nutritional traits of oats from multi-environment trials (Hu et al., 2021). Recently, advanced statistical machine-learning algorithms have been developed to incorporate the multi-omics intermediaries with efficient computing performance to leverage the GP models (Hu et al., 2021; Wang et al., 2023a, 2023b). Nonetheless, model overfitting and spatial-temporal features accompanying the intermediaries should be cautiously considered during implementation of the omics profiles of plants in GP research (Yan and Wang, 2023).

STATISTICAL MACHINE-LEARNING METHODS

Statistical methods play a central role in GP, since the effect of DNA markers is estimated by modeling the mathematical relationships between the genotypic and phenotypic data provided in the TRS. Thereafter, evaluated markers are provided with specific weights to their phenotypic effect that allow the genomic breeding values of candidate individuals in the BS to be estimated. Hence, GP is a statistical machine-learning approach that aims to train, develop, and analyze the performance of models with the data from the TRS (Tong and Nikoloski, 2021; Montesinos López et al., 2022a, 2022b, 2023). Meuwissen et al. (2001) by simulating the effect of approximately 50 000 marker haplotypes with a modified linear least squares regression, BLUP, and two Bayesian statistical methods (BayesA and BayesB).

Several statistical machine-learning methods have been proposed for GP over the last 20 years (Figure 4). As most of the available datasets in genomics for plant and animal breeding applications try to find the relationship between the response variable (output) and thousands or even millions of SNP markers as inputs (or predictors, p), the framework for training these models is where more inputs are available than observations (observations, n), that is, $p >> n$, which presents a major challenge. This phenomenon leads the advent of different penalization (regularization) approaches (Meuwissen et al., 2001; De Los Campos et al., 2013). Hence, the different regularization mechanisms applied to estimate genome-wide SNP marker effects from a regression with large p with small n has led to the emergence of numerous statistical machine-learning approaches (Montesinos López et al., 2022a, 2022b). These statistical machine-learning algorithms perform differently, and their suitability and performance depend on coping with several factors that arise from the multi-dimensionality of genome-wide SNP markers and the genetic complexity of multi-factorial traits (De Los Campos et al., 2010). Consequently, no statistical machine-learning model can be singled out as outperforming other available algorithms and giving the highest possible GP accuracy that has been verified by numerous empirical and simulation researches and its theoretical support given by the "no-free-lunch" theorem (Azodi et al., 2019; Montesinos López et al., 2022a, 2022b). These statistical models can be grouped as parametric, semi-parametric, and non-parametric models (Montesinos López et al., 2022a, 2022b).

Parametric models are developed from the assumption that the independent or predictor variables take a predetermined function with the dependent or response variables. Some examples of parametric models are linear multiple regression, non-linear regression, logistic regression, multinomial regression, and Poisson regression (Montesinos López et al., 2022a, 2022b). Non-parametric models are a class of statistical and machine-learning models that do not make explicit assumptions about the functional form or distribution of the underlying data. Predictors are not predefined in this class of models but are instead crafted on the basis of insights extracted from the data (Montesinos López et al., 2022a, 2022b). Unlike parametric models, which assume specific mathematical forms for relationships between variables (e.g., linear regression), non-parametric models offer more flexibility by allowing the data to determine the structure of the model. These models are particularly useful when dealing with complex or unknown relationships, as they can adapt to various data patterns without requiring predefined parameter specifications. Non-parametric models include methods such as kernel density estimation, k-nearest neighbors, decision trees, gradient boosting machine, and random forest. A semi-parametric model is a statistical machine-learning approach where a portion of the predictors is not constrained to predetermined mathematical forms, while another portion adheres to known functional relationships with the response variable. This blend of flexibility and structure is exemplified by equations such as

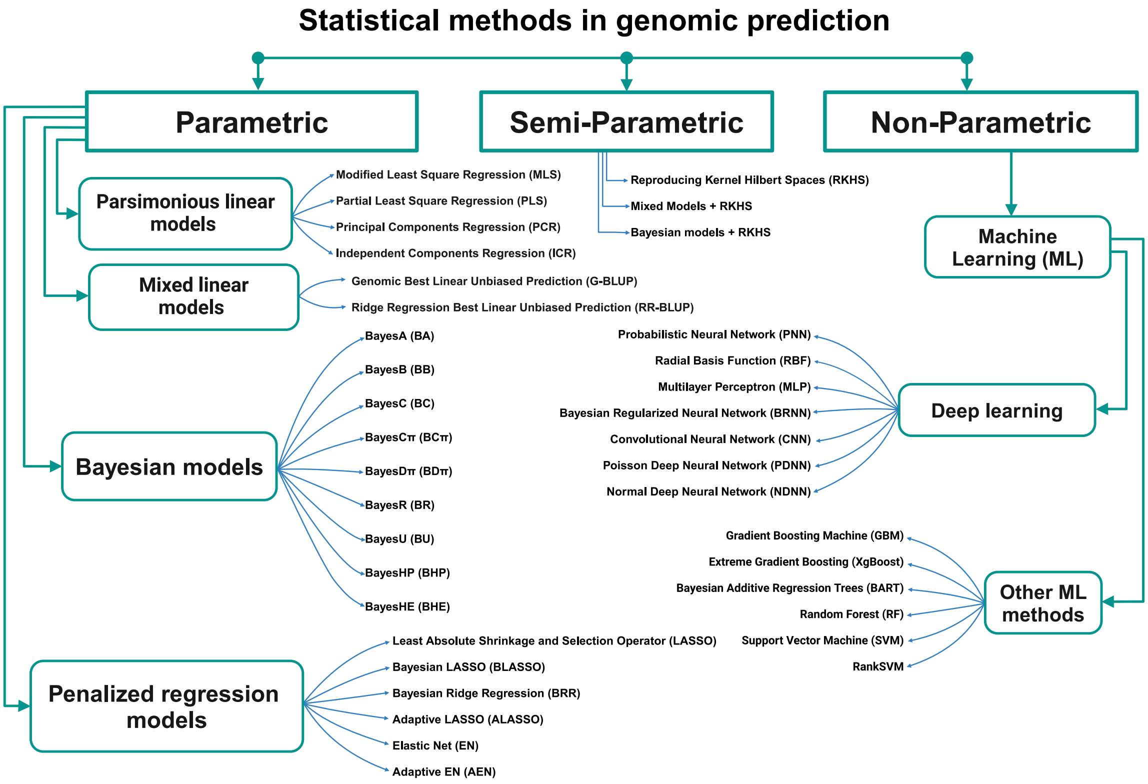
Statistical methods in genomic prediction
Parametric
Semi-Parametric
Non-Parametric
Parsimonious linear models
Modified Least Square Regression (MLS)
Partial Least Square Regression (PLS)
Principal Components Regression (PCR)
Independent Components Regression (ICR)
Mixed linear models
Genomic Best Linear Unbiased Prediction (G-BLUP)
Ridge Regression Best Linear Unbiased Prediction (RR-BLUP)
Bayesian models
BayesA (BA)
BayesB (BB)
BayesC (BC)
BayesCπ (BCπ)
BayesDπ (BDπ)
BayesR (BR)
BayesU (BU)
BayesHP (BHP)
BayesHE (BHE)
Penalized regression models
Least Absolute Shrinkage and Selection Operator (LASSO)
Bayesian LASSO (BLASSO)
Bayesian Ridge Regression (BRR)
Adaptive LASSO (ALASSO)
Elastic Net (EN)
Adaptive EN (AEN)
Reproducing Kernel Hilbert Spaces (RKHS)
Mixed Models + RKHS
Bayesian models + RKHS
Machine Learning (ML)
Deep learning
Probabilistic Neural Network (PNN)
Radial Basis Function (RBF)
Multilayer Perceptron (MLP)
Bayesian Regularized Neural Network (BRNN)
Convolutional Neural Network (CNN)
Poisson Deep Neural Network (PDNN)
Normal Deep Neural Network (NDNN)
Other ML methods
Gradient Boosting Machine (GBM)
Extreme Gradient Boosting (XgBoost)
Bayesian Additive Regression Trees (BART)
Random Forest (RF)
Support Vector Machine (SVM)
RankSVM

Figure 4. List of the statistical machine-learning models currently in use for genomic prediction.

All these statistical machine-learning models are classified into three major categories: parametric, semi-parametric, and non-parametric. The parametric statistical machine-learning models include modified least square regression (MLS) (Meuwissen et al., 2001), partial least square regression (PLS) (Montesinos López et al., 2022a; 2022b), principal components regression (PCR) (Solberg et al., 2009), independent components regression (ICR) (Azevedo et al., 2013), genomic best linear unbiased prediction (G-BLUP) (Vanraden, 2008), BayesA (BA) and BayesB (BB) (Meuwissen et al., 2001), BayesC (BC) (George and McCulloch, 1993), BayesCπ (BCπ) and BayesDπ (BDπ) (Habier et al., 2011), BayesR (BR) (Erbe et al., 2012), BayesU (BU) (Pong-Wong and Woolliams, 2014), BayesHP (BHP) and BayesHE (BHE) (Shi et al., 2021), least absolute shrinkage and selection operator (LASSO) (Usai et al., 2009), adaptive LASSO (ALASSO) (Zou, 2006), Bayesian LASSO (BLASSO) (Park and Casella, 2008), ridge regression best linear unbiased prediction (RR-BLUP) (Meuwissen et al., 2001), Bayesian ridge regression (BRR) (Pérez et al., 2010), elastic net (EN) (Zou and Hastie, 2005), and adaptive EN (AEN) (Zou and Zhang, 2009). The semi-parametric method includes the reproducing kernel Hilbert space (RKHS) model (Gianola et al., 2006) and the mixed and Bayesian models combined with the RKHS model. The non-parametric method comprises gradient boosting machine (GBM) (Li et al., 2018), extreme gradient boosting (XgBoost) (Chen and He, 2014), support vector machine (SVM) (Maenhout et al., 2007), rankSVM (Blondel et al., 2015), Bayesian additive regression trees (BART) (Waldmann, 2016), random forest (RF) (Chen and Ishwaran, 2012), probabilistic neural network (PNN) (González-Camacho et al., 2016), radial basis function (RBF) (Chen and Ishwaran, 2012), multilayer perceptron (MLP) (Gianola et al., 2011), Bayesian regularized neural network (BRNN) (Pérez-Rodríguez et al., 2012), convolutional neural network (CNN) (Ma et al., 2018), Poisson deep neural network (PDNN), and normal deep neural network (NDNN) (Montesinos-López et al., 2020).

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + m(x) + \epsilon, \qquad \text{(Equation 1)}$$

$$y = \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) + m(x) + \epsilon. \qquad \text{(Equation 2)}$$

In the context of GP models, a classical example is Bayesian or mixed models with linear components for environmental effects and non-linear (Gaussian kernel, or other types of kernels) components for genotype effects (Montesinos López et al., 2022a, 2022b). Essentially, semi-parametric models represent a combination of both parametric and non-parametric modeling techniques. Most of the currently available statistical machine-learning models classified in the three mentioned groups are presented in Figure 4.

Modeling genotype × environment interaction

In studies involving multiple environments, genetic association and prediction models are usually developed from summarized phenotypic data across environments or separate models for each environment. Another approach to account for multiple environments is using an environment index, which for example can be derived from environmental conditions such as temperature and photoperiod (Guo et al., 2020a, 2020b, 2020c, 2020d). Growing degree days was earlier proposed as a promising example of an environment index for capturing flowering time plasticity in rice (Guo et al. 2020a). Li et al. (2021) proposed that a carefully developed environment index can replace phenotypic means obtained through conventional measurements and can model observed phenotype and also predict phenotypic performance in new environments, and they tested their hypothesis on three different traits in wheat and oat field trials. Similarly, in sorghum, diurnal temperature range during the rapid growth period was found to be an effective environment index (Mu et al., 2022). Taken together, these studies highlight the importance of studying phenotypic plasticity under G×E interaction and exploring derived environmental indices for modeling and predicting phenotypes in untested environments.

The reaction norm model

Multi-environment trials for assessing G×E play an important role in plant breeding for selecting high-performing and stable lines across environments. For instance, the multi-environment linear mixed models accounting for correlated environmental structures within the G-BLUP framework increased accuracy when predicting the performance of unobserved phenotypes using pedigree and molecular markers (Zhang et al., 2015). Burgueño et al. (2012) proposed and effectively applied a marker and pedigree G-BLUP model for assessing G×E, while Heslot et al. (2014) incorporated crop modeling data on the genomic G×E prediction. Jarquín et al. (2014) developed a reaction norm model, an extension of the G-BLUP model, where the main and interaction effects of markers and environmental covariates are introduced using highly dimensional random variance-covariance structures of markers and environmental covariables. The model has been successfully applied in GP prediction of breeding values using pedigree and genomic relationships (Pérez-Rodríguez et al., 2015; Velu et al., 2018).

Here, the baseline model for the phenotypes (y_{ij}) can be described as

$$y_{ij} \;=\; \mu + E_i + L_j + EL_{ij} + e_{ij}, \qquad \text{(Equation 3)}$$

where μ is the overall mean, E_i $(i = 1,\ldots,I)$ is the random effect of the ith environment, L_j is the random effect of the jth line (j = 1,…,J), EL_{ij} is the interaction between the ith environment and the jth line, and e_{ij} is the random error term. The assumptions are as follows: $E_i \overset{\text{iid}}{\sim} N(0,\,\sigma_E^2)$, $L_j \overset{\text{iid}}{\sim} N(0,\,\sigma_L^2)$, $EL_{ij} \overset{\text{iid}}{\sim} N(0,\,\sigma_{EL}^2)$, and $e_{ij} \overset{\text{iid}}{\sim} N(0,\sigma_e^2)$, with $N(.,.)$ denoting a normal density and iid standing for independent and identically distributed. Markers can be introduced in Equation 3 such that the effect of line (L_j) can be replaced by g_j defined by the regression on marker covariates (it approximates the genetic value of the jth line). The vector containing the genomic values is $g \sim N(0, \boldsymbol{G}\sigma_g^2)$, where σ_g^2 is the genomic variance and $\boldsymbol{G}$ is a genomic relationship matrix (Vanraden, 2008). Furthermore, the effects of line (L_j) can be replaced by a_j, with $\boldsymbol{a} \sim N(0, \boldsymbol{A}\sigma_a^2)$, where $\boldsymbol{A}$ is the additive relationship matrix derived from pedigree and σ_a^2 is the additive variance.

The marker × environment interaction model

The marker × environment (M×E) interaction model proposed by Lopez-Cruz et al. (2015) breaks down the marker effects into components that are common across environments (stability) and environment-specific deviations (interaction). This model borrows information across environments while allowing marker effects to change across environments. This method can be implemented using both shrinkage and variable selection methods and thus can be used to identify genomic regions with stable effect across environments and regions that are responsible for G× E. However, it is noteworthy that the M×E model is best suited for joint analysis of positively correlated environments (Lopez-Cruz et al., 2015). Crossa et al. (2016a) successfully applied the M× E GP model to predict untested individuals and dissect genomic regions with stable effect across environments and with environment-specific effect.

IMPLEMENTING GENOMIC PREDICTION AT DIFFERENT BREEDING STAGES

There are various uses of GP in breeding crops. The first is in pre-breeding, either to search for desired accessions based on their GEBVs in a gene bank (Crossa et al., 2016b; Dzievit et al., 2021; Bohra et al., 2022; El Hanafi et al., 2023) or to identify elite parents for further crossing (Gaynor et al., 2017; Chung and Liao, 2022). GP allows a cost-effective approach for selecting interesting germplasm held in gene banks (Yu et al., 2016), thus increasing the use of this germplasm—particularly those lacking pedigree information and data evaluation—in plant breeding (Jiang et al., 2021). It also speeds up the introgression of exotic germplasm into the elite breeding pool (Crossa et al., 2016b), as shown recently in wheat improvement (Schulthess et al., 2022). GP may also be used for increasing genetic gains by selecting promising germplasm at early stages (Kadam et al., 2016; Rembe et al., 2022) or for feeding them into a genomic recurrent selection (GRS) approach (Bassi et al., 2016; Biswas et al., 2023), as well as for accelerating the cultivar development pipeline (Ballén-Taborda et al., 2022). GRS facilitates the recycling of parents in a breeding program. The success of GP in any of these breeding stages, however, relies mainly on the trait architecture and its heritability.

A challenge faced by plant breeding is to predict performance across sites over years or cropping seasons. GP may allow estimation of the robustness of desired productivity or quality traits across the target population of environments. Such an approach improves the efficiency of multi-environment testing and its further use in the cultivar development pipeline because it eliminates mediocre breeding lines in the early stages, thus saving time and resources. In this regard, as shown by Atanda et al. (2021b), sparse testing using GP may also be a valuable approach for increasing the number of trial environments without increasing costs but keeping the selection intensity in the early stages of evaluation. Montesinos-López et al. (2023b) showed that a significant gain in the number of new lines to be evaluated could be obtained by using sparse testing methods without a relevant increase of required resources. The authors demonstrated that with a conventional block design capacity to evaluate only 225 lines, the number could be increased to 269, 308, and 475 with a sparse testing design using 85%, 75%, and 50% as training increasing the number of lines by 19.56%, 36.89%, and 111.11%, respectively.

GP has further found extensive application in predicting heterosis, encompassing both high-parent and mid-parent heterosis, across a diverse range of crops, including maize (Albrecht et al., 2011, 2014; Riedelsheimer et al., 2013; Beyene et al., 2015, 2019; Cantelmo et al., 2017; Zhang et al., 2022), rice (Xu et al., 2014, 2018; Huang et al., 2015; Cui et al., 2020), barley (Philipp et al., 2016; Li et al., 2017), wheat (Basnet et al., 2019; Zhao et al., 2021), sorghum (Sapkota et al., 2022; Kent et al., 2023; Maulana et al., 2023), ryegrass (Grinberg et al., 2016), and pumpkin (Wu et al., 2019). Notably, the predictive scope of GP extends beyond conventional traits such as yield and its components (Grinberg et al., 2016; He et al., 2016; Philipp et al., 2016; Wu et al., 2019) to encompass a wider spectrum of characteristics, such as biotic and abiotic stress tolerances (Lorenz et al., 2012; Arojju et al., 2018), nutrient utilization efficiency (Zhao et al., 2020), and biofortification of crops with several micronutrients (Velu et al., 2016; Mageto et al., 2020; Rakotondramanana et al., 2022; Tadesse et al., 2023).

ACHIEVEMENTS

The task of applying GS in breeding is to enhance genetic gains per year at a lower cost and in less time compared to the conventional breeding methods. Given a vector of true breeding values of an individual $\mathbf{a}' = [a_1\ a_2 \ldots\ a_t]$ and the vector of economic weights $\mathbf{w}' = [w_1\ w_2 \ldots\ w_t]$ for t traits, the net genetic merit is $\mathrm{H} = \mathbf{w}'\mathbf{a}'$. The response to multi-trait genetic gains can be written as $\mathrm{H} = (k\sigma_{\mathrm{H}}\rho_{\mathrm{H,I}})/L$, where k is intensity of selection (the standardized selection differential), σ_{H} is the standard deviation of H, $\rho_{\mathrm{H,I}}$ is the correlation between H and any phenotypic or genomic index I, and L is the time required for I to complete one selection cycle (in a standard breeding program this takes several years). The selection response is the most important breeder's equation, and factors that increase the numerator or decrease the denominator of R will increase the overall genetic gains of the target traits. Simulation and empirical results have shown that GS can increase genetic gains by shortening the breeding interval cycle (L) (rapid selection cycle) or increasing testing efficiency by performing sparse field evaluation (Tessema et al., 2020; Xu et al., 2020; Atanda et al., 2022; Dreisigacker et al., 2023). To achieve a shorter interval cycle (l), the most favorable situation for GS is prediction within full-sib families, since the biparental populations have very high LD between marker alleles and QTL alleles with no pedigree, family, or group structure. Estimated prediction accuracies for biparental populations should thus be considered the maximum attainable in closed rapid-cycle marker-only selection. Several research confirmed the efficiency of GS for early-generation rapid cycling (Massman et al., 2013; Zhang et al., 2017c; Bonnett et al., 2022; Dreisigacker et al., 2023).

Two showcases are provided to elucidate the ongoing empirical research facilitated by GS from public and private breeding programs.

Showcase 1: Genetic gains of maize in Africa

Most GS results in maize have been achieved by rapid cycling of biparental populations. For example, the $F_{2:3}$ segregating populations were crossed with a tester, usually from the opposite heterotic group. CIMMYT's Global Maize Program designed a GS rapid cycle of multi-parental crosses. Fifteen elite tropical maize lines were crossed in diallel fashion to form cycle 0 (C_0) comprising 1000 plants, which were genotyped with 1 000 000 genotyping-by-sequencing (GBS) SNP markers and phenotyped at three locations in Mexico. The best phenotypic plants were selected to form the parents for GS cycle 1 (C_1). The C_1 parents were intercrossed and the progeny was genotyped with the same GBS markers as used for the C_0 population. Genomic-enabled prediction for grain yield for the individuals in the C_1 population was performed in each of the three environments; based on the predicted values, selection was made to form the parents of the C_2 population. As before, the parents were intercrossed and genotyped to form the C_2 population, and plants were selected based on the GP for grain yield. GP and GS were performed for two more cycles. Two cycles per year were performed; and at the end of the second year, seeds from cycles C_0, $C_{1,}$ $C_{2,}$ $C_{3,}$ and C_4 were collected, assembled, and sown at three locations in Mexico (Agua Fria, Cotaxtla, and Tlaltizapan). Fifty entries were sown per genomic cycle at each location, together with two widely used commercial tropical maize hybrids. The average genetic grain yield gains were 0.134 t ha^{-1} with C_0 producing 6.653 t ha^{-1}. Grain yield of C_1 was slightly lower (6.488), and cycles $C_{2,}$ $C_{3,}$ and C_4 produced means of 7.022, 6.879, and 7.126 t ha^{-1}, respectively. Cycles C_2 and C_4 were significantly different from the rest (least significant difference at the 0.05 probability level). Results from two other locations in Mexico are being processed, and the complete results of this multi-parental maize rapid selection cycle are yet to be published.

In addition, Beyene et al. (2015) previously reported significant genetic gains in maize grain yield through GS in eight CIMMYT tropical biparental maize populations in sub-Saharan Africa under drought conditions. They revealed that the average gain from GS per cycle across the eight populations was 0.086 t ha^{-1}, while the C_3-derived hybrids produced significantly higher average grain yields than C_0-derived hybrids. However, the average gain per cycle using marker-assisted recurrent selection across 10 populations was only 0.045 t ha^{-1} per cycle under similar environmental conditions.

Showcase 2: Two-part GS-assisted breeding at Lantmännen Lantbruk, Sweden

The breeding-cycle duration is arguably the single factor that has the largest effect on gain per time (Cobb et al., 2019). The genetic gain per unit time is of fundamental importance, particularly for breeding programs to maintain their competitive advantage, and is also crucial for attempting to adapt new cultivars to a rapidly changing environment (Budhlakoti et al., 2022). In a conventional breeding program of an inbred crop, such as wheat, barley, or oats, new parents are typically selected during the advanced yield trial stage, which results in a breeding cycle of around 5−8 years.

In Lantmännen, the GS-assisted breeding program of inbred crops is split into two parts: the first part is the GS-enabled recurrent selection, also called "population improvement"; and the second part is inbred line development, also called "product development," in which selected lines undergo testing in advanced field trials. This strategy significantly reduces the breeding-cycle time by selecting new parents at an early stage based on their genomic estimated breeding values. Simulation research supports this two-part strategy, outperforming both the conventional selection as well as "standard" GS (i.e., GS only applied at the preliminary yield trial stage) by significantly increasing genetic gain per unit time (Gaynor et al., 2017). Meanwhile, the two-part GS-assisted breeding strategy brings challenging issues for breeding programs. First, genotyping a large number (up 100 000) of early-generation individuals for high-density SNP markers could be expensive, particularly for small breeding programs. Second, a closed-loop two-part strategy, where no new allelic variation is introduced, leads to loss of both genetic diversity and prediction accuracy over time, with a negative impact on long-term genetic gain (Gaynor et al., 2017).

However, for self-pollinating crops where LD normally extends over longer genomic distances, rapid-cycling GS using a low-density marker set can deliver gains of similar magnitudes as high-density genotyping, even without marker imputation (A. Ceplitis, Lantmännen Lantbruk, Svalöv, Sweden, unpublished data). In addition, the negative effect on prediction accuracy that results from repeated rounds of recombination and the concomitant divergence of LD structure between the training and breeding populations can be alleviated by introducing inbred lines from the product development part as crossing parents in the population improvement part. Such a modified two-part strategy can maintain long-term genetic gain while simultaneously increasing prediction accuracy over time (A. Ceplitis, Lantmännen Lantbruk, Svalöv, Sweden, unpublished data).

The two-part breeding strategy was specifically developed for inbred line crops. Extending the strategy to outcrossing population crops, such as many forage species, which are characterized by significant inbreeding depression and rapid LD decay, is an area of active research. For these types of crops, preliminary results from simulation research indicate that a two-part GS strategy is superior over conventional phenotypic selection and other alternative GS scenarios in terms of accumulated genetic gain, particularly when prediction models include dominance effects (A. Ceplitis, Lantmännen Lantbruk, Svalöv, Sweden, unpublished data).

OUTLOOK

In this review, we delved into the fundamental building blocks of GP methodology and traced its evolution over more than two decades, illustrating its transformative impact on plant breeding. We elucidated how this methodology plays a pivotal role across various breeding stages, aiding in the selection of superior candidate individuals for further crossing, all while minimizing or even eliminating the need for extensive phenotyping over many consecutive breeding generations. This comprehensive review underscores the transformative impact of GP on the enhancement of crop genetic improvement, particularly in revolutionizing cross-breeding. The utilization of high-throughput genomic technology enables a thorough analysis of the entire crop genome, facilitating the identification of promising breeding germplasm associated with desirable traits for subsequent selection. By leveraging extensive sets of genomic and phenotypic data, GS methods predict breeding values for specific traits, thus circumventing the need for laborious and resource-intensive field trials. This streamlined approach speeds up the breeding process, thereby facilitating the selection of superior germplasm with coveted attributes such as increased yield, resistance to pathogens and pests, and adaptability to the dynamic environmental changes, which are often exacerbated by ongoing global warming.

A pivotal strength of GP lies in its capacity to unravel the intricate genetic architecture of traits. In contrast to cross-breeding methods heavily reliant on phenotypic observations influenced by both genetic and environmental factors, GP delves directly into the genetic makeup of plants, offering a more precise and reliable evaluation of their potential performance. This not only simplifies the identification of favorable alleles but also enables plant breeders to consider gene interactions and environmental influences in the target trait(s), thus resulting in the development of more robust and resilient crop germplasm. From this improved germplasm pool, the selection and further release of desired cultivars become more targeted and effective. As GP of breeding values progresses, its integration with machine learning and artificial intelligence emerges as a promising frontier in crop genetic improvement. The synergy of extensive genomic data and advanced computational models allows for the discernment of subtle genetic patterns and interactions previously overlooked. This holistic approach opens avenues for enhancing crop productivity, sustainability, and resilience in the face of challenges such as climate change and global food and nutrition security. Ultimately, GP of breeding values stands as a cutting-edge approach empowering plant breeders to make informed decisions, thus promising a new wave of innovation in agriculture.

This review explored the impact of various factors on the accuracy of GP with empirical research on wheat, maize, and potato as examples of different reproduction systems. We emphasized that GP, as a predictive tool, relies on the assurance of consistently high or, at the very least, commendable prediction accuracy. Nevertheless, it is worth noting that achieving such precision is not always feasible, owing to the myriad factors that influence its efficacy. We elucidated these factors and offered insights into how they can be optimized to enhance the practical application of GP methodology. Moreover, we expound upon how GS can harness the integration of omics and environmental data to further enhance its accuracy, broadening its scope and applicability. In conclusion, our review underscores that GS can significantly elevate genetic gains per unit of time within crop-breeding programs, but to increase its efficiency it is of paramount importance to integrate all factors that affect GP methodology to fully harness the potential of this groundbreaking predictive data-driven approach.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Molecular Plant Online*.

FUNDING

This research was supported by SLU Grogrund (#SLU-LTV.2020.1.1.1-654) and an Einar and Inga Nilsson Foundation grant. J.I.y.S. was supported by grant PID2021-123718OB-I00 funded by MCIN/AEI/10.13039/501 100 011 033 and by "ERDF A way of making Europe," CEX2020-000999-S. R.R.V. was supported by Novo Nordisk Fonden (0074727) and SLU's Centre for Biological Control. In addition, J.I.y.S. and J.F.-G. were supported by the Beatriz Galindo Program BEAGAL 18/00115.

AUTHOR CONTRIBUTIONS

A.A. and A. Chawade conceived the study. A.A., J.Å., O.A.M.-L., J.I.y.S., J.F.-G., W.T., J.C., A. Ceplitis, and R.O. contributed to writing the original manuscript draft; A.A., J.I.y.S., and J.F.-G. created figures; A. Chawade, A.S.C., R.R.V., A. Ceplitis, and R.O. reviewed and edited the manuscript. All authors approved the final manuscript.

DECLARATION OF INTERESTS

A. Ceplitis and J.Å. are employed by Lantmännen Lantbruk.

Received: November 3, 2023
Revised: January 22, 2024
Accepted: March 8, 2024

SUPPORTING CITATIONS

Adams et al. (2023), Adeyemo et al. (2020), Alemu et al. (2021a), 2021b, Ali et al. (2020), Allier et al. (2020), Belamkar et al. (2018), Ben-Sadoun et al.

(2020), Brauner et al. (2018), Bustos-Korts et al. (2016), Byrne et al. (2020), Caruana et al. (2019), Crossa et al. (2013), Cuevas et al. (2022), Cullis et al. (2006), Cullis et al. (2020), Daetwyler et al. (2014), de Bem Oliveira et al. (2020), de Oliveira et al. (2020), de Verdal et al. (2023), Edmondson (2020), Enciso-Rodriguez et al. (2018), Endelman et al. (2018), Ertiro et al. (2020), Fradgley et al. (2023), García-Barrios et al. (2023), Gill et al. (2021), Gowda et al. (2015), Habyarimana et al. (2017), Han et al. (2018), Hao et al. (2019), Heffner et al. (2010), Holland et al. (2020), Jarquin et al. (2021), Juliana et al. (2019b), Juliana et al. (2020), Kadam et al. (2021), Karaman et al. (2016), Lado et al. (2013), Lyra et al. (2018), Mangin et al. (2019), Mendonça and Fritsche-Neto (2020), Mitchell (2000), Momen and Morota (2018), Olatoye et al. (2020), Ornella et al. (2012), Ortiz et al. (2022), Ortiz et al. (2023a), Ortiz et al. (2023b), Owens et al. (2014), Pace et al. (2015), Pandey et al. (2023), Rincent et al. (2017), Rio et al. (2021), Rio et al. (2022), Rogers et al. (2022), Roth et al. (2020), Saint Pierre et al. (2016), Sehgal et al. (2020), Selga et al. (2020), Selga et al. (2021), Semagn et al. (2022a), 2022b, Shahi et al. (2022), Shahinnia et al. (2022), Sitonik et al. (2019), Sirsat et al. (2022), Sood et al. (2020), Sood et al. (2023), Stich and Van Inghelandt (2018), Sukumaran et al. (2017), Sun et al. (2017), Sverrisdóttir et al. (2017), Tayeh et al. (2015), Technow et al. (2013), Tomar et al. (2021a), 2021b, Tsai et al. (2020), Vélez-Torres et al. (2018), Wang et al. (2020), Wilson et al. (2021), Yu et al. (2020), Yu et al. (2023), Yuan et al. (2019), Zakieh et al. (2023), Zhao et al. (2013).

REFERENCES

Adams, J., De Vries, M., and Van Eeuwijk, F. (2023). Efficient Genomic Prediction of Yield and Dry Matter in Hybrid Potato. Plants **12**:2617. https://doi.org/10.3390/plants12142617.

Adeyemo, E., Bajgain, P., Conley, E., Sallam, A., and Anderson, J. (2020). Optimizing training population size and content to improve prediction accuracy of FHB-related traits in wheat. Agronomy **10**:543. https://doi.org/10.3390/agronomy10040543.

Akdemir, D. (2017). STPGA: Selection of Training Populations by Genetic Algorithm. Preprint at. https://doi.org/10.1101/111989. https://cran.r-project.org/web/packages/STPGA/STPGA.pdf.

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. Sci. Rep. **9**, 1446. https://doi.org/10.1038/s41598-018-38081-6.

Akdemir, D., Sanchez, J.I., and Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. Genet. Sel. Evol. **47**, 38. https://doi.org/10.1186/s12711-015-0116-6.

Akdemir, D., Rio, S., and Isidro Y Sánchez, J. (2021). TrainSel: An R Package for Selection of Training Populations. Front. Genet. **12**, 655287. https://doi.org/10.3389/fgene.2021.655287.

Al Kalaldeh, M., Gibson, J., Duijvesteijn, N., Daetwyler, H.D., Macleod, I., Moghaddar, N., Lee, S.H., and Van Der Werf, J.H.J. (2019). Using imputed whole-genome sequence data to improve the accuracy of genomic prediction for parasite resistance in Australian sheep. Genet. Sel. Evol. **51**, 32. https://doi.org/10.1186/s12711-019-0476-4.

Albrecht, T., Wimmer, V., Auinger, H.-J., Erbe, M., Knaak, C., Ouzunova, M., Simianer, H., and Schön, C.C. (2011). Genome-based prediction of testcross values in maize. Theor. Appl. Genet. **123**:339–350. https://doi.org/10.1007/s00122-011-1587-7.

Albrecht, T., Auinger, H.-J., Wimmer, V., Ogutu, J.O., Knaak, C., Ouzunova, M., Piepho, H.-P., and Schön, C.C. (2014). Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor. Appl. Genet. **127**:1375–1386. https://doi.org/10.1007/s00122-014-2305-z.

Alemu, A., Batista, L., Singh, P.K., Ceplitis, A., and Chawade, A. (2023). Haplotype-tagged SNPs improve genomic prediction accuracy for Fusarium head blight resistance and yield-related traits in wheat. Theor. Appl. Genet. **136**, 92. https://doi.org/10.1007/s00122-023-04352-8.

Alemu, A., Suliman, S., Hagras, A., Thabet, S., Al-Abdallat, A., Abdelmula, A.A., and Tadesse, W. (2021a). Multi-model genome-wide association and genomic prediction analysis of 16 agronomic, physiological and quality related traits in ICARDA spring wheat. Euphytica **217**, 205. https://doi.org/10.1007/s10681-021-02933-6.

Alemu, A., Brazauskas, G., Gaikpa, D.S., Henriksson, T., Islamov, B., Jørgensen, L.N., Koppel, M., Koppel, R., Liatukas, Ž., Svensson, J.T., and Chawade, A. (2021b). Genome-Wide Association Analysis and Genomic Prediction for Adult-Plant Resistance to Septoria Tritici Blotch and Powdery Mildew in Winter Wheat. Front. Genet. **12**, 661742. https://doi.org/10.3389/fgene.2021.661742.

Ali, M., Zhang, Y., Rasheed, A., Wang, J., and Zhang, L. (2020). Genomic Prediction for Grain Yield and Yield-Related Traits in Chinese Winter Wheat. Int. J. Mol. Sci. **21**:1342. https://doi.org/10.3390/ijms21041342.

Allier, A., Teyssèdre, S., Lehermeier, C., Charcosset, A., and Moreau, L. (2020). Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs. Theor. Appl. Genet. **133**:201–215. https://doi.org/10.1007/s00122-019-03451-9.

Anilkumar, C., Muhammed Azharudheen, T.P., Sah, R.P., Sunitha, N.C., Devanna, B.N., Marndi, B.C., and Patra, B.C. (2023). Gene based markers improve precision of genome-wide association studies and accuracy of genomic predictions in rice breeding. Heredity **130**:335–345. https://doi.org/10.1038/s41437-023-00599-5.

Araus, J.L., and Cairns, J.E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci. **19**:52–61. https://doi.org/10.1016/j.tplants.2013.09.008.

Araus, J.L., Kefauver, S.C., Zaman-Allah, M., Olsen, M.S., and Cairns, J.E. (2018). Translating High-Throughput Phenotyping into Genetic Gain. Trends Plant Sci. **23**:451–466. https://doi.org/10.1016/j.tplants.2018.02.001.

Arojju, S.K., Conaghan, P., Barth, S., Milbourne, D., Casler, M.D., Hodkinson, T.R., Michel, T., and Byrne, S.L. (2018). Genomic prediction of crown rust resistance in Lolium perenne. BMC Genet. **19**, 35. https://doi.org/10.1186/s12863-018-0613-z.

Arruda, M.P., Brown, P.J., Lipka, A.E., Krill, A.M., Thurber, C., and Kolb, F.L. (2015). Genomic Selection for Predicting Fusarium Head Blight Resistance in a Wheat Breeding Program. Plant Genome **8**, eplantgenome2015.01.0003. https://doi.org/10.3835/plantgenome2015.01.0003.

Atanda, S.A., Govindan, V., Singh, R., Robbins, K.R., Crossa, J., and Bentley, A.R. (2022). Sparse testing using genomic prediction improves selection for breeding targets in elite spring wheat. Theor. Appl. Genet. **135**:1939–1950. https://doi.org/10.1007/s00122-022-04085-0.

Atanda, S.A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., Gowda, M., Dreher, K., Zhang, X., Prasanna, B.M., et al. (2021a). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. Theor. Appl. Genet. **134**:279–294. https://doi.org/10.1007/s00122-020-03696-9.

Atanda, S.A., Olsen, M., Crossa, J., Burgueño, J., Rincent, R., Dzidzienyo, D., Beyene, Y., Gowda, M., Dreher, K., Boddupalli, P.M., et al. (2021b). Scalable Sparse Testing Genomic Selection Strategy for Early Yield Testing Stage. Front. Plant Sci. **12**, 658978. https://doi.org/10.3389/fpls.2021.658978.

Azevedo, C.F., Resende, M.D.V.D., Silva, F.F.E., Lopes, P.S., and Guimarães, S.E.F. (2013). Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. Pesq. agropec. bras. **48**:619–626. https://doi.org/10.1590/s0100-204x2013000600007.

Azodi, C.B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., and Shiu, S.-H. (2019). Benchmarking Parametric and Machine

Learning Models for Genomic Prediction of Complex Traits. G3 (Bethesda). **9**:3691–3702. https://doi.org/10.1534/g3.119.400498.

Ballén-Taborda, C., Lyerly, J., Smith, J., Howell, K., Brown-Guedira, G., Babar, M.A., Harrison, S.A., Mason, R.E., Mergoum, M., Murphy, J.P., et al. (2022). Utilizing genomics and historical data to optimize gene pools for new breeding programs: A case study in winter wheat. Front. Genet. **13**, 964684. https://doi.org/10.3389/fgene.2022.964684.

Ballesta, P., Bush, D., Silva, F.F., and Mora, F. (2020). Genomic Predictions Using Low-Density SNP Markers, Pedigree and GWAS Information: A Case Study with the Non-Model Species Eucalyptus cladocalyx. Plants **9**:99. https://doi.org/10.3390/plants9010099.

Basnet, B.R., Crossa, J., Dreisigacker, S., Pérez-Rodríguez, P., Manes, Y., Singh, R.P., Rosyara, U.R., Camarillo-Castillo, F., and Murua, M. (2019). Hybrid Wheat Prediction Using Genomic, Pedigree, and Environmental Covariables Interaction Models. Plant Genome **12**, 180051. https://doi.org/10.3835/plantgenome2018.07.0051.

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

Belamkar, V., Guttieri, M.J., Hussain, W., Jarquín, D., El-basyoni, I., Poland, J., Lorenz, A.J., and Baenziger, P.S. (2018). Genomic Selection in Preliminary Yield Trials in a Winter Wheat Breeding Program. G3 (Bethesda). **8**:2735–2747. https://doi.org/10.1534/g3.118.200415.

Ben-Sadoun, S., Rincent, R., Auzanneau, J., Oury, F.X., Rolland, B., Heumez, E., Ravel, C., Charmet, G., and Bouchet, S. (2020). Economical optimization of a breeding scheme by selective phenotyping of the calibration set in a multi-trait context: application to bread making quality. Theor. Appl. Genet. **133**:2197–2212. https://doi.org/10.1007/s00122-020-03590-4.

Bentley, A.R., Scutari, M., Gosman, N., Faure, S., Bedford, F., Howell, P., Cockram, J., Rose, G.A., Barber, T., Irigoyen, J., et al. (2014). Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. Theor. Appl. Genet. **127**:2619–2633. https://doi.org/10.1007/s00122-014-2403-y.

Bermingham, M.L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A.F., Wilson, J.F., Agakov, F., Navarro, P., and Haley, C.S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci. Rep. **5**, 10312. https://doi.org/10.1038/srep10312.

Bernardo, R. (1994). Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Sci. **34**:20–25. https://doi.org/10.2135/cropsci1994.0011183X003400010003x.

Bernardo, R. (2008). Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. Crop Sci. **48**:1649–1664. https://doi.org/10.2135/cropsci2008.03.0131.

Bernardo, R. (2016). Bandwagons I, too, have known. Theor. Appl. Genet. **129**:2323–2332. https://doi.org/10.1007/s00122-016-2772-5.

Bernardo, R., and Yu, J. (2007). Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci. **47**:1082–1090. https://doi.org/10.2135/cropsci2006.11.0690.

Berro, I., Lado, B., Nalin, R.S., Quincke, M., and Gutiérrez, L. (2019). Training Population Optimization for Genomic Selection. Plant Genome **12**:1–14. https://doi.org/10.3835/plantgenome2019.04.0028.

Beyene, Y., Gowda, M., Olsen, M., Robbins, K.R., Pérez-Rodríguez, P., Alvarado, G., Dreher, K., Gao, S.Y., Mugo, S., Prasanna, B.M., and Crossa, J. (2019). Empirical Comparison of Tropical Maize Hybrids Selected Through Genomic and Phenotypic Selections. Front. Plant Sci. **10**, 1502. https://doi.org/10.3389/fpls.2019.01502.

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., Sehabiague, P., Makumbi, D., Magorokosho, C., Oikeh, S., et al. (2015). Genetic Gains in Grain Yield Through Genomic Selection in Eight Bi-parental Maize Populations under Drought Stress. Crop Sci. **55**:154–163. https://doi.org/10.2135/cropsci2014.07.0460.

Biswas, P.S., Ahmed, M.M.E., Afrin, W., Rahman, A., Shalahuddin, A.K.M., Islam, R., Akter, F., Syed, M.A., Sarker, M.R.A., Ifterkharuddaula, K.M., and Islam, M.R. (2023). Enhancing genetic gain through the application of genomic selection in developing irrigated rice for the favorable ecosystem in Bangladesh. Front. Genet. **14**, 1083221. https://doi.org/10.3389/fgene.2023.1083221.

Blondel, M., Onogi, A., Iwata, H., and Ueda, N. (2015). A Ranking Approach to Genomic Selection. PLoS One **10**, e0128570. https://doi.org/10.1371/journal.pone.0128570.

Bohra, A., Kilian, B., Sivasankar, S., Caccamo, M., Mba, C., McCouch, S.R., and Varshney, R.K. (2022). Reap the crop wild relatives for breeding future crops. Trends Biotechnol. **40**:412–431. https://doi.org/10.1016/j.tibtech.2021.08.009.

Bolormaa, S., Gore, K., Van Der Werf, J.H.J., Hayes, B.J., and Daetwyler, H.D. (2015). Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Anim. Genet. **46**:544–556. https://doi.org/10.1111/age.12340.

Bonnett, D., Li, Y., Crossa, J., Dreisigacker, S., Basnet, B., Pérez-Rodríguez, P., Alvarado, G., Jannink, J.L., Poland, J., and Sorrells, M. (2022). Response to Early Generation Genomic Selection for Yield in Wheat. Front. Plant Sci. **12**, 718611. https://doi.org/10.3389/fpls.2021.718611.

Borlaug, N.E. (2002). Feeding a world of 10 billion people: The miracle ahead. In Vitro Cell Dev. Biol. Plant **38**:221–228. https://doi.org/10.1079/ivp2001279.

Brauner, P.C., Müller, D., Schopp, P., Böhm, J., Bauer, E., Schön, C.C., and Melchinger, A.E. (2018). Genomic Prediction Within and Among Doubled-Haploid Libraries from Maize Landraces. Genetics **210**:1185–1196. https://doi.org/10.1534/genetics.118.301286.

Burgueño, J., de los Campos, G., Weigel, K., and 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.

Budhlakoti, N., Kushwaha, A.K., Rai, A., Chaturvedi, K.K., Kumar, A., Pradhan, A.K., Kumar, U., Kumar, R.R., Juliana, P., Mishra, D.C., et al. (2022). Genomic Selection: A Tool for Accelerating the Efficiency of Molecular Breeding for Development of Climate-Resilient Crops. Front. Genet. **13**, 832153. https://doi.org/10.3389/fgene.2022.832153.

Bustos-Korts, D., Malosetti, M., Chapman, S., Biddulph, B., and van Eeuwijk, F. (2016). Improvement of predictive ability by uniform coverage of the target genetic space. G3 (Bethesda). **6**:3733–3747. https://doi.org/10.1534/g3.116.035410.

Butler, D., Smith, A., and Cullis, B. (2013). On Model Based Design of Comparative Experiments (National Institute for Applied Statistics Research Australia, University of Wollongong). https://documents.uow.edu.au/content/groups/public/@web/@inf/@math/documents/doc/uow273977.pdf.

Byrne, S., Meade, F., Mesiti, F., Griffin, D., Kennedy, C., and Milbourne, D. (2020). Genome-Wide Association and Genomic Prediction for Fry Color in Potato. Agronomy **10**:90. https://doi.org/10.3390/agronomy10010090.

Cabrera-Bosquet, L., Crossa, J., Von Zitzewitz, J., Serret, M.D., and Araus, J.L. (2012). High-throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding ConvergeF. J. Integr. Plant Biol. **54**:312–320. https://doi.org/10.1111/j.1744-7909.2012.01116.x.

Callister, A.N., Bermann, M., Elms, S., Bradshaw, B.P., Lourenco, D., and Brawner, J.T. (2022). Accounting for population structure in genomic predictions of Eucalyptus globulus. G3 (Bethesda). **12**, jkac180. https://doi.org/10.1093/g3journal/jkac180.

Cantelmo, N.F., Von Pinho, R.G., and Balestre, M. (2017). Genome-wide prediction for maize single-cross hybrids using the G-BLUP model and validation in different crop seasons. Mol. Breed. **37**, 51. https://doi.org/10.1007/s11032-017-0651-7.

Caruana, B.M., Pembleton, L.W., Constable, F., Rodoni, B., Slater, A.T., and Cogan, N.O.I. (2019). Validation of Genotyping by Sequencing Using Transcriptomics for Diversity and Application of Genomic Selection in Tetraploid Potato. Front. Plant Sci. **10**, 670. https://doi.org/10.3389/fpls.2019.00670.

Chen, T., and He, T. (2014). Higgs boson discovery with boosted trees. In Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning-Volume 42, pp. 69–80.

Chen, X., and Ishwaran, H. (2012). Random forests for genomic data analysis. Genomics **99**:323–329. https://doi.org/10.1016/j.ygeno.2012.04.003.

Chen, Z.-Q., Klingberg, A., Hallingbäck, H.R., and Wu, H.X. (2023). Preselection of QTL markers enhances accuracy of genomic selection in Norway spruce. BMC Genom. **24**, 147. https://doi.org/10.1186/s12864-023-09250-3.

Cheng, H., Garrick, D.J., and Fernando, R.L. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. J. Anim. Sci. Biotechnol. **8**, 38. https://doi.org/10.1186/s40104-017-0164-6.

Chung, P.-Y., and Liao, C.-T. (2022). Selection of parental lines for plant breeding via genomic prediction. Front. Plant Sci. **13**. https://doi.org/10.3389/fpls.2022.934767.

Clark, S.A., Hickey, J.M., Daetwyler, H.D., and van der Werf, J.H.J. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. **44**:4. https://doi.org/10.1186/1297-9686-44-4.

Cobb, M. (2017). 60 years ago, Francis Crick changed the logic of biology. PLoS Biol. **15**, e2003243. https://doi.org/10.1371/journal.pbio.2003243.

Cobb, J.N., Juma, R.U., Biswas, P.S., et al. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. Theor. Appl. Genet. **132**:627–645. https://doi.org/10.1007/s00122-019-03317-0.

Combs, E., and Bernardo, R. (2013). Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. Plant Genome **6**. https://doi.org/10.3835/plantgenome2012.11.0030.

Crain, J., Mondal, S., Rutkoski, J., Singh, R.P., and Poland, J. (2018). Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. Plant Genome **11**, 170043. https://doi.org/10.3835/plantgenome2017.05.0043.

Crespo-Herrera, L., Howard, R., Piepho, H.-P., Pérez-Rodríguez, P., Montesinos-Lopez, O., Burgueño, J., Singh, R., Mondal, S., Jarquín, D., and Crossa, J. (2021). Genome-enabled prediction for sparse testing in multi-environmental wheat trials. Plant Genome **14**, e20151. https://doi.org/10.1002/tpg2.20151.

Crossa, J., de los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., and Pérez-Rodríguez, P. (2016a). 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.

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., et al. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity **112**:48–60. https://doi.org/10.1038/hdy.2013.16.

Crossa, J., Campos, G.d.l., Pérez, P., Gianola, D., Burgueño, J., Araus, J.L., Makumbi, D., Singh, R.P., Dreisigacker, S., Yan, J., et al. (2010). Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics **186**:713–724. https://doi.org/10.1534/genetics.110.118521.

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J.M., Chen, C., De Los Campos, G., Burgueño, J., Windhausen, V.S., Buckler, E., et al. (2013). Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing. G3 (Bethesda). **3**:1903–1926. https://doi.org/10.1534/g3.113.008227.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci. **22**:961–975. https://doi.org/10.1016/j.tplants.2017.08.011.

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., Vikram, P., Sansaloni, C., Petroli, C., Akdemir, D., et al. (2016b). Genomic Prediction of Gene Bank Wheat Landraces. G3 (Bethesda). **6**:1819–1834. https://doi.org/10.1534/g3.116.029637.

Cuevas, J., Reslow, F., Crossa, J., and Ortiz, R. (2022). Modeling genotype × environment interaction for single and multitrait genomic prediction in potato (Solanum tuberosum L.). G3 (Bethesda). **13**, jkac322. https://doi.org/10.1093/g3journal/jkac322.

Cui, Y., Li, R., Li, G., Zhang, F., Zhu, T., Zhang, Q., Ali, J., Li, Z., and Xu, S. (2020). Hybrid breeding of rice via genomic selection. Plant Biotechnol. J. **18**:57–67. https://doi.org/10.1111/pbi.13170.

Cullis, B.R., Smith, A.B., and Coombes, N.E. (2006). On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. **11**:381–393. https://doi.org/10.1198/108571106X154443.

Cullis, B.R., Smith, A.B., Cocks, N.A., and Butler, D.G. (2020). The Design of Early-Stage Plant Breeding Trials Using Genetic Relatedness. J. Agric. Biol. Environ. Stat. **25**:553–578. https://doi.org/10.1007/s13253-020-00403-5.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., and Woolliams, J.A. (2010). The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics **185**:1021–1031. https://doi.org/10.1534/genetics.110.116855.

Daetwyler, H.D., Kemper, K.E., van der Werf, J.H.J., and Hayes, B.J. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population1. J. Anim. Sci. **90**:3375–3384. https://doi.org/10.2527/jas.2011-4557.

Daetwyler, H.D., Bansal, U.K., Bariana, H.S., Hayden, M.J., and Hayes, B.J. (2014). Genomic prediction for rust resistance in diverse wheat landraces. Theor. Appl. Genet. **127**:1795–1803. https://doi.org/10.1007/s00122-014-2341-8.

de Bem Oliveira, I., Amadeu, R.R., Ferrão, L.F.V., and Muñoz, P.R. (2020). Optimizing whole-genomic prediction for autotetraploid blueberry breeding. Heredity **125**:437–448. https://doi.org/10.1038/s41437-020-00357-x.

de los Campos, G., Gianola, D., and Rosa, G.J.M. (2009). Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation1. J. Anim. Sci. **87**:1883–1887. https://doi.org/10.2527/jas.2008-1259.

de Los Campos, G., Gianola, D., Rosa, G.J.M., Weigel, K.A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. **92**:295–308. https://doi.org/10.1017/s0016672310000285.

de Los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P.L. (2013). Whole-Genome Regression and

Prediction Methods Applied to Plant and Animal Breeding. Genetics **193**:327–345. https://doi.org/10.1534/genetics.112.143313.

de los Campos, G., Veturi, Y., Vazquez, A.I., Lehermeier, C., and Pérez-Rodríguez, P. (2015). Incorporating Genetic Heterogeneity in Whole-Genome Regressions Using Interactions. J. Agric. Biol. Environ. Stat. **20**:467–490. https://doi.org/10.1007/s13253-015-0222-5.

de Oliveira, A.A., Resende, M.F.R., Ferrão, L.F.V., Amadeu, R.R., Guimarães, L.J.M., Guimarães, C.T., Pastina, M.M., and Margarido, G.R.A. (2020). Genomic prediction applied to multiple traits and environments in second season maize hybrids. Heredity **125**:60–72. https://doi.org/10.1038/s41437-020-0321-0.

de Verdal, H., Baertschi, C., Frouin, J., Quintero, C., Ospina, Y., Alvarez, M.F., Cao, T.-V., Bartholomé, J., and Grenier, C. (2023). Optimization of Multi-Generation Multi-location Genomic Prediction Models for Recurrent Genomic Selection in an Upland Rice Population. Rice **16**:43. https://doi.org/10.18167/DVN1/A7DEHI.

Desta, Z.A., and 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.

Dreisigacker, S., Pérez-Rodríguez, P., Crespo-Herrera, L., Bentley, A., and Crossa, J. (2023). Results From Rapid-Cycle Recurrent Genomic Selection in Spring Bread Wheat. G3 Genes|Genomes|Genetics **13**, jkad025. https://doi.org/10.1093/g3journal/jkad025.

Dzievit, M.J., Guo, T., Li, X., and Yu, J. (2021). Comprehensive analytical and empirical evaluation of genomic prediction across diverse accessions in maize. Plant Genome **14**:e20160. https://doi.org/10.1002/tpg2.20160.

Edmondson, R.N. (2020). Multi-level Block Designs for Comparative Experiments. J. Agric. Biol. Environ. Stat. **25**:500–522. https://doi.org/10.1007/s13253-020-00416-0.

Edriss, V., Gao, Y., Zhang, X., Jumbo, M.B., Makumbi, D., Olsen, M.S., Crossa, J., Packard, K.C., and Jannink, J.L. (2017). Genomic Prediction in a Large African Maize Population. Crop Sci. **57**:2361–2371. https://doi.org/10.2135/cropsci2016.08.0715.

Edwards, S.M., Buntjer, J.B., Jackson, R., Bentley, A.R., Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman, E., et al. (2019). The effects of training population design on genomic prediction accuracy in wheat. Theor. Appl. Genet. **132**:1943–1952. https://doi.org/10.1007/s00122-019-03327-y.

El Hanafi, S., Jiang, Y., Kehel, Z., Schulthess, A.W., Zhao, Y., Mascher, M., Haupt, M., Himmelbach, A., Stein, N., Amri, A., and Reif, J.C. (2023). Genomic predictions to leverage phenotypic data across genebanks. Front. Plant Sci. **14**, 1227656. https://doi.org/10.3389/fpls.2023.1227656.

Enciso-Rodriguez, F., Douches, D., Lopez-Cruz, M., Coombs, J., and De Los Campos, G. (2018). Genomic Selection for Late Blight and Common Scab Resistance in Tetraploid Potato (*Solanum tuberosum*). G3 (Bethesda). **8**:2471–2481. https://doi.org/10.1534/g3.118.200273.

Endelman, J.B., Carley, C.A.S., Bethke, P.C., Coombs, J.J., Clough, M.E., da Silva, W.L., De Jong, W.S., Douches, D.S., Frederick, C.M., Haynes, K.G., et al. (2018). Genetic Variance Partitioning and Genome-Wide Prediction with Allele Dosage Information in Autotetraploid Potato. Genetics **209**:77–87. https://doi.org/10.1534/genetics.118.300685.

Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A., and Goddard, M.E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. **95**:4114–4129. https://doi.org/10.3168/jds.2011-5019.

Ertiro, B.T., Labuschagne, M., Olsen, M., Das, B., Prasanna, B.M., and Gowda, M. (2020). Genetic Dissection of Nitrogen Use Efficiency in Tropical Maize Through Genome-Wide Association and Genomic Prediction. Front. Plant Sci. **11**, 474. https://doi.org/10.3389/fpls.2020.00474.

Esfandyari, H., Sørensen, A.C., and Bijma, P. (2015). A crossbred reference population can improve the response to genomic selection for crossbred performance. Genet. Sel. Evol. **47**, 76. https://doi.org/10.1186/s12711-015-0155-z.

Falconer, D., and Mackay, T. (1996). Introduction to Quantitative Genetics, 4 (Benjamin Cummings).

FAO. (2023). The State of Food Security and Nutrition in the World 2023. Urbanization, Agrifood Systems Transformation and Healthy Diets across the Rural–Urban Continuum (FAOSTAT).

Fernández-González, J., Akdemir, D., and Isidro y Sánchez, J. (2023). A comparison of methods for training population optimization in genomic selection. Theor. Appl. Genet. **136**:30. https://doi.org/10.1007/s00122-023-04265-6.

Fernández-González, Haquin, B., Combes, E., Bernard, K., Allard, A., and Isidro y Sánchez, J. (2024). Maximizing efficiency in sunflower breeding through historical data optimization. Plant Methods **20**, 42. https://doi.org/10.1186/s13007-024-01151-0.

Filho, D.F., Filho, J.S.D.S.B., Regitano, L.C.D.A., Alencar, M.M.D., Alves, R.R., and Meirelles, S.L.C. (2019). Tournaments between markers as a strategy to enhance genomic predictions. PLoS One **14**, e0217283. https://doi.org/10.1371/journal.pone.0217283.

Flint-Garcia, S.A., Thornsberry, J.M., and Buckler, E.S. (2003). Structure of Linkage Disequilibrium in Plants. Annu. Rev. Plant Biol. **54**:357–374. https://doi.org/10.1146/annurev.arplant.54.031902.134907.

Fradgley, N., Gardner, K.A., Bentley, A.R., Howell, P., Mackay, I.J., Scott, M.F., Mott, R., and Cockram, J. (2023). Multi-trait ensemble genomic prediction and simulations of recurrent selection highlight importance of complex trait genetic architecture for long-term genetic gains in wheat. In Silico Plants **5**. https://doi.org/10.1093/insilicoplants/diad002.

Galán, R.J., Bernal-Vasquez, A.-M., Jebsen, C., Piepho, H.-P., Thorwarth, P., Steffan, P., Gordillo, A., and Miedaner, T. (2020). Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye. Theor. Appl. Genet. **133**:3001–3015. https://doi.org/10.1007/s00122-020-03651-8.

García-Barrios, G., Crossa, J., Cruz-Izquierdo, S., Aguilar-Rincón, V.H., Sandoval-Islas, J.S., Corona-Torres, T., Lozano-Ramírez, N., Dreisigacker, S., He, X., Singh, P.K., and Pacheco-Gil, R.A. (2023). Genomic Prediction of Resistance to Tan Spot, Spot Blotch and Septoria Nodorum Blotch in Synthetic Hexaploid Wheat. Int. J. Mol. Sci. **24**, 10506. https://doi.org/10.3390/ijms241310506.

Gaynor, R.C., Gorjanc, G., Bentley, A.R., Ober, E.S., Howell, P., Jackson, R., Mackay, I.J., and Hickey, J.M. (2017). A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines. Crop Sci. **57**:2372–2386. https://doi.org/10.2135/cropsci2016.09.0742.

George, E.I., and McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. J. Am. Stat. Assoc. **88**:881–889. https://doi.org/10.1080/01621459.1993.10476353.

Gianola, D., Fernando, R.L., and Stella, A. (2006). Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures. Genetics **173**:1761–1776. https://doi.org/10.1534/genetics.105.049510.

Gianola, D., Okut, H., Weigel, K.A., and Rosa, G.J. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. **12**:87. https://doi.org/10.1186/1471-2156-12-87.

Gill, H.S., Halder, J., Zhang, J., Brar, N.K., Rai, T.S., Hall, C., Bernardo, A., Amand, P.S., Bai, G., Olson, E., et al. (2021). Multi-Trait

Multi-Environment Genomic Prediction of Agronomic Traits in Advanced Breeding Lines of Winter Wheat. Front. Plant Sci. **12**, 709545. https://doi.org/10.3389/fpls.2021.709545.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica **136**:245–257. https://doi.org/10.1007/s10709-008-9308-0.

González-Camacho, J.M., Crossa, J., Pérez-Rodríguez, P., Ornella, L., and Gianola, D. (2016). Genome-enabled prediction using probabilistic neural network classifiers. BMC Genom. **17**, 208. https://doi.org/10.1186/s12864-016-2553-1.

Gowda, M., Das, B., Makumbi, D., Babu, R., Semagn, K., Mahuku, G., Olsen, M.S., Bright, J.M., Beyene, Y., and Prasanna, B.M. (2015). Genome-wide association and genomic prediction of resistance to maize lethal necrosis disease in tropical maize germplasm. Theor. Appl. Genet. **128**:1957–1968. https://doi.org/10.1007/s00122-015-2559-0.

Grinberg, N.F., Lovatt, A., Hegarty, M., Lovatt, A., Skøt, K.P., Kelly, R., Blackmore, T., Thorogood, D., King, R.D., Armstead, I., et al. (2016). Implementation of Genomic Prediction in Lolium perenne (L.) Breeding Populations. Front. Plant Sci. **7**, 133. https://doi.org/10.3389/fpls.2016.00133.

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., McMullen, M.D., Holland, J.B., Szalma, S.J., Wisser, R.J., and Yu, J. (2019). Optimal Designs for Genomic Selection in Hybrid Crops. Mol. Plant **12**:390–401. https://doi.org/10.1016/j.molp.2018.12.022.

Guo, T., Mu, Q., Wang, J., Vanous, A.E., Onogi, A., Iwata, H., Li, X., and Yu, J. (2020a). Dynamic effects of interacting genes underlying rice flowering-time phenotypic plasticity and global adaptation. Genome Res. **30**:673–683. https://doi.org/10.1101/gr.255703.119.

Guo, J., Khan, J., Pradhan, S., Shahi, D., Khan, N., Avci, M., Mcbreen, J., Harrison, S., Brown-Guedira, G., Murphy, J.P., et al. (2020b). Multi-Trait Genomic Prediction of Yield-Related Traits in US Soft Wheat under Variable Water Regimes. Genes **11**:1270. https://doi.org/10.3390/genes11111270.

Guo, R., Dhliwayo, T., Mageto, E.K., Palacios-Rojas, N., Lee, M., Yu, D., Ruan, Y., Zhang, A., San Vicente, F., Olsen, M., et al. (2020c). Genomic Prediction of Kernel Zinc Concentration in Multiple Maize Populations Using Genotyping-by-Sequencing and Repeat Amplification Sequencing Markers. Front. Plant Sci. **11**, 534. https://doi.org/10.3389/fpls.2020.00534.

Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., and 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.

Guo, Z., Tucker, D.M., Basten, C.J., Gandhi, H., Ersoz, E., Guo, B., Xu, Z., Wang, D., and Gay, G. (2014). The impact of population structure on genomic prediction in stratified populations. Theor. Appl. Genet. **127**:749–762. https://doi.org/10.1007/s00122-013-2255-x.

Guo, Z., Zou, C., Liu, X., Wang, S., Li, W.-X., Jeffers, D., Fan, X., Xu, M., and Xu, Y. (2020d). Complex Genetic System Involved in Fusarium Ear Rot Resistance in Maize as Revealed by GWAS, Bulked Sample Analysis, and Genomic Prediction. Plant Dis. **104**:1725–1735. https://doi.org/10.1094/pdis-07-19-1552-re.

Habier, D., Fernando, R.L., and Dekkers, J.C.M. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics **177**:2389–2397. https://doi.org/10.1534/genetics.107.081190.

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

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., and Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. **42**:5. https://doi.org/10.1186/1297-9686-42-5.

Habyarimana, E., Parisi, B., and Mandolino, G. (2017). Genomic prediction for yields, processing and nutritional quality traits in cultivated potato (*Solanum tuberosum* L.). Plant Breed. **136**:245–252. https://doi.org/10.1111/pbr.12461.

Haile, J.K., N'Diaye, A., Sari, E., Walkowiak, S., Rutkoski, J.E., Kutcher, H.R., and Pozniak, C.J. (2020). Potential of Genomic Selection and Integrating "Omics" Data for Disease Evaluation in Wheat. Crop Breeding, Genetics and Genomics **2**, e200016. https://doi.org/10.20900/cbgg20200016.

Haley, C.S., and Visscher, P.M. (1998). Strategies to Utilize Marker-Quantitative Trait Loci Associations. J. Dairy Sci. **81**:85–97. https://doi.org/10.3168/jds.S0022-0302(98)70157-2.

Han, S., Miedaner, T., Utz, H.F., Schipprack, W., Schrag, T.A., and Melchinger, A.E. (2018). Genomic prediction and GWAS of Gibberella ear rot resistance traits in dent and flint lines of a public maize breeding program. Euphytica **214**, 6. https://doi.org/10.1007/s10681-017-2090-2.

Hao, Y., Wang, H., Yang, X., Zhang, H., He, C., Li, D., Li, H., Wang, G., Wang, J., and Fu, J. (2019). Genomic Prediction using Existing Historical Data Contributing to Selection in Biparental Populations: A Study of Kernel Oil in Maize. Plant Genome **12**, 180025. https://doi.org/10.3835/plantgenome2018.05.0025.

Hayes, B., and Goddard, M.E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. **33**:209–229. https://doi.org/10.1186/1297-9686-33-3-209.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., and Goddard, M.E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. **92**:433–443. https://doi.org/10.3168/jds.2008-1646.

He, S., Schulthess, A.W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., Ebmeyer, E., Reif, J.C., and Jiang, Y. (2016). Genomic selection in a commercial winter wheat population. Theor. Appl. Genet. **129**:641–651. https://doi.org/10.1007/s00122-015-2655-1.

Heffner, E.L., Sorrells, M.E., and Jannink, J.L. (2009). Genomic Selection for Crop Improvement. Crop Sci. **49**:1–12. https://doi.org/10.2135/cropsci2008.08.0512.

Heffner, E.L., Lorenz, A.J., Jannink, J.-L., and Sorrells, M.E. (2010). Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Sci. **50**:1681–1690. https://doi.org/10.2135/cropsci2009.11.0662.

Helton, J.C., and Davis, F.J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliab. Eng. Syst. Saf. **81**:23–69. https://doi.org/10.1016/S0951-8320(03)00058-9.

Heslot, N., and Feoktistov, V. (2020). Optimization of Selective Phenotyping and Population Design for Genomic Prediction. J. Agric. Biol. Environ. Stat. **25**:579–600. https://doi.org/10.1007/s13253-020-00415-1.

Heslot, N., Akdemir, D., Sorrells, M.E., and 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.

Hickey, J.M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B.M., Grondona, M., Zambelli, A., Windhausen, V.S., Mathews, K., and Gorjanc, G. (2014). Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. Crop Sci. **54**:1476–1488. https://doi.org/10.2135/cropsci2013.03.0195.

Holland, J.B., Marino, T.P., Manching, H.C., and Wisser, R.J. (2020). Genomic prediction for resistance to Fusarium ear rot and fumonisin contamination in maize. Crop Sci. **60**:1863–1875. https://doi.org/10.1002/csc2.20163.

Hu, H., Campbell, M.T., Yeats, T.H., Zheng, X., Runcie, D.E., Covarrubias-Pazaran, G., Broeckling, C., Yao, L., Caffe-Treml, M., Gutiérrez, L.A., et al. (2021). Multi-omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations. Theor. Appl. Genet. **134**:4043–4054. https://doi.org/10.1007/s00122-021-03946-4.

Hu, X., Xie, W., Wu, C., and 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.

Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., Li, W., Zhan, Q., Cheng, B., Xia, J., et al. (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nat. Commun. **6**:6258. https://doi.org/10.1038/ncomms7258.

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M.E. (2015). Training set optimization under population structure in genomic selection. Theor. Appl. Genet. **128**:145–158. https://doi.org/10.1007/s00122-014-2418-4.

Isidro y Sánchez, J., and Akdemir, D. (2021). Training Set Optimization for Sparse Phenotyping in Genomic Selection: A Conceptual Overview. Front. Plant Sci. **12**. https://doi.org/10.3389/fpls.2021.715910.

Jannink, J.-L., Lorenz, A.J., and 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.

Jarquin, D., de Leon, N., Romay, C., Bohn, M., Buckler, E.S., Ciampitti, I., Edwards, J., Ertl, D., Flint-Garcia, S., Gore, M.A., et al. (2021). Utility of Climatic Information via Combining Ability Models to Improve Genomic Prediction for Yield Within the Genomes to Fields Maize Project. Front. Genet. **11**, 592769. https://doi.org/10.3389/fgene.2020.592769.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. **127**:595–607. https://doi.org/10.1007/s00122-013-2243-1.

Janss, L., De Los Campos, G., Sheehan, N., and Sorensen, D. (2012). Inferences from Genomic Models in Stratified Populations. Genetics **192**:693–704. https://doi.org/10.1534/genetics.112.141143.

Jiang, Y., Weise, S., Graner, A., and Reif, J.C. (2021). Using Genome-Wide Predictions to Assess the Phenotypic Variation of a Barley (Hordeum sp.) Gene Bank Collection for Important Agronomic Traits and Passport Information. Front. Plant Sci. **11**, 604781. https://doi.org/10.3389/fpls.2020.604781.

Johnson, M.E., Moore, L.M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. J. Stat. Plann. Inference **26**:131–148. https://doi.org/10.1016/0378-3758(90)90122-B.

Juliana, P., Singh, R.P., Braun, H.-J., Huerta-Espino, J., Crespo-Herrera, L., Govindan, V., Mondal, S., Poland, J., and Shrestha, S. (2020). Genomic Selection for Grain Yield in the CIMMYT Wheat Breeding Program—Status and Perspectives. Front. Plant Sci. **11**, 564183. https://doi.org/10.3389/fpls.2020.564183.

Juliana, P., Montesinos-López, O.A., Crossa, J., Mondal, S., González Pérez, L., Poland, J., Huerta-Espino, J., Crespo-Herrera, L., Govindan, V., Dreisigacker, S., et al. (2019a). Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. Theor. Appl. Genet. **132**:177–194. https://doi.org/10.1007/s00122-018-3206-3.

Juliana, P., Poland, J., Huerta-Espino, J., Shrestha, S., Crossa, J., Crespo-Herrera, L., Toledo, F.H., Govindan, V., Mondal, S., Kumar, U., et al. (2019b). Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat. Genet. **51**:1530–1539. https://doi.org/10.1038/s41588-019-0496-6.

Jung, M., Roth, M., Aranzana, M.J., Auwerkerken, A., Bink, M., Denancé, C., Dujak, C., Durel, C.-E., Font i Forcada, C., Cantin, C.M., et al. (2020). The apple REFPOP—a reference population for genomics-assisted breeding in apple. Hortic. Res. **7**, 189. https://doi.org/10.1038/s41438-020-00408-8.

Kadam, D.C., Potts, S.M., Bohn, M.O., Lipka, A.E., and Lorenz, A.J. (2016). Genomic Prediction of Single Crosses in the Early Stages of a Maize Hybrid Breeding Pipeline. G3 (Bethesda). **6**:3443–3453. https://doi.org/10.1534/g3.116.031286.

Kadam, D.C., Rodriguez, O.R., and Lorenz, A.J. (2021). Optimization of training sets for genomic prediction of early-stage single crosses in maize. Theor. Appl. Genet. **134**:687–699. https://doi.org/10.1007/s00122-020-03722-w.

Kaler, A.S., Purcell, L.C., Beissinger, T., and Gillman, J.D. (2022). Genomic prediction models for traits differing in heritability for soybean, rice, and maize. BMC Plant Biol. **22**, 87. https://doi.org/10.1186/s12870-022-03479-y.

Karaman, E., Cheng, H., Firat, M.Z., Garrick, D.J., and Fernando, R.L. (2016). An Upper Bound for Accuracy of Prediction Using GBLUP. PLoS One **11**, e0161054. https://doi.org/10.1371/journal.pone.0161054.

Kent, M.A., Fonseca, J.M.O., Klein, P.E., Klein, R.R., Hayes, C.M., and Rooney, W.L. (2023). Use of genomic prediction to screen sorghum B-lines in hybrid testcrosses. Plant Genome **16**:e20369. https://doi.org/10.1002/tpg2.20369.

Kim, G.W., Hong, J.-P., Lee, H.-Y., Kwon, J.-K., Kim, D.-A., and Kang, B.-C. (2022). Genomic selection with fixed-effect markers improves the prediction accuracy for Capsaicinoid contents in *Capsicum annuum*. Horticulture Research **9**. https://doi.org/10.1093/hr/uhac204.

Knoch, D., Werner, C.R., Meyer, R.C., Riewe, D., Abbadi, A., Lücke, S., Snowdon, R.J., and Altmann, T. (2021). Multi-omics-based prediction of hybrid performance in canola. Theor. Appl. Genet. **134**:1147–1165. https://doi.org/10.1007/s00122-020-03759-x.

Lado, B., Matus, I., Rodríguez, A., Inostroza, L., Poland, J., Belzile, F., del Pozo, A., Quincke, M., Castro, M., and von Zitzewitz, J. (2013). Increased Genomic Prediction Accuracy in Wheat Breeding Through Spatial Adjustment of Field Trial Data. G3 (Bethesda). **3**:2105–2114. https://doi.org/10.1534/g3.113.007807.

Laloë, D. (1993). Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. **25**:557–576.

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics **124**:743–756. https://doi.org/10.1093/genetics/124.3.743.

Lehermeier, C., Schön, C.C., and De Los Campos, G. (2015). Assessment of Genetic Heterogeneity in Structured Plant Populations Using Multivariate Whole-Genome Regression Models. Genetics **201**:323–337. https://doi.org/10.1534/genetics.115.177394.

Lemeunier, P., Paux, E., Babi, S., Auzanneau, J., Goudemand-Dugué, E., Ravel, C., and Rincent, R. (2022). Training population optimization for genomic selection improves the predictive ability of a costly measure in bread wheat, the gliadin to glutenin ratio. Euphytica **218**:111. https://doi.org/10.1007/s10681-022-03062-4.

Li, B., Zhang, N., Wang, Y.-G., George, A.W., Reverter, A., and Li, Y. (2018). Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. Front. Genet. **9**. https://doi.org/10.3389/fgene.2018.00237.

Li, X., Guo, T., Wang, J., Bekele, W.A., Sukumaran, S., Vanous, A.E., Mcnellie, J.P., Tibbs-Cortes, L.E., Lopes, M.S., Lamkey, K.R., et al. (2021). An integrated framework reinstating the environmental dimension for GWAS and genomic selection in crops. Mol. Plant **14**:874–887. https://doi.org/10.1016/j.molp.2021.03.010.

Li, Z., Gao, N., Martini, J.W.R., and Simianer, H. (2019). Integrating Gene Expression Data Into Genomic Prediction. Front. Genet. **10**. https://doi.org/10.3389/fgene.2019.00126.

Li, Z., Philipp, N., Spiller, M., Stiewe, G., Reif, J.C., and Zhao, Y. (2017). Genome-Wide Prediction of the Performance of Three-Way Hybrids in Barley. Plant Genome **10**, plantgenome2016. https://doi.org/10.3835/plantgenome2016.05.0046.

Liu, X., Wang, H., Wang, H., Guo, Z., Xu, X., Liu, J., Wang, S., Li, W.-X., Zou, C., Prasanna, B.M., et al. (2018). Factors affecting genomic selection revealed by empirical evidence in maize. The Crop Journal **6**:341–352. https://doi.org/10.1016/j.cj.2018.03.005.

Lopez-Cruz, M., and de los Campos, G. (2021). Optimal breeding-value prediction using a sparse selection index. Genetics **218**, iyab030. https://doi.org/10.1093/genetics/iyab030.

Lopez-Cruz, M., Beyene, Y., Gowda, M., Crossa, J., Pérez-Rodríguez, P., and de los Campos, G. (2021). Multi-generation genomic prediction of maize yield using parametric and non-parametric sparse selection indices. Heredity **127**:423–432. https://doi.org/10.1038/s41437-021-00474-1.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.-L., Singh, R.P., Autrique, E., and de los Campos, G. (2015). Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker × Environment Interaction Genomic Selection Model. G3 (Bethesda). **5**:569–582. https://doi.org/10.1534/g3.114.016097.

Lopez-Cruz, M., Dreisigacker, S., Crespo-Herrera, L., Bentley, A.R., Singh, R., Poland, J., Shrestha, S., Huerta-Espino, J., Govindan, V., Juliana, P., et al. (2022). Sparse kernel models provide optimization of training set design for genomic prediction in multiyear wheat breeding data. Plant Genome **15**, e20254. https://doi.org/10.1002/tpg2.20254.

Lorenz, A., and Nice, L. (2017). Training Population Design and Resource Allocation for Genomic Selection in Plant Breeding. In Genomic Selection for Crop Improvement (Springer International Publishing), pp. 7–22. https://doi.org/10.1007/978-3-319-63170-7_2.

Lorenz, A.J., and Smith, K.P. (2015). Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. Crop Sci. **55**:2657–2667. https://doi.org/10.2135/cropsci2014.12.0827.

Lorenz, A.J., Smith, K.P., and Jannink, J.L. (2012). Potential and Optimization of Genomic Selection for Fusarium Head Blight Resistance in Six-Row Barley. Crop Sci. **52**:1609–1621. https://doi.org/10.2135/cropsci2011.09.0503.

Lorenzana, R.E., and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. **120**:151–161. https://doi.org/10.1007/s00122-009-1166-3.

Lyra, D.H., Granato, Í.S.C., Morais, P.P.P., Alves, F.C., Dos Santos, A.R.M., Yu, X., Guo, T., Yu, J., and Fritsche-Neto, R. (2018). Controlling population structure in the genomic prediction of tropical maize hybrids. Mol. Breed. **38**, 126. https://doi.org/10.1007/s11032-018-0882-2.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., and 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.

Maenhout, S., De Baets, B., Haesaert, G., and Van Bockstaele, E. (2007). Support vector machine regression for the prediction of maize hybrid performance. Theor. Appl. Genet. **115**:1003–1013. https://doi.org/10.1007/s00122-007-0627-9.

Mageto, E.K., Crossa, J., Pérez-Rodríguez, P., Dhliwayo, T., Palacios-Rojas, N., Lee, M., Guo, R., San Vicente, F., Zhang, X., and Hindu, V. (2020). Genomic Prediction with Genotype by Environment Interaction Analysis for Kernel Zinc Concentration in Tropical Maize Germplasm. G3 (Bethesda). **10**:2629–2639. https://doi.org/10.1534/g3.120.401172.

Mangin, B., Rincent, R., Rabier, C.-E., Moreau, L., and Goudemand-Dugue, E. (2019). Training set optimization of genomic prediction by means of EthAcc. PLoS One **14**, e0205629. https://doi.org/10.1371/journal.pone.0205629.

Martini, J.W.R., Gao, N., and Crossa, J. (2022). Incorporating Omics Data in Genomic Prediction. In Methods in Molecular Biology (Springer US), pp. 341–357. https://doi.org/10.1007/978-1-0716-2205-6_12.

Massman, J.M., Gordillo, A., Lorenzana, R.E., and Bernardo, R. (2013). Genomewide predictions from maize single-cross data. Theor. Appl. Genet. **126**:13–22. https://doi.org/10.1007/s00122-012-1955-y.

Maulana, F., Perumal, R., Serba, D.D., and Tesso, T. (2023). Genomic prediction of hybrid performance in grain sorghum (Sorghum bicolor L.). Front. Plant Sci. **14**, 1139896. https://doi.org/10.3389/fpls.2023.1139896.

Melchinger, A.E., Fernando, R., Stricker, C., Schön, C.C., and Auinger, H.-J. (2023). Genomic prediction in hybrid breeding: I. Optimizing the training set design. Theor. Appl. Genet. **136**:176. https://doi.org/10.1007/s00122-023-04413-y.

Mendonça, L.d.F., and Fritsche-Neto, R. (2020). The accuracy of different strategies for building training sets for genomic predictions in segregating soybean populations. Crop Sci. **60**:3115–3126. https://doi.org/10.1002/csc2.20267.

Merrick, L.F., Herr, A.W., Sandhu, K.S., Lozada, D.N., and Carter, A.H. (2022). Optimizing Plant Breeding Programs for Genomic Selection. Agronomy **12**:714. https://doi.org/10.3390/agronomy12030714.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics **157**:1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Mitchell, T.J. (2000). An Algorithm for the Construction of "D-Optimal" Experimental Designs. Technometrics **42**:48–54. https://doi.org/10.2307/1271431.

Momen, M., and Morota, G. (2018). Quantifying genomic connectedness and prediction accuracy from additive and non-additive gene actions. Genet. Sel. Evol. **50**:45. https://doi.org/10.1186/s12711-018-0415-9.

Montesinos-López, O.A., Mosqueda-González, B.A., Salinas-Ruiz, J., Montesinos-López, A., and Crossa, J. (2023a). Sparse multi-trait genomic prediction under balanced incomplete block design. Plant Genome **16**, e20305. https://doi.org/10.1002/tpg2.20305.

Montesinos-López, O.A., Montesinos-López, J.C., Singh, P., Lozano-Ramirez, N., Barrón-López, A., Montesinos-López, A., and Crossa, J. (2020). A Multivariate Poisson Deep Learning Model for Genomic Prediction of Count Data. G3 (Bethesda). **10**:4177–4190. https://doi.org/10.1534/g3.120.401631.

Montesinos-López, O.A., Montesinos-López, A., Kismiantini Roman-Gallardo, A., Roman-Gallardo, A., Gardner, K., Lillemo, M., Fritsche-Neto, R., and Crossa, J. (2022a). Partial Least Squares Enhances Genomic Prediction of New Environments. Front. Genet. **13**, 920689. https://doi.org/10.3389/fgene.2022.920689.

Montesinos-López, O.A., Saint Pierre, C., Gezan, S.A., Bentley, A.R., Mosqueda-González, B.A., Montesinos-López, A., Van Eeuwijk, F., Beyene, Y., Gowda, M., Gardner, K., et al. (2023b). Optimizing Sparse Testing for Genomic Prediction of Plant Breeding Crops. Genes **14**:927. https://doi.org/10.3390/genes14040927.

Montesinos-López, O.A., and Montesinos-López, A. (2023). Designing optimal training sets for genomic prediction using adversarial validation with probit regression. Plant Breed. **142**:594–606. https://doi.org/10.1111/pbr.13124.

Montesinos López, O.A., Montesinos López, A., and Crossa, J. (2022b). General Elements of Genomic Selection and Statistical Learning. In Multivariate Statistical Machine Learning Methods for Genomic Prediction (Springer International Publishing), pp. 1–34. https://doi.org/10.1007/978-3-030-89010-0_1.

Montesinos López, O.A., Mosqueda González, B.A., Montesinos López, A., and Crossa, J. (2023). Statistical Machine-Learning Methods for Genomic Prediction Using the SKM Library. Genes **14**:1003. https://doi.org/10.3390/genes14051003.

Moreira, F.F., Oliveira, H.R., Volenec, J.J., Rainey, K.M., and Brito, L.F. (2020). Integrating High-Throughput Phenotyping and Statistical Genomic Methods to Genetically Improve Longitudinal Traits in Crops. Front. Plant Sci. **11**, 681. https://doi.org/10.3389/fpls.2020.00681.

Morgante, F., Huang, W., Maltecca, C., and Mackay, T.F.C. (2018). Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. Heredity **120**:500–514. https://doi.org/10.1038/s41437-017-0043-0.

Nejati-Javaremi, A., Smith, C., and Gibson, J.P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. **75**:1738–1745. https://doi.org/10.2527/1997.7571738x.

Mu, Q., Guo, T., Li, X., and Yu, J. (2022). Phenotypic plasticity in plant height shaped by interaction between genetic loci and diurnal temperature range. New Phytol. **233**:1768–1779. https://doi.org/10.1111/nph.17904.

Neyhart, J.L., Tiede, T., Lorenz, A.J., and Smith, K.P. (2017). Evaluating Methods of Updating Training Data in Long-Term Genomewide Selection. G3 (Bethesda). **7**:1499–1510. https://doi.org/10.1534/g3.117.040550.

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. G3 (Bethesda). **8**:2889–2899. https://doi.org/10.1534/g3.118.200311.

Olatoye, M.O., Clark, L.V., Labonte, N.R., Dong, H., Dwiyanti, M.S., Anzoua, K.G., Brummer, J.E., Ghimire, B.K., Dzyubenko, E., Dzyubenko, N., et al. (2020). Training Population Optimization for Genomic Selection in Miscanthus. G3 (Bethesda). **10**:2465–2476. https://doi.org/10.1534/g3.120.401402.

Ornella, L., Singh, S., Perez, P., Burgueño, J., Singh, R., Tapia, E., Bhavani, S., Dreisigacker, S., Braun, H.J., Mathews, K., et al. (2012). Genomic Prediction of Genetic Values for Resistance to Wheat Rusts. Plant Genome **5**:136. https://doi.org/10.3835/plantgenome2012.07.0017.

Ortiz, R., Crossa, J., Reslow, F., Perez-Rodriguez, P., and Cuevas, J. (2022). Genome-Based Genotype × Environment Prediction Enhances Potato (Solanum tuberosum L.) Improvement Using Pseudo-Diploid and Polysomic Tetraploid Modeling. Front. Plant Sci. **13**. https://doi.org/10.3389/fpls.2022.785196.

Ortiz, R., Reslow, F., Vetukuri, R., García-Gil, M.R., Pérez-Rodríguez, P., and Crossa, J. (2023a). Inbreeding Effects on the Performance and Genomic Prediction for Polysomic Tetraploid Potato Offspring Grown at High Nordic Latitudes. Genes **14**:1302. https://doi.org/10.3390/genes14061302.

Ortiz, R., Reslow, F., Montesinos-López, A., Huicho, J., Pérez-Rodríguez, P., Montesinos-López, O.A., and Crossa, J. (2023b). Partial least squares enhance multi-trait genomic prediction of potato cultivars in new environments. Sci. Rep. **13**, 9947. https://doi.org/10.1038/s41598-023-37169-y.

Ou, J.-H., and Liao, C.-T. (2019). Training set determination for genomic selection. Theor. Appl. Genet. **132**:2781–2792. https://doi.org/10.1007/s00122-019-03387-0.

Owens, B.F., Lipka, A.E., Magallanes-Lundback, M., Tiede, T., Diepenbrock, C.H., Kandianis, C.B., Kim, E., Cepela, J., Mateos-Hernandez, M., Buell, C.R., et al. (2014). A Foundation for Provitamin A Biofortification of Maize: Genome-Wide Association and Genomic Prediction Models of Carotenoid Levels. Genetics **198**:1699–1716. https://doi.org/10.1534/genetics.114.169979.

Pace, J., Yu, X., and Lübberstedt, T. (2015). Genomic prediction of seedling root length in maize (Zea mays L.). Plant J. **83**:903–912. https://doi.org/10.1111/tpj.12937.

Pandey, J., Scheuring, D.C., Koym, J.W., Endelman, J.B., and Vales, M.I. (2023). Genomic selection and genome-wide association studies in tetraploid chipping potatoes. Plant Genome **16**:e20297. https://doi.org/10.1002/tpg2.20297.

Park, T., and Casella, G. (2008). The Bayesian Lasso. J. Am. Stat. Assoc. **103**:681–686. https://doi.org/10.1198/016214508000000337.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J.M., Crossa, J., Manès, Y., and Dreisigacker, S. (2012). Comparison Between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat. G3 (Bethesda). **2**:1595–1605. https://doi.org/10.1534/g3.112.003665.

Pérez-Rodríguez, P., Crossa, J., Bondalapati, K., De Meyer, G., Pita, F., and Campos, G.d.l. (2015). A Pedigree-Based Reaction Norm Model for Prediction of Cotton Yield in Multienvironment Trials. Crop Sci. **55**:1143–1151. https://doi.org/10.2135/cropsci2014.08.0577.

Pérez, P., De Los Campos, G., Crossa, J., and 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.

Philipp, N., Liu, G., Zhao, Y., He, S., Spiller, M., Stiewe, G., Pillen, K., Reif, J.C., and Li, Z. (2016). Genomic Prediction of Barley Hybrid Performance. Plant Genome **9**. https://doi.org/10.3835/plantgenome2016.02.0016.

Pong-Wong, R., and Woolliams, J. (2014). Bayes U: A Genomic Prediction Method Based on the Horseshoe Prior (World Congress of Genetics Applied to Livestock Production).

Pszczola, M., and Calus, M.P.L. (2016). Updating the reference population to achieve constant genomic prediction reliability across generations. Animal **10**:1018–1024. https://doi.org/10.1017/S1751731115002785.

Pszczola, M., Strabel, T., Mulder, H.A., and Calus, M.P.L. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. **95**:389–400. https://doi.org/10.3168/jds.2011-4338.

Rakotondramanana, M., Tanaka, R., Pariasca-Tanaka, J., Stangoulis, J., Grenier, C., and Wissuwa, M. (2022). Genomic prediction of zinc-biofortification potential in rice gene bank accessions. Theor. Appl. Genet. **135**:2265–2278. https://doi.org/10.1007/s00122-022-04110-2.

Rembe, M., Zhao, Y., Wendler, N., Oldach, K., Korzun, V., and Reif, J.C. (2022). The Potential of Genome-Wide Prediction to Support Parental Selection, Evaluated with Data from a Commercial Barley Breeding Program. Plants **11**:2564. https://doi.org/10.3390/plants11192564.

Riedelsheimer, C., Endelman, J.B., Stange, M., Sorrells, M.E., Jannink, J.-L., and Melchinger, A.E. (2013). Genomic Predictability of Interconnected Biparental Maize Populations. Genetics **194**:493–503. https://doi.org/10.1534/genetics.113.150227.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., and 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.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V.M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., et al. (2012). Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (*Zea mays* L.). Genetics **192**:715–728. https://doi.org/10.1534/genetics.112.141473.

Rincent, R., Charcosset, A., and Moreau, L. (2017). Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. Theor. Appl. Genet. **130**:2231–2247. https://doi.org/10.1007/s00122-017-2956-7.

Rio, S., Gallego-Sánchez, L., Montilla-Bascón, G., Canales, F.J., Isidro Y Sánchez, J., and Prats, E. (2021). Genomic prediction and training set optimization in a structured Mediterranean oat population. Theor. Appl. Genet. **134**:3595–3609. https://doi.org/10.1007/s00122-021-03916-w.

Rio, S., Mary-Huard, T., Moreau, L., and Charcosset, A. (2019). Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. Theor. Appl. Genet. **132**:81–96. https://doi.org/10.1007/s00122-018-3196-1.

Rio, S., Akdemir, D., Carvalho, T., and Sánchez, J.I.Y. (2022). Assessment of genomic prediction reliability and optimization of experimental designs in multi-environment trials. Theor. Appl. Genet. **135**:405–419. https://doi.org/10.1007/s00122-021-03972-2.

Rogers, A.R., Bian, Y., Krakowsky, M., Peters, D., Turnbull, C., Nelson, P., and Holland, J.B. (2022). Genomic prediction for the Germplasm Enhancement of Maize project. Plant Genome **15**:e20267. https://doi.org/10.1002/tpg2.20267.

Roth, M., Muranty, H., Di Guardo, M., Guerra, W., Patocchi, A., and 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.

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L.G., Crossa, J., Reynolds, M., and Singh, R. (2016). Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. G3 (Bethesda). **6**:2799–2808. https://doi.org/10.1534/g3.116.032888.

Saint Pierre, C., Burgueño, J., Crossa, J., Fuentes Dávila, G., Figueroa López, P., Solís Moya, E., Ireta Moreno, J., Hernández Muela, V.M., Zamora Villa, V.M., Vikram, P., et al. (2016). Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. Sci. Rep. **6**, 27312. https://doi.org/10.1038/srep27312.

Sapkota, S., Boatwright, J.L., Kumar, N., Myers, M., Cox, A., Ackerman, A., Caughman, W., Brenton, Z.W., Boyles, R.E., and Kresovich, S. (2022). Genomic prediction of hybrid performance for agronomic traits in sorghum. G3 (Bethesda). **13**, jkac311. https://doi.org/10.1093/g3journal/jkac311.

Sarinelli, J.M., Murphy, J.P., Tyagi, P., Holland, J.B., Johnson, J.W., Mergoum, M., Mason, R.E., Babar, A., Harrison, S., Sutton, R., et al. (2019). Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. Theor. Appl. Genet. **132**:1247–1261. https://doi.org/10.1007/s00122-019-03276-6.

Schrag, T.A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., and 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., Kale, S.M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., Jiang, Y., Beukert, U., Serfling, A., Himmelbach, A., et al. (2022). Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. Nat. Genet. **54**:1544–1552. https://doi.org/10.1038/s41588-022-01189-7.

Sehgal, D., Rosyara, U., Mondal, S., Singh, R., Poland, J., and Dreisigacker, S. (2020). Incorporating Genome-Wide Association Mapping Results Into Genomic Prediction Models for Grain Yield and Yield Stability in CIMMYT Spring Bread Wheat. Front. Plant Sci. **11**, 197. https://doi.org/10.3389/fpls.2020.00197.

Selga, C., Koc, A., Chawade, A., and Ortiz, R. (2020). A Bioinformatics Pipeline to Identify a Subset of SNPs for Genomics-Assisted Potato Breeding. Plants **10**:30. https://doi.org/10.3390/plants10010030.

Selga, C., Reslow, F., Pérez-Rodríguez, P., and Ortiz, R. (2021). The power of genomic estimated breeding values for selection when using a finite population size in genetic improvement of tetraploid potato. G3 (Bethesda). **12**, jkab362. https://doi.org/10.1093/g3journal/jkab362.

Semagn, K., Iqbal, M., Jarquin, D., Randhawa, H., Aboukhaddour, R., Howard, R., Ciechanowska, I., Farzand, M., Dhariwal, R., Hiebert, C.W., et al. (2022a). Genomic Prediction Accuracy of Stripe Rust in Six Spring Wheat Populations by Modeling Genotype by Environment Interaction. Plants **11**:1736. https://doi.org/10.3390/plants11131736.

Semagn, K., Iqbal, M., Jarquin, D., Crossa, J., Howard, R., Ciechanowska, I., Henriquez, M.A., Randhawa, H., Aboukhaddour, R., McCallum, B.D., et al. (2022b). Genomic Predictions for Common Bunt, FHB, Stripe Rust, Leaf Rust, and Leaf Spotting Resistance in Spring Wheat. Genes **13**:565. https://doi.org/10.3390/genes13040565.

Shahi, D., Guo, J., Pradhan, S., Khan, J., Avci, M., Khan, N., Mcbreen, J., Bai, G., Reynolds, M., Foulkes, J., and Babar, M.A. (2022). Multi-trait genomic prediction using in-season physiological parameters increases prediction accuracy of complex traits in US wheat. BMC Genom. **23**, 298. https://doi.org/10.1186/s12864-022-08487-8.

Shahinnia, F., Geyer, M., Schürmann, F., Rudolphi, S., Holzapfel, J., Kempf, H., Stadlmeier, M., Löschenberger, F., Morales, L., Buerstmayr, H., et al. (2022). Genome-wide association study and genomic prediction of resistance to stripe rust in current Central and Northern European winter wheat germplasm. Theor. Appl. Genet. **135**:3583–3595. https://doi.org/10.1007/s00122-022-04202-z.

Shi, S., Li, X., Fang, L., Liu, A., Su, G., Zhang, Y., Luobu, B., Ding, X., and Zhang, S. (2021). Genomic Prediction Using Bayesian Regression Models With Global–Local Prior. Front. Genet. **12**, 628205. https://doi.org/10.3389/fgene.2021.628205.

Silva, F.F., Jerez, E.A.Z., De Resende, M.D.V., Viana, J.M.S., Azevedo, C.F., Lopes, P.S., Nascimento, M., De Lima, R.O., and Guimarães, S.E.F. (2018). Bayesian model combining linkage and linkage disequilibrium analysis for low density-based genomic selection in animal breeding. J. Appl. Anim. Res. **46**:873–878. https://doi.org/10.1080/09712119.2017.1415903.

Sirsat, M.S., Oblessuc, P.R., and Ramiro, R.S. (2022). Genomic Prediction of Wheat Grain Yield Using Machine Learning. Agriculture **12**:1406. https://doi.org/10.3390/agriculture12091406.

Sitonik, C., Suresh, L.M., Beyene, Y., Olsen, M.S., Makumbi, D., Oliver, K., Das, B., Bright, J.M., Mugo, S., Crossa, J., et al. (2019). Genetic architecture of maize chlorotic mottle virus and maize lethal necrosis through GWAS, linkage analysis and genomic prediction in tropical maize germplasm. Theor. Appl. Genet. **132**:2381–2399. https://doi.org/10.1007/s00122-019-03360-x.

Solberg, T.R., Sonesson, A.K., Woolliams, J.A., and Meuwissen, T.H.E. (2009). Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. **41**:29. https://doi.org/10.1186/1297-9686-41-29.

Soller, M., and Plotkin-Hazan, J. (1977). The use marker alleles for the introgression of linked quantitative alleles. Theor. Appl. Genet. **51**:133–137. https://doi.org/10.1007/bf00273825.

Sood, S., Lin, Z., Caruana, B., Slater, A.T., and Daetwyler, H.D. (2020). Making the most of all data: Combining non-genotyped and genotyped potato individuals with HBLUP. Plant Genome **13**:e20056. https://doi.org/10.1002/tpg2.20056.

Sood, S., Bhardwaj, V., Bairwa, A., Dalamu Sharma, S., Sharma, S., Sharma, A.K., Kumar, A., Lal, M., and Kumar, V. (2023). Genome-wide association mapping and genomic prediction for late blight and potato cyst nematode resistance in potato (Solanum tuberosum L.). Front. Plant Sci. **14**, 1211472. https://doi.org/10.3389/fpls.2023.1211472.

Stich, B., and Van Inghelandt, D. (2018). Prospects and Potential Uses of Genomic Prediction of Key Performance Traits in Tetraploid Potato. Front. Plant Sci. **9**, 159. https://doi.org/10.3389/fpls.2018.00159.

Sukumaran, S., Crossa, J., Jarquin, D., Lopes, M., and Reynolds, M.P. (2017). Genomic Prediction with Pedigree and Genotype × Environment Interaction in Spring Wheat Grown in South and West Asia, North Africa, and Mexico. G3 (Bethesda). **7**:481–495. https://doi.org/10.1534/g3.116.036251.

Sun, J., Rutkoski, J.E., Poland, J.A., Crossa, J., Jannink, J.L., and Sorrells, M.E. (2017). Multitrait, Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. Plant Genome **10**. https://doi.org/10.3835/plantgenome2016.11.0111.

Sverrisdóttir, E., Byrne, S., Sundmark, E.H.R., Johnsen, H.Ø., Kirk, H.G., Asp, T., Janss, L., and Nielsen, K.L. (2017). Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. Theor. Appl. Genet. **130**:2091–2108. https://doi.org/10.1007/s00122-017-2944-y.

Sverrisdóttir, E., Sundmark, E.H.R., Johnsen, H.Ø., Kirk, H.G., Asp, T., Janss, L., Bryan, G., and Nielsen, K.L. (2018). The Value of Expanding the Training Population to Improve Genomic Selection Models in Tetraploid Potato. Front. Plant Sci. **9**, 1118. https://doi.org/10.3389/fpls.2018.01118.

Tadesse, W., Gataa, Z.E., Rachdad, F.E., Baouchi, A.E., Kehel, Z., and Alemu, A. (2023). Single- and multi-trait genomic prediction and genome-wide association analysis of grain yield and micronutrient-related traits in ICARDA wheat under drought environment. Mol. Genet. Genom. **298**:1515–1526. https://doi.org/10.1007/s00438-023-02074-6.

Tadesse, W., Sanchez-Garcia, M., Assefa, S.G., Amri, A., Bishaw, Z., Ogbonnaya, F.C., and Baum, M. (2019). Genetic Gains in Wheat Breeding and Its Role in Feeding the World. rop Breed Genet Genom **1**, e190005. https://doi.org/10.20900/cbgg20190005.

Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., Chabert-Martinello, M., Magnin-Robert, J.-B., Marget, P., Aubert, G., and Burstin, J. (2015). Genomic Prediction in Pea: Effect of Marker Density and Training Population Size and Composition on Prediction Accuracy. Front. Plant Sci. **6**, 941. https://doi.org/10.3389/fpls.2015.00941.

Technow, F., Bürger, A., and Melchinger, A.E. (2013). Genomic Prediction of Northern Corn Leaf Blight Resistance in Maize with Combined or Separated Training Sets for Heterotic Groups. G3 (Bethesda). **3**:197–203. https://doi.org/10.1534/g3.112.004630.

Tessema, B.B., Liu, H., Sørensen, A.C., Andersen, J.R., and Jensen, J. (2020). Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat. Front. Genet. **11**, 578123. https://doi.org/10.3389/fgene.2020.578123.

Tibbs Cortes, L., Zhang, Z., and Yu, J. (2021). Status and prospects of genome-wide association studies in plants. Plant Genome **14**:e20077. https://doi.org/10.1002/tpg2.20077.

Tomar, V., Dhillon, G.S., Singh, D., Singh, R.P., Poland, J., Chaudhary, A.A., Bhati, P.K., Joshi, A.K., and Kumar, U. (2021a). Evaluations of Genomic Prediction and Identification of New Loci for Resistance to Stripe Rust Disease in Wheat (Triticum aestivum L.). Front. Genet. **12**, 710485. https://doi.org/10.3389/fgene.2021.710485.

Tomar, V., Singh, D., Dhillon, G.S., Chung, Y.S., Poland, J., Singh, R.P., Joshi, A.K., Gautam, Y., Tiwari, B.S., and Kumar, U. (2021b). Increased Predictive Accuracy of Multi-Environment Genomic Prediction Model for Yield and Related Traits in Spring Wheat (Triticum aestivum L.). Front. Plant Sci. **12**, 720123. https://doi.org/10.3389/fpls.2021.720123.

Tong, H., and Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. J. Plant Physiol. **257**, 153354. https://doi.org/10.1016/j.jplph.2020.153354.

Tsai, H.-Y., Janss, L.L., Andersen, J.R., Orabi, J., Jensen, J.D., Jahoor, A., and Jensen, J. (2020). Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. Sci. Rep. **10**, 3347. https://doi.org/10.1038/s41598-020-60203-2.

Usai, M.G., Goddard, M.E., and Hayes, B.J. (2009). LASSO with cross-validation for genomic selection. Genet. Res. **91**:427–436. https://doi.org/10.1017/s0016672309990334.

Van Den Berg, I., Boichard, D., Guldbrandtsen, B., and Lund, M.S. (2016). Using Sequence Variants in Linkage Disequilibrium with Causative Mutations to Improve Across-Breed Prediction in Dairy Cattle: A Simulation Study. G3 (Bethesda). **6**:2553–2561. https://doi.org/10.1534/g3.116.027730.

Vanraden, P.M. (2008). Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. **91**:4414–4423. https://doi.org/10.3168/jds.2007-0980.

Varshney, R.K., Bohra, A., Yu, J., Graner, A., Zhang, Q., and Sorrells, M.E. (2021). Designing Future Crops: Genomics-Assisted Breeding Comes of Age. Trends Plant Sci. **26**:631–649. https://doi.org/10.1016/j.tplants.2021.03.010.

Vélez-Torres, M., García-Zavala, J.J., Hernández-Rodríguez, M., Lobato-Ortiz, R., López-Reynoso, J.J., Benítez-Riquelme, I., Mejía-Contreras, J.A., Esquivel-Esquivel, G., Molina-Galán, J.D., Pérez-Rodríguez, P., and Zhang, X. (2018). Genomic prediction of the general combining ability of maize lines (*Zea mays* L.) and the performance of their single crosses. Plant Breed. **137**:379–387. https://doi.org/10.1111/pbr.12597.

Velu, G., Crossa, J., Singh, R.P., Hao, Y., Dreisigacker, S., Perez-Rodriguez, P., Joshi, A.K., Chatrath, R., Gupta, V., Balasubramaniam, A., et al. (2016). Genomic prediction for grain zinc and iron concentrations in spring wheat. Theor. Appl. Genet. **129**:1595–1605. https://doi.org/10.1007/s00122-016-2726-y.

Velu, G., Singh, R.P., Crespo-Herrera, L., Juliana, P., Dreisigacker, S., Valluru, R., Stangoulis, J., Sohu, V.S., Mavi, G.S., Mishra, V.K., et al. (2018). Genetic dissection of grain zinc concentration in spring wheat for mainstreaming biofortification in CIMMYT wheat breeding. Sci. Rep. **8**, 13526. https://doi.org/10.1038/s41598-018-31951-z.

Waldmann, P. (2016). Genome-wide prediction using Bayesian additive regression trees. Genet. Sel. Evol. **48**:42. https://doi.org/10.1186/s12711-016-0219-8.

Wang, D., Salah El-Basyoni, I., Stephen Baenziger, P., Crossa, J., Eskridge, K.M., and Dweikat, I. (2012). Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. Heredity **109**:313–319. https://doi.org/10.1038/hdy.2012.44.

Wang, K., Abid, M.A., Rasheed, A., Crossa, J., Hearne, S., and Li, H. (2023a). DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol. Plant **16**:279–293. https://doi.org/10.1016/j.molp.2022.11.004.

Wang, N., Wang, H., Zhang, A., Liu, Y., Yu, D., Hao, Z., Ilut, D., Glaubitz, J.C., Gao, Y., Jones, E., et al. (2020). Genomic prediction across years in a maize doubled haploid breeding program to accelerate early-stage testcross testing. Theor. Appl. Genet. **133**:2869–2879. https://doi.org/10.1007/s00122-020-03638-5.

Wang, S., Wei, J., Li, R., Qu, H., Chater, J.M., Ma, R., Li, Y., Xie, W., and Jia, Z. (2019). Identification of optimal prediction models using multi-omic data for selecting hybrid rice. Heredity **123**:395–406. https://doi.org/10.1038/s41437-019-0210-6.

Wang, W., Guo, W., Le, L., Yu, J., Wu, Y., Li, D., Wang, Y., Wang, H., Lu, X., Qiao, H., et al. (2023b). Integration of high-throughput phenotyping, GWAS, and predictive models reveals the genetic architecture of plant height in maize. Mol. Plant **16**:354–373. https://doi.org/10.1016/j.molp.2022.11.016.

Weber, S.E., Frisch, M., Snowdon, R.J., and Voss-Fels, K.P. (2023). Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. Front. Plant Sci. **14**, 1217589. https://doi.org/10.3389/fpls.2023.1217589.

Werner, C.R., Gaynor, R.C., Gorjanc, G., Hickey, J.M., Kox, T., Abbadi, A., Leckband, G., Snowdon, R.J., and Stahl, A. (2020). How Population Structure Impacts Genomic Selection Accuracy in Cross-Validation: Implications for Practical Breeding. Front. Plant Sci. **11**, 592977. https://doi.org/10.3389/fpls.2020.592977.

Westhues, M., Schrag, T.A., Heuer, C., Thaller, G., Utz, H.F., Schipprack, W., Thiemann, A., Seifert, F., Ehret, A., Schlereth, A., et al. (2017). Omics-based hybrid prediction in maize. Theor. Appl. Genet. **130**:1927–1939. https://doi.org/10.1007/s00122-017-2934-0.

Whittaker, J.C., Thompson, R., and Denham, M.C. (2000). Marker-assisted selection using ridge regression. Genet. Res. **75**:249–252. https://doi.org/10.1017/s0016672399004462.

Wientjes, Y.C.J., Veerkamp, R.F., and Calus, M.P.L. (2013). The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics **193**:621–631. https://doi.org/10.1534/genetics.112.146290.

Wilson, S., Zheng, C., Maliepaard, C., Mulder, H.A., Visser, R.G.F., van der Burgt, A., and van Eeuwijk, F. (2021). Understanding the Effectiveness of Genomic Prediction in Tetraploid Potato. Front. Plant Sci. **12**, 672417. https://doi.org/10.3389/fpls.2021.672417.

Windhausen, V.S., Atlin, G.N., Hickey, J.M., Crossa, J., Jannink, J.-L., Sorrells, M.E., Raman, B., Cairns, J.E., Tarekegne, A., Semagn, K., et al. (2012). Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and Environments. G3 (Bethesda). **2**:1427–1436. https://doi.org/10.1534/g3.112.003699.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. **14**:507–515. https://doi.org/10.1038/nrg3457.

Wu, P.Y., Stich, B., Weisweiler, M., Shrestha, A., Erban, A., Westhoff, P., and Inghelandt, D.V. (2022). Improvement of prediction ability by integrating multi-omic datasets in barley. BMC Genom. **23**, 200. https://doi.org/10.1186/s12864-022-08337-7.

Wu, P.Y., Tung, C.W., Lee, C.Y., and Liao, C.T. (2019). Genomic Prediction of Pumpkin Hybrid Performance. Plant Genome **12**, 180082. https://doi.org/10.3835/plantgenome2018.10.0082.

Wu, P.-Y., Ou, J.-H., and Liao, C.-T. (2023). Sample size determination for training set optimization in genomic prediction. Theor. Appl. Genet. **136**:57. https://doi.org/10.1007/s00122-023-04254-9.

Wu, X.-L., Xu, J., Feng, G., Wiggans, G.R., Taylor, J.F., He, J., Qian, C., Qiu, J., Simpson, B., Walker, J., and Bauck, S. (2016). Optimal Design of Low-Density SNP Arrays for Genomic Prediction: Algorithm and Applications. PLoS One **11**, e0161719. https://doi.org/10.1371/journal.pone.0161719.

Würschum, T., Maurer, H.P., Weissmann, S., Hahn, V., and Leiser, W.L. (2017). Accuracy of within- and among-family genomic prediction in triticale. Plant Breed. **136**:230–236. https://doi.org/10.1111/pbr.12465.

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

Xu, Y., Xu, C., and Xu, S. (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. Heredity **119**:174–184. https://doi.org/10.1038/hdy.2017.27.

Xu, Y., Wang, X., Ding, X., Zheng, X., Yang, Z., Xu, C., and Hu, Z. (2018). Genomic selection of agronomic traits in hybrid rice using an NCII population. Rice **11**. https://doi.org/10.1186/s12284-018-0223-4.

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., Prasanna, B.M., Olsen, M.S., Wang, G., and Zhang, A. (2020). Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. Plant Commun. **1**, 100005. https://doi.org/10.1016/j.xplc.2019.100005.

Yan, J., and Wang, X. (2023). Machine learning bridges omics sciences and plant breeding. Trends Plant Sci. **28**:199–210. https://doi.org/10.1016/j.tplants.2022.08.018.

Yu, G., Cui, Y., Jiao, Y., Zhou, K., Wang, X., Yang, W., Xu, Y., Yang, K., Zhang, X., Li, P., et al. (2023). Comparison of sequencing-based and array-based genotyping platforms for genomic prediction of maize hybrid performance. The Crop Journal **11**:490–498. https://doi.org/10.1016/j.cj.2022.09.004.

Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S.E., Roozeboom, K.L., Wang, D., Wang, M.L., Pederson, G.A., et al. (2016). Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat. Plants **2**, 16150. https://doi.org/10.1038/nplants.2016.150.

Yu, X., Leiboff, S., Li, X., Guo, T., Ronning, N., Zhang, X., Muehlbauer, G.J., Timmermans, M.C.P., Schnable, P.S., Scanlon, M.J., and Yu, J. (2020). Genomic prediction of maize microphenotypes provides insights for optimizing selection and mining diversity. Plant Biotechnol. J. **18**:2456–2465. https://doi.org/10.1111/pbi.13420.

Yuan, Y., Cairns, J.E., Babu, R., Gowda, M., Makumbi, D., Magorokosho, C., Zhang, A., Liu, Y., Wang, N., Hao, Z., et al. (2019). Genome-Wide Association Mapping and Genomic Prediction Analyses Reveal the Genetic Architecture of Grain Yield and Flowering Time Under Drought and Heat Stress Conditions in Maize. Front. Plant Sci. **9**, 1919. https://doi.org/10.3389/fpls.2018.01919.

Zakieh, M., Alemu, A., Henriksson, T., Pareek, N., Singh, P.K., and Chawade, A. (2023). Exploring GWAS and genomic prediction to improve Septoria tritici blotch resistance in wheat. Sci. Rep. **13**, 15651. https://doi.org/10.1038/s41598-023-42856-x.

Zenke-Philippi, C., Thiemann, A., Seifert, F., Schrag, T., Melchinger, A.E., Scholten, S., and Frisch, M. (2016). Prediction of hybrid performance in maize with a ridge regression model employed to DNA markers and mRNA transcription profiles. BMC Genom. **17**, 262. https://doi.org/10.1186/s12864-016-2580-y.

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., Cui, Z., Ruan, Y., Burgueño, J., San Vicente, F., et al. (2017a). Effect of Trait Heritability, Training Population Size and Marker Density on Genomic Prediction Accuracy Estimation in 22 bi-parental Tropical Maize Populations. Front. Plant Sci. **8**, 1916. https://doi.org/10.3389/fpls.2017.01916.

Zhang, A., Pérez-Rodríguez, P., San Vicente, F., Palacios-Rojas, N., Dhliwayo, T., Liu, Y., Cui, Z., Guan, Y., Wang, H., Zheng, H., et al. (2022). Genomic prediction of the performance of hybrids and the combining abilities for line by tester trials in maize. The Crop Journal **10**:109–116. https://doi.org/10.1016/j.cj.2021.04.007.

Zhang, H., Yin, L., Wang, M., Yuan, X., and Liu, X. (2019). Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. Front. Genet. **10**, 189. https://doi.org/10.3389/fgene.2019.00189.

Zhang, J., Naik, H.S., Assefa, T., Sarkar, S., Reddy, R.V.C., Singh, A., Ganapathysubramanian, B., and Singh, A.K. (2017b). Computer vision and machine learning for robust phenotyping in genome-wide studies. Sci. Rep. **7**, 44048. https://doi.org/10.1038/srep44048.

Zhang, X., Pérez-Rodríguez, P., Burgueño, J., Olsen, M., Buckler, E., Atlin, G., Prasanna, B.M., Vargas, M., San Vicente, F., and Crossa, J. (2017c). Rapid Cycling Genomic Selection in a Multiparental Tropical Maize Population. G3 (Bethesda). **7**:2315–2326. https://doi.org/10.1534/g3.117.043141.

Zhang, X., Pérez-Rodríguez, P., Semagn, K., Beyene, Y., Babu, R., López-Cruz, M.A., San Vicente, F., Olsen, M., Buckler, E., Jannink, J.L., et al. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity **114**:291–299. https://doi.org/10.1038/hdy.2014.99.

Zhao, X., Nie, G., Yao, Y., Ji, Z., Gao, J., Wang, X., and Jiang, Y. (2020). Natural variation and genomic prediction of growth, physiological traits, and nitrogen-use efficiency in perennial ryegrass under low-nitrogen stress. J. Exp. Bot. **71**:6670–6683. https://doi.org/10.1093/jxb/eraa388.

Zhao, Y., Zeng, J., Fernando, R., and Reif, J.C. (2013). Genomic Prediction of Hybrid Wheat Performance. Crop Sci. **53**:802–810. https://doi.org/10.2135/cropsci2012.08.0463.

Zhao, Y., Li, Z., Liu, G., Jiang, Y., Maurer, H.P., Würschum, T., Mock, H.-P., Matros, A., Ebmeyer, E., Schachschneider, R., et al. (2015). Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc. Natl. Acad. Sci. USA **112**:15624–15629. https://doi.org/10.1073/pnas.1514547112.

Zhao, Y., Thorwarth, P., Jiang, Y., Philipp, N., Schulthess, A.W., Gils, M., Boeven, P.H.G., Longin, C.F.H., Schacht, J., Ebmeyer, E., et al. (2021). Unlocking big data doubled the accuracy in predicting the grain yield in hybrid wheat. Sci. Adv. **7**, eabf9106. https://doi.org/10.1126/sciadv.abf9106.

Zhong, S., Dekkers, J.C.M., Fernando, R.L., and Jannink, J.-L. (2009). Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics **182**:355–364. https://doi.org/10.1534/genetics.108.098277.

Zhu, C., Gore, M., Buckler, E.S., and Yu, J. (2008). Status and Prospects of Association Mapping in Plants. Plant Genome **1**:5–20. https://doi.org/10.3835/plantgenome2008.02.0089.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. J. Am. Stat. Assoc. **101**:1418–1429. https://doi.org/10.1198/016214506000000735.

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. J. Roy. Stat. Soc. B Stat. Methodol. **67**:301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Zou, H., Cobb, J.N., Juma, R.U., Biswas, P.S., Arbalaez, J.D., Rutkoski, J., Atlin, G., Hagen, T., Quinn, M., and Ng, E.H. (2009). On the adaptive elastic-net with a diverging number of parameters. Ann. Stat. **37**:1733–1751.

Molecular Plant, Volume *17*

Supplemental information

Genomic selection in plant breeding: Key factors shaping two decades of progress

Admas Alemu, Johanna Åstrand, Osval A. Montesinos-López, Julio Isidro y Sánchez, Javier Fernández-Gónzalez, Wuletaw Tadesse, Ramesh R. Vetukuri, Anders S. Carlsson, Alf Ceplitis, José Crossa, Rodomiro Ortiz, and Aakash Chawade

Genomic selection in plant breeding: key factors shaping two decades of progress

Admas Alemu[1*], Johanna Åstrand[1,2], Osval A Montesinos-López[3], Julio Isidro y Sánchez[4], Javier Fernández-Gónzalez[4], Wuletaw Tadesse[5], Ramesh R. Vetukuri[1], Anders S. Carlsson[1], Alf Ceplitis[2], José Crossa[6], Rodomiro Ortiz[1*], Aakash Chawade[1]

[1] Department of Plant Breeding, Swedish University of Agricultural Sciences, Alnarp, Sweden
[2] Lantmännen Lantbruk, Svalöv, Sweden
[3] Facultad de Telemática, Univ. de Colima, Colima, Colima 28040, México
[4] Centro de Biotecnología y Genómica de Plantas (CBGP, UPM-INIA), Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Campus de Montegancedo-UPM, 28223, Madrid, Spain
[5] International Center for Agricultural Research in the Dry Areas (ICARDA), Rabat, Morocco
[6] International Maize and Wheat Improvement Center (CIMMYT), Km 45, Carretera México-Veracruz, Texcoco 52640, México, Mexico

[*]Corresponding authors: admas.alemu.abebe@slu.se and rodomiro.ortiz@slu.se

Supplementary table 1. Genomic prediction research for yield, yield-related traits, and plant resistance to pathogens in wheat. SE, single-environment; ME, multi-environment; CV, cross-validation; YLD, grain yield; DH, days to heading; DM, days to maturity; TKW, thousand kernel weight; GRSP, grain number per spike; GFP, grain filling period; PH, plant height; SPL, spike length; LS, Leaf spot; TS, Tan spot; SB, Spot bloch; SNB, Septoria nodorum blotch; FHB, Fusarium head blight; STB, Septoria tritici blotch; PM, Powdery mildew; YR, Yellow rust; SR, Stem rust.

Plant materials	Field experiment	Training population (TRS)	Breeding/validation population (BS)	Cross-validation methods	Statistical models	No. of SNP markers	Traits	Prediction accuracy	Reference
Elite and early-generation breeding lines	Field trials at a single location for 3 years	1,334 elite breeding lines	1924 F$_2$ individuals developed from 21 parents selected from TRS	Within and across tested trials using the two populations	G-BLUP RR-BLUP RKHS	29,999	YLD	0.2 – 0.3	Bonnett *et al.* (2022)
Advanced wheat lines	Field experiments in 5 trials	Percentage composition varied depending on scenarios from a total of 803 lines	Percentage composition varied depending on scenarios	CV1, CV2 and leave-one-location-out with 7 different models	G-BLUP	Not specified	YLD	0.2 – 0.6	Saint Pierre et al. (2016)
Advanced wheat lines	Field trials at two locations for 2 years	Percentage composition of applied 141 genotypes, varied depending on scenarios	Percentage composition of applied 141 genotypes, varied depending on scenarios	Five-fold CV (SE), untested lines in tested environments (ME_CV1), tested lines in untested environments (ME_CV2)	G-BLUP	14,563	YLD DH DM TKW	0.08 – 0.81 0.02 – 0.83 0.02 – 0.83 0.32 – 0.88	Tomar et al. (2021b)
Breeding lines from preliminary to advanced yield trials	36 yield trials at one location for 7 years	Four fold of varied number of breeding lines across three yield trial stages	One fold of varied number of breeding lines across three yield trial stages	Five-fold CV within stage, across-stage and across-environment scenarios	G-BLUP	5,399 – 11,982	YLD	0.31 – 0.56	Juliana et al. (2020)
Varieties and advanced lines	6 field trials at 3 locations for two years	Four fold of 192 varieties and advanced lines	One fold of 192 varieties and advanced lines	Within experiment five-fold CV	RR-BLUP	10,577	YLD TKW GRSP DH DM GFP	0.35 – 0.62 0.38 – 0.58 0.23 – 0.4 0.24 – 0.47 0.29 – 0.56 0.28 – 0.42	Alemu et al. (2021a)
Preliminary and advanced yield trials (F$_{3:6}$ and F$_{3:7}$)	Several field trials at 6–10 locations for 5 years	The number varied from 570 to 1,072 across 9 scenarios	Varied from 28 – 560 across scenarios	Various CVs for within and across-environment	G-BLUP	27,000	YLD	0.17 – 0.52	Belamkar et al. (2018)
F$_{2:4}$ lines developed from elite breeders' germplasm	Four field trials from three environments tested in two years	2,563 – 2,787 lines	205 – 429 lines	Leave-one-cross-out and cross-dependent and random ten-fold CVs	G-BLUP	24,498	YLD	0.13 – 0.54	Edwards *et al.* (2019)

Supplementary table 1. cont…

Plant materials	Field experiment	Training population (TRS)	Breeding/validation population (BS)	Cross-validation methods	Statistical models	No. of SNP/other markers	Traits	Prediction accuracy	Reference
Double haploids, F_5 and advanced lines	Three distinct datasets from field trials from two environments tested in two years	The four fold of 501 F5s, 759 DHs from two pops and 447 lines	The one fold of 501 F5s, 759 DHs from two pops and 447 lines	Five-fold cross-validation within the dataset	12 different machine learning and Bayesian models	11,089	YLD	0.2 – 0.72	Sirsat et al. (2022)
Diverse wheat accessions	Six field trials collected from three locations tested for three seasons	133 (four fold) wheat accessions	33 (one fold) genotypes	Five-fold	8 BLUPs, Bayesian and machine learning-based models	11,997 (initial) Further filtered with 0, 20, 40, 60 and 80% missing and 0, 5 and 10% MAF values	YLD TKW DH SPL PH	0.25 – 0.59 0.6 – 0.65 0.16 – 0.41 0.15 – 0.42 0.37 – 0.62	Ali et al. (2020)
Multi-parent advanced generation inter-cross (MAGIC) population	Two field trials from two locations tested in two years	90% of the total 504 MAGIC population	10% of the 504 MAGIC population	Multi- and single-trait based ten-fold methods	LASSO and RF	55,000	YLD	0.2 – 0.38	Fradgley et al. (2023)
Single-cross hybrids (6,675) and inbred lines (6,283)	Varied number of multi-environmental tests across the six experimental series comprising different sizes of elite lines and hybrid progenies	Highly varied across experimental series	Various size of populations depending on the CVs and scenarios	Within experiment, five-fold and chessboard CV methods	G-BLUP	10,522	YLD	0.04 – 0.31 (hybrid) 0.08 – 0.48 (inbred)	Zhao *et al.* (2021)
Six panels of lines from CIMMYT's global bread wheat breeding programme	Several field trials from 11 locations tested for several years	613 – 2,719 lines depending on several applied scenarios	153 – 980 lines depending on the scenarios	Five-fold	G-BLUP and BayesB	78,606 (initial) Further filtered with different levels of missing values, LD and their combinations	YLD	0.15 – 0.5	Juliana et al. (2019b)
Elite wheat lines	12 field trials tested for two years	90% of the 306 elite lines after generating 50 random partitions of the datasets	The remaining 10%	Ten-fold	Seven models including parametric, semi-parametric and non-parametric methods.	1,717 DArT markers	YLD DH	0.02 – 0.69	Pérez-Rodríguez *et al.* (2012)
Advanced breeding lines	In a single environment with two water supply levels tested for two years	86% of the 384 wheat lines	The remaining 14% of the total population	Seven-fold cross-validation	RR-BLUP and Gaussian kernel	102,324	YLD	0.38 – 0.63	Lado et al. (2013)
Advanced breeding lines	In three environments tested for four years	Three crossing sets from the total four block crosses comprising 1,325 advanced lines	One crossing set from a total of four blocks comprising 1,325 advanced lines	Block-cross validation using crossing set as blocks	RR-BLUP and Bayesian Power Lasso	9,290	YLD	0.21 – 0.31	Tsai et al. (2020)

Supplementary table 1. cont…

Plant materials	Field experiment	Training population (TRS)	Breeding/validation population (BS)	Cross-validation methods	Statistical models	No. of SNP/other markers	Traits	Prediction accuracy	Reference
Advanced breeding lines	Field trial in a single environment for five years under different drought and heat stress levels	90% of the total population (4,302)	10% of the total population (4,302)	Ten-fold	G-BLUP	8,443 SNPs and 501 SNP-haplotypes	YLD	0.28 – 0.58	Sehgal et al. (2020)
Breeding lines	Field trials in a single environment tested for two years under different abiotic stresses	Four fold from the total (630) number of breeding lines	One fold from the total (630) number of breeding lines	Five-fold	G-BLUP and RR-BLUP	12,083	YLD	0.17 – 0.62	Sun et al. (2017)
Breeding lines	Field trial at two locations for two years under different drought levels	Four fold from the total (237) number of breeding lines	One fold from the total (237) number of breeding lines	Five-fold	Five models including G-BLUP, Bayesian and deep learning methods	27,957	YLD TKW	-0.23 – 0.59 0.2 – 0.88	Guo et al. (2020a)
Advanced and elite breeding lines	Field trial at five locations for two years	80% of a total of 314 lines	20% of a total of 314 lines	Five-fold	RR-BLUP and Bayesian Multi Trait Gaussian model	10,290	YLD PH DH	0.03 – 0.71 0.16 – 0.54 0.16 – 0.58	Gill et al. (2021)
Advanced elite lines	Field trials at 9 locations for two years	80% of a total of 287 lines	20% of a total of 287 lines	Multi-environment-based CV1 and CV2	Bayesian Multi Trait Gaussian model	15,000	YLD TKW	-0.05 – 0.54 -0.03 – 0.88	Sukumaran et al. (2017)
Elite lines	Field trial at two locations for two years	165 (70%) from a total of 236 lines	71 (30%) from a total of 236 lines	CV1 and CV2	BRR	27,466	YLD	0.17 – 0.50	Shahi et al. (2022)
Hybrids developed from 22 female and 13 male wheat lines	Field trial at four locations for one year	The size varied depending on the CV method	The size varied depending on the CV method	$CV_{\text{five-fold}}$ $CV_{\text{unrelated}}$ CV_{male} CV_{female}	RR-BLUP and four Bayesian models	1,201	YLD	0.28 – 0.63	Zhao et al. (2013)
681 and 701 RILs from 5 populations during off-season and main season, respectively	Two seasons in a single environment for natural infection (SR) and a single controlled environment for artificial infection (YR)	Several sets of training population sizes across different CV scenarios	Several sets of test population sizes across different CV scenarios	Tenfold CV within and across populations and environments	RR-BLUP BLASSO SVM	1,400 DArT markers	SR YR	0.26 – 0.85 -0.09 – 0.63	Ornella et al. (2012)
Wheat landraces	Under natural infection evaluated for 4 years at a single location	90% of a total of 206 wheat landraces	10% of a total of 206 wheat landraces	Five-fold	G-BLUP BayesR	5,568 and major genes	SR YR LR	0.27 – 0.38 0.3 – 0.46 0.33 – 0.48	Daetwyler et al. (2014)
Cultivars and breeding lines	Natural infection evaluated for two years at four locations	90% of sets of populations with 230, 1,065, 1,001 and 175 individuals	10% of sets of populations with 230, 1,065, 1,001 and 175 individuals	Five-fold	OLS G-BLUP	6,860	SR	-0.01 – 0.64	Shahinnia et al. (2022)

Supplementary table 1. cont…

Plant materials	Field experiment	Training population (TRS)	Breeding/validation population (BS)	Cross-validation methods	Statistical models	No. of SNP markers	Traits	Prediction accuracy	Reference
Cultivars and RILs	Natural infection for several years and at several locations differ in diseases	80% of a total of 578 for CV1 and CV2 and varied size for CV0 methods	20% of a total of 578 for CV1 and CV2 and varied size for CV0 methods	Multi-environment-based C1, C2 and CV0 methods	G-BLUP	3158, 5732, and 23,795 across populations	Bunt FHB LR YR LS	0.02 – 0.87 0.41 – 0.62 0.11 – 0.77 0.25 – 0.77 0.04 – 0.75	Semagn et al. (2022b)
Advanced breeding lines	Under artificial infection in a single controlled environment	80% of a total of 316 lines	20% of a total of 316 lines	Five-fold	RR-BLUP	10,120	STB	0.49 – 0.58	Zakieh et al. (2023)
Advanced breeding lines	Natural infection for three consecutive years in a single environment	80% of a total of 141 advanced breeding lines	20% of a total of 141 advanced breeding lines	Five-fold	G-BLUP RR-BLUP RKHS RF	14,563	YR	0.59 – 0.63	Tomar et al. (2021a)
Advanced breeding lines and old cultivars	From a single controlled experiment with artificial infection	80% of 272 advanced lines and 147 old cultivars for pooled population or either of this number for across-population scenarios	20% of 272 advanced lines and 147 old cultivars for pooled population or either of this number for across-population scenarios	Five-fold cross-validation for pooled or within-population	RR-BLUP and 6 other Bayesian models	6421 (initial) Further reduced set of SNPs	FHB	0.45 – 0.52	Alemu *et al.* (2023)
Wheat cultivars, modern varieties, double haploids and RILs	Field evaluation under natural infection for 3 – 8 nurseries with the 6 populations	80% of a total of 1,104 individuals from 6 populations for CV1 and CV2 and varied size for CV0 methods	20% of a total of 1,104 individuals from 6 populations for CV1 and CV2 and varied size for CV0 methods	Multi-environment-based C1, C2 and CV0 methods	G-BLUP	The number varied from 1,058 to 23,795 across the 6 populations	YR	0.19 – 0.84	Semagn et al. (2022a)
Synthetic hexaploid wheats	In a greenhouse with artificial inoculation	70% of a total of 400 lines	70% of a total of 400 lines	Fifty-fold	G-BLUP	6548	TS SB SNB	0.39 – 0.67 0.27 – 0.45 0.4 – 0.55	García-Barrios et al. (2023)
Landraces and old cultivars	Natural infection under field conditions for two years and at four locations	80% of a total of 175 genotypes within environment	20% of a total of 175 genotypes within environment	Five-fold within environment method	RR-BLUP	7,401	STB PM	0.15 – 0.66 0.18 – 0.83	(Alemu et al., 2021b)

Supplementary table 2. Genomic prediction research for yield, yield-related traits and plant resistance to pathogens in maize. AD, anthesis date; ASI, anthesis-silking interval; SEN, senescence; MLN, Maize lethal necrosis; TRL; total root length; SEL, secondary root length; PRL, primary root length; NCLB, Northern corn leaf blight; FER, Fusarium ear rot; MCMV, Maize chlorotic mottle virus; GER, Gibberella ear rot; DON, Deoxynivalenol concentration; Zn, Zinc content; FUM, Fumonisin; OC, Oil content; YLD, grain yield; PH, plant height; GMC, grain moisture content; EH, ear height; EW, ear weight.

Plant materials	**Field experiment**	**Training population (TRS)**	**Breeding/validation population (BS)**	**Cross-validation methods**	**Statistical models**	**No. of SNP markers**	**Traits**	**Prediction accuracy**	**Reference**
DH and inbred lines	Field evaluation in four severe drought stress and five well-watered trials	70% of a total population (504 DH lines and 296 inbred lines)	30% of a total population (504 DH lines and 296 inbred lines)	Hold-out cross-validation	G-BLUP RKHS	158,281 (Exp-I) 235,265 (Exp-II)	YLD	0.37 – 0.61	Crossa et al. (2013)
Diverse breeding lines	Field evaluation in a total of 156 trials	The size of the training population varied greatly across scenarios and the two CV methods	The size of the test population varied across scenarios and the two CV methods	Multi-environment CV1 and CV2 methods	G-BLUP	66,000	YLD	-0.03 – 0.36	Edriss et al. (2017)
Inbred progenies comprising RILs and DH lines	Field trials in five environments (location – year combination)	Different set of population sizes across scenarios	Different set of population sizes across scenarios	Leave-one-out cross-validation (LOOCV) with five different scenarios	G-BLUP	2,296	YLD	0.28 – 0.77	Kadam *et al.* (2016)
Diverse inbred lines and F_2 progenies	Field trials at six (experiment 1) and four (experiment 2) locations for two years	Size of training population varied across environments	Size of test population varied across environments	Five-fold	RR-BLUP	18,695	YLD AD ASI	-0.2 – 0.54 -0.03 – 0.54 -0.4 – 0.57	Windhausen *et al.* (2012)
Diverse tropical/subtropical inbred lines	Artificial infection under field conditions for 3 seasons in a single location	80% of a total of 615 lines	20% of a total of 615 lines	Five-fold	RR-BLUP	259,000 (Exp-I) 264,000 (Exp-II)	MLN	0.41 – 0.56	Gowda et al. (2015)
Inbred lines	Field trials in 43 environments (location-year combination)	The size varied across scenarios	The size varied across scenarios	Multi-environment-based CV1, CV2, CV0 and C00 methods	G-BLUP	477,845	YLD	0.1 – 0.53	Jarquin et al. (2021)
Inbred lines	Nine carotenoid compounds measured from grain samples of 252 maize lines	80% of a total of 252 lines	20% of a total of 252 lines	Five-fold	RR-BLUP LASSO EN	284,187 SNPs and other subsets of SNPs	24 carotenoid-related traits	0.21 – 0.71	Owens et al. (2014)
Inbred lines	Field evaluation at 9 optimum and 13 low-nitrogen stressed locations	Size varied across management and populations	Size varied across management and populations	Five-fold	RR-BLUP	5,929	YLD AD ASI SEN	0.2 – 0.42 0.62 – 0.65 0.59 – 0.71 0.06 – 0.52	Ertiro et al. (2020)
Inbred lines	Field evaluation with six trait-environment combinations	384 inbred lines	2,431 inbred lines	Hold-out with 60/40% training test populations split	RR-BLUP	186,849	TRL SEL PRL	0.21 – 0.54 0.32 – 0.43 0.31 – 0.44	Pace et al. (2015)

Supplementary table 2. cont….

Plant materials	Field experiment	Training population (TRS)	Breeding/validation population (BS)	Cross-validation methods	Statistical models	No. of SNP markers	Traits	Prediction accuracy	Reference
Inbred lines (collaborative panel)	Field trials at seven locations for one year	386 lines	338 lines	Five-fold within training population	BRR	40,478	YLD	0.4 – 0.55	Allier et al. (2020)
Tropical maize inbred lines	Field evaluation with artificial inoculation for three years in a single location	Different size across trials and scenarios	Different size across trials and scenarios	Five-fold	RR-BLUP	4,000	MLN MCMV	0.52 – 0.87 0.21 – 0.78	Sitonik et al. (2019)
DH lines	Varied number of trials across three years to evaluate subsets of DH lines for yield	Training population size varied in different applied scenarios	Test population size varied in different applied scenarios	Five-fold	RKHS	6,137	YLD	0.23 – 0.56	Wang et al. (2020)
Elite maize breeding lines	Field evaluation with artificial inoculation for two years at two locations	Size varied from 32 to 184 across prediction scenarios	22 lines	Five-fold	G-BLUP BayesB	23,797	GER DON	-0.17 – 0.6 -0.06 – 0.66	Han et al. (2018)
Tropical maize germplasm	Field trials for 3 years at four locations for grain sample to analyse ZN content	Size varied across scenarios and panels	Size varied across scenarios and panels	Multi-environment-based five-fold CV1 and CV2 methods	EG-BLUP	170,798 – 181,889	Zn	-0.05 – 0.72	Mageto *et al.* (2020)
Tropical maize germplasm	Field trials in three environments (location-year combinations)	80% of a total of 300, 108 and 143 lines from the three panels	20% of a total of 300, 108 and 143 lines from the three panels	Five-fold	RR-BLUP	2,795 – 6,150	Zn	-0.13 – 0.65	Guo et al. (2020b)
$S_{0:1}$ lines	Field evaluation with artificial inoculation for two years at three locations	Size varied across panels, years and scenarios	Size varied across panels and scenarios	Five-fold	G-BLUP Bayes Cπ BLASSO XgBoost	6,086	FER FUM	0.24 – 0.46 0.39 – 0.67	Holland et al. (2020)
Diverse inbred lines	Field trials in four environments (location-year combinations) to evaluate kernel oil content	351	134	Five-fold within training population	RR-BLUP BayesA BLASSO BayesC RKHS	44,624	OC	0.4 – 0.68	Hao et al. (2019)
Hybrids	Field evaluation for four years in a single location	Size varied according to CV scenarios	Size varied according to CV scenarios	Five-fold within environments and multi-environment-based CV1 and CV2 methods	G-BLUP	Not mentioned	YLD	-0.05 – 0.7	De Oliveira et al. (2020)
Inbred and hybrid datasets	Field evaluation in five locations for one year	75% of the 452 hybrid and 128 inbred line datasets	25% of the 452 hybrid and 128 inbred datasets	75/25 training-test test split	G-BLUP	52,700	YLD	0.3 – 0.77	Lyra et al. (2018)
Tropical and subtropical inbred lines	Field evaluation in multiple locations four years (15 environments)	80% of a total of 300 lines	20% of a total of 300 lines	Five-fold	RR-BLUP	10,108	YLD ASI AD	0.04 – 0.77 -0.2 – 0.77 0.20 – 0.82	Yuan et al. (2019)

Supplementary table 2. cont…

Plant materials	Field experiment	Training population (TRS)	Breeding/validation population (BS)	Cross-validation methods	Statistical models	No. of SNP markers	Traits	Prediction accuracy	Reference
DH lines derived from flint landraces and elite lines	Field evaluation across four agro-ecologically diverse locations for one year	Size varied across scenarios	Size varied across scenarios	Different methods including LOOCV	G-BLUP	32,492	YLD OC	-0.53 – 0.53 -0.24 – 0.71	Brauner et al. (2018)
Inbred lines	Field evaluation under natural infection at two locations for two years	Ranging from 10 to 90% of a total of 509 lines	Ranging from 10 to 90% of a total of 509 lines	Five-fold	G-BLUP BayesA BayesB BayesC	37,801	FER	0.2 – 0.43	Guo et al. (2020c)
Breeding lines	Field evaluation for three years at two locations	Twenty-four single cross-combinations	Four single cross-combinations	75/25% training-test sets split within the training population	RKHS	328,127	YLD	0.49 – 0.61	Vélez-Torres et al. (2018)
Lines from the Germplasm Enhancement of Maize (GEM) project	Multi-location and year trial with 45 and 31 environments (year-location) from two GEM programmes	Varied from 182 – 1,491 lines grouped into 6 datasets	The size varied from 589 to 2,080 lines	Across population	G-BLUP	40,000	YLD	0.36 – 0.75	Rogers et al. (2022)
Inbred lines	Field evaluation for two years in a single location	80% of a total of 149 inbred lines	20% of a total of 149 inbred lines	Five-fold	G-BLUP BayesB LASSO EN PLS RKHS XgBoost RF	102,654, 41,855, 11,255 and 1,319 from GBS, 40K, 10K and 1K SNP array platforms, respectively	GY PH EH EW	0.08 – 0.71	Technow et al. (2013)
DH lines derived from flint landraces and elite lines	Field evaluation across four agro-ecologically diverse locations for one year	Size varied across scenarios	Size varied across scenarios	Different methods including LOOCV	G-BLUP	32,492	YLD OC	0.29 – 0.48 0.57 – 0.73 0.51 – 0.66 0.47 – 0.62	Yu et al. (2023)

Supplementary table 3. Genomic prediction research for tuber yield, size, quality and plant resistance to pathogens in potato. FC, fry color; CS, crisp score; PI, Phytophthora infestans infection; LB, Potato late blight; EB, Potato early blight; CSc, Common scab; PM, days to plant maturity; TSC, tuber starch content; TSY, tuber starch yield; TYLD, tuber yield; SC, specific gravity; TS, tuber shape; TL, tuber length; FCo, tuber flesh color; DMC, tuber dry matter content; TN, tuber number; NTS, number of tuber stem; TRS, tuber reducing sugar; T40, tubers size below 40 mm; T40-50, tubers size between 40 – 50 mm; T50-60, tubers size between 50 – 60 mm; T60, tubers size above 60 mm.

Plant materials	Field experiment	Training population (TRS)	Breeding/validation population (BS)	Cross-validation methods	Statistical models	No. of SNP markers	Traits	Prediction accuracy	Reference
Diverse tetraploid potato comprising mainly advanced breeding clones	Field evaluation for three years in a single location	Sample size of training set varied	Sample size varied	Five-fold	G-BLUP BayesA BayesCπ BLASSO		PI PM TSC TSY TYLD	0.66 – 0.86 0.63 – 0.71 0.63 – 0.86 0.29 – 0.5 0.43 – 0.55	Stich and Van Inghelandt (2018)
Potato lines	Field evaluation in a single location for three years	Varied from 45 – 219 across three years	56 lines	Not applied	RR-BLUP BayesA BLASSO RF	46,406	FC	0.11 – 0.77	Byrne et al. (2020)
Worldwide cultivars	Field evaluation in three locations for two years	105	42	Not applied	G-BLUP RKHS BLASSO BayesA BAYES Cπ	39,000	TN TYLD TL DM	0.31 – 0.42 0.55 – 0.59 0.55 – 0.57 0.73 – 0.77	Wilson et al. (2021)
Breeding clones and released cultivars	Field evaluation of 256 lines in three locations for two years	Size varied according to the cross-validation scenarios	Size varied based on the cross-validation scenarios	Five-fold and leave-one-environment out methods	PLS	2,503	TYLD TSC TRS T40 T40-50 T50-60 T60	0.55 – 0.80 0.48 – 0.94 − 0.06 – 0.79 0.60 – 0.8 0.25 – 0.67 -0.14 – 0.69 0.5 – 0.082	Ortiz et al. (2023b)
Biparental progenies and breeding clones	Field evaluation in a single location and environment plus unbalanced phenotypic data from 1997 - 2014	The size varied depending the traits and CV methods	The size varied depending the traits and CV methods	Eight-fold and leave-sibs-out methods	G-BLUP BayesA BayesC	171,859	TSC FC	0.68 – 0.73 0.47 – 0.56	Sverrisdóttir et al. (2017)
Breeding clones and released cultivars	Field evaluation in three locations for one year	Varied according to validation scenarios	Varied according to validation scenarios	Several schemes	Reaction norm with GxE	>2,000	TYLD TSC TRS T40 T40-50 T50-60 T60	0.40 – 0.80 0.48 – 0.88 0.47 – 0.58 0.48 – 0.71 0.24 – 0.62 -0.12 – 0.73 0.53 – 0.82	Cuevas et al. (2022)

Supplementary table 3. cont…

Plant materials	Field experiment	Training population (TRS)	Breeding/validation population (BS)	Cross-validation methods	Statistical models	No. of SNP markers	Traits	Prediction accuracy	Reference
Breeding clones	Field evaluation for upto six years in a single location	Varied across populations and traits	Varied across populations and traits	Different methods	G-BLUP	Varied across populations	TYLD FC	0.53 – 0.55 0.43 – 0.46	Endelman et al. (2018)
Europe-wide potato cultivars	Field evaluation for two years in a single location	90% of the total 190 cultivars	10% of the total 190 cultivars	10-fold	G-BLUP BL RKHS	50,107 SilicoDArT markers	TYLD DM TN NTS	0.37 – 41 0.54 – 0.68 0.13 – 0.16 0.01 – 0.03	Habyarimana et al. (2017)
Breeding clones	Field evaluation for two years in a single location	80% of the total 762 lines	20% of the total 762 lines	5-fold for within training population	G-BLUP	167,637	DM FC	0.75 – 0.83 0.39 – 0.79	Sverrisdóttir *et al.* (2018)
Inbred and hybrid progenies	Field evaluation in single location for one year	70% of the studied populations	30% of the studied populations	10-fold	G-BLUP	2,000	TYLD TRS T40-50 T50-60 T60	-0.22 – 0.31 -0.28 – 0.38 -0.12 – 0.35 -0.26 – 0.29 -0.39 – 0.20	Ortiz et al. (2023a)
Early generation and advanced tetraploid potato genotypes	Field evaluation with artificial inoculation for seven (late blight) and nine (common scab) years in one location	Varied depending the CV scenarios	Varied depending the CV scenarios	Multi-environment based 5-fold and one-year-out methods	G-BLUP BRR BayesB	8303	LB CSc	0.24 – 0.31 0.22 – 0.29	Enciso-Rodriguez et al. (2018)
Breeding clones	Field evaluation in a single location and year	N-1 through LOOCV (N=92)	A single individual according to LOOCV method	Leave-one-out (LOOCV)	BRR BayesA, BayesB BayesC BLASSO	9,180	LB TN	0.24 0.20	Selga et al. (2020)
Diverse potato accessions	Field evaluation with natural inoculation for three years in a single location	Training population size varied across the three years (107-152)	Test population size varied across the three years (41-73)	5-fold	G-BLUP	1,20,622	LB	-0.15 – 0.68	Sood et al. (2023)
tetraploid potato clones	Field evaluation in three locations for one year	Size varied depending on scenarios	Size varied depending on scenarios	5-fold for within population and across population methods	BRR	10,546	TYLD TN LB	0.05 – 0.75 0.05 – 0.72 0.16 – 0.29	Selga et al. (2021)
Chipping potato clones	Field evaluation for three years in one location	Different size with varied scenarios	Different size	Across populations	G-BLUP	14,401	FC TYLD	0.77 0.24	Pandey et al. (2023)
Potato hybrids	Field evaluation in four locations for two years	Differ according to prediction scenarios	Differ according to prediction scenarios	Not applied	G-BLUP	704	TYLD TN DM	0.46 – 0.48 0.36 – 0.51 0.49 – 0.58	Adams et al. (2023)

Supplementary table 3. cont…

Plant materials	Field experiment	Training population (TRS)	Breeding/validation population (BS)	Cross-validation methods	Statistical models	No. of SNP markers	Traits	Prediction accuracy	Reference
Potato cultivars	Field evaluation for three years in a single location	117 cultivars	50 cultivars	117/50 training-test split	G-BLUP	180,550	DM	0.25	Sood et al. (2020)
							EB	0.14	
							TS	0.28	
							CS	0.52	
							FCo	0.77	
Potato cultivars	Field evaluation for 6 years in a single location	80% of the total 169 cultivars	20% of the total 169 cultivars	5-fold	BayesA	183,848	FCo	0.80 – 0.81	Caruana et al. (2019)
					BayesB		DM	0.23 – 0.29	
							CS	0.37 – 0.46	
Breeding clones and cultivars	Field evaluation at three locations for one year	Varied depending on CV scenarios	Varied depending on CV scenarios	70/30 training-test split and multi-environment-based CV2 method	G-BLUP	2000	TYLD	0.31 – 0.59	Ortiz et al. (2022)
					GK		TSC	0.60 – 0.73	
							T40	0.38 – 0.58	
							T40-50	0.27 – 0.46	
							T50-60	0.27 – 0.53	
							T60	0.46 – 0.57	
							LB	0.61 – 0.63	
							TRS	0.32 – 0.39	

Supplementary table 4. Summary of the available training set optimization methods in genomic prediction. *CS*, candidate set. *TP*, target population, usually corresponding to either the *CS* in untargeted optimization or to the test set (*TS*) in targeted optimization. *nTP*, number of genotypes in the *TP*. *nTRS*, number of genotypes in the training set (*TRS*). *i*, a given individual belonging to the *TP*. *j*, a given individual belonging to the *TRS*. *sum*(·), sum of all elements in a vector or matrix. *diag*(·), returns the main diagonal of a matrix. *Tr*(·), trace of a matrix. | · |, absolute value of a scalar or determinant of a matrix. *min*(·), returns the smallest value of a vector. *max*(·), returns the largest value of a vector. *ln*(·), natural logarithm. *Matrix*$_{[set1,set2]}$, a subset of the *Matrix* corresponding to the rows in *set1* and columns in *set2* is taken. $arg\ \underset{input}{min}(objective)$ and $arg\ \underset{input}{max}(objective)$ refer to heuristic application to find the input that minimizes or maximizes the objective. *X,* design matrix linking the response variable to the fixed effects. In most cases the only fixed effect is the intercept (which results in *X* being a vector of ones), while in D-opt the fixed effects are the genotypic effects. *Z*, design matrix linking the response variable to the random genotypic effects. *G*, genomic relationship matrix. *I*, Identity matrix, assuming the dimensions compatible with the specified operations. *R*, residual variance-covariance matrix, usually the identity. *W*, marker matrix in numeric form with individuals in rows and markers in columns. The markers are often replaced by several of their primary principal components to accelerate computations. w_i , column vector for the markers of *TP* individual *i*. β, matrix of regression coefficients that allow to predict the genotypic values of *TP* individuals (matrix rows) as a linear combination of CS genotypic values (matrix columns). $\tilde{\beta}$, matrix of sparse regression coefficients. Ω_G, genetic covariance matrix between trials. Ω_{E}, error covariance matrix between trials. σ_ϵ^2, residual variance, usually assumed to be 1. σ_g^2, genotypic variance, usually assumed to be 1. $\lambda = \frac{\sigma_\epsilon^2}{\sigma_g^2} \cdot \lambda_1$, parameter controlling the degree of sparsity in SSI. Usually found by cross-validation within the CS, which would require CS phenotypic information. *a* and *b*, weighting parameters in Avg_GRM_MinMax, larger than 0, usually assumed to be 1. In the "observations" column, four different statements are made: a) type of the method, either parametric or non-parametric; b) suited for either targeted and/or untargeted scenarios; c) can be used to optimize either the TRS size, composition and/or distribution in the field; and d) other remarks. Notes: some of the parametric methods described here can also be implemented with contrasts (see supplementary file 1, Note 1). Parametric methods such as A-opt, CDmean and PEVmean could also be applied to optimize TRS distribution in the field. This requires dropping the assumption $\mathrm{R} = \mathrm{I}$, which would alter the equations as follows: $\mathrm{V} = \mathrm{ZGZ}' + \mathrm{R}$; $\mathrm{M} = \mathrm{V}^{-1} - \mathrm{V}^{-1}\mathrm{X}(\mathrm{X}'\mathrm{V}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{V}^{-1}$ (more details in **Supplementary file 1, Note 2.5**).

Method name(s)	Mechanism of action	Observations	References
PEVmean	$M = I - X(X'X)^{-1}X'$ $PEV = (Z'MZ + \lambda G^{-1})^{-1}\sigma_\epsilon^2$ $\arg\min_{Z}\left(\frac{\mathrm{Tr}(PEV_{[TP,TP]})}{nTP}\right)$	a) Parametric b) Targeted or untargeted c) Optimize composition	Rincent et al. (2012) Isidro et al. (2015) Neyhart et al. (2017) Mangin et al. (2019) Rio et al. (2021) Kadam et al. (2021)
PEV^{ridge}mean	$PEV^{ridge} = \left(W_{[TP,All]}\right)\left[\left(W_{[TRS,All]}\right)'\right.$ $\left.\left(W_{[TRS,All]}\right) + \lambda I\right]^{-}(W_{[TP,All]})'$ $\arg\min_{TRS}\left(\frac{\mathrm{Tr}(PEV^{ridge})}{nTP}\right)$	a) Parametric b) Targeted or untargeted c) Optimize composition	Akdemir et al. (2015) Akdemir and Isidro-Sánchez (2019) Sarinelli et al. (2019) Ou and Liao (2019) Guo et al. (2019) Heslot and Feoktistov (2020) Mendonça and Fritsche-Neto (2020) Montesinos-López and Montesinos-López (2023)
		d) Used in conjunction with clustering	Ou and Liao (2019)
CDmean	$M = I - X(X'X)^{-1}X'$ $CD = \frac{G - \lambda(Z'MZ + \lambda G^{-1})^{-1}}{G}$ $\arg\max_{Z}\left(\frac{\mathrm{Tr}(CD_{[TP,TP]})}{nTP}\right)$	a) Parametric b) Targeted or untargeted c) Optimize composition	Rincent et al. (2017) Isidro *et al.* (2015) Tayeh et al. (2015) Bustos-Korts et al. (2016) Neyhart *et al.* (2017) Ou and Liao (2019) Mangin *et al.* (2019) Olatoye et al. (2020) de Verdal et al. (2023) Rio *et al.* (2021) Lemeunier et al. (2022) Atanda et al. (2021a) Kadam *et al.* (2021)
		d) Approach not reliant on any heuristic. Faster but it can't maximize TRS diversity	Atanda et al. (2021b)
		d) Used in conjunction with clustering	Rincent *et al.* (2017) Ou and Liao (2019)
		d) Using additive + dominance kernel	Momen and Morota (2018)
		d) Extended formula derived from multi-trait model	Ben-Sadoun et al. (2020)
CDmean multi-environment	$GxE = \Omega_G \otimes G$ $R = \Omega_E \otimes I$ $V = \mathrm{Z\ GxE\ Z' + R}$ $M = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$ $CD = \frac{GxE\ Z'MZ\ GxE}{GxE}$ $\arg\max_{Z}\left(\frac{\mathrm{Tr}(CD_{[TP,TP]})}{nTP}\right)$	a) Parametric b) Targeted or untargeted c) Optimize composition and distribution. d) TP refers to genotype-trial combinations, not only genotypes.	Rio et al. (2022)
CDmin	$M = I - X(X'X)^{-1}X'CD = \frac{GZ'MZG}{G}$ $\arg\max_{Z}(\min(\mathrm{diag}(CD)))$	a) Parametric b) Targeted or untargeted c) Optimize composition and/or distribution.	Akdemir et al. (2021)
CDmean without projection	$Vinv = (G_{[TRS,TRS]} + \lambda I)^{-1}$ $CD = \left(\frac{G_{[All,TRS]}(Vinv - (VinvVinv))}{sum(VinvG_{[TRS,All]})}\right)/G$ $\arg\max_{TRS}\left(\frac{\mathrm{Tr}(CD_{[TP,TP]})}{nTP}\right)$	a) Parametric b) Targeted or untargeted c) Optimize composition Only suitable when the only fixed effect is the intercept. Has been used stand-alone or in conjunction with clustering	Fernández-González et al. (2023)
	Assuming that the best CDmean fitness value is obtained when the entire CS is used as TRS, find the TRS size at which an acceptable, user-defined, fitness reduction happens.	c) Optimize Size	Fernández-González et al. (2023)

Supplementary table 4. Cont…

Method name(s)	Mechanism of action	Observations	References
CD^{ridge}mean	$PEV^{ridge} = W_{[TP,All]}(W_{[TRS,All]}{}'$ $W_{[TRS,All]} + \lambda I)^{-}W_{[TP,All]}{}'$ $CD^{ridge} = \frac{PEV^{ridge}}{W_{[TP,All]}W_{[TP,All]}{}'}$ $\arg \min_{TRS}\left(\frac{\mathrm{Tr}(CD^{ridge})}{nTP}\right)$	a) Parametric b) Targeted or untargeted c) Optimize composition	Akdemir and Isidro-Sánchez (2019) Roth et al. (2020) Guo *et al.* (2019)
CDMEAN2 PCA_CDmean	$D_1 = \mathrm{diag}\left(W_{[TP,All]}W'_{[TP,All]}\right)$ $X_1 = W'_{[TRS,All]}W_{[TRS,All]}$ $X_2 = (X_1 + \lambda I)^{-1}$ $D_2 = \mathrm{diag}\left(W_{[TP,All]}X_2X_1X_2W'_{[TP,All]}\right)$ $CDMEAN2 = \mathrm{sum}\left(\frac{D_2}{D_1}\right)/nTP$ $\arg \min_{TRS}(CDMEAN2)$	a) Parametric b) Targeted or untargeted c) Optimize composition	Fernández-González et al. (2023) Fernández-González et al. (2024)
A-opt	$M = I - X(X'X)^{-1}X'$ $PEV = (Z'MZ + \lambda G^{-1})^{-1}\sigma_\epsilon^2$ $Aopt = \frac{2}{nTP-1}\left(\mathrm{Tr}(PEV_{[TP,TP]}) - \frac{1}{nTP}\mathrm{sum}(PEV_{[TP,TP]})\right)$ $\arg \min_{Z}(Aopt)$	a) Parametric b) Targeted or untargeted c) Optimize composition and distribution	Cullis et al. (2006) Butler et al. (2013) Cullis et al. (2020)
A-opt^{ridge}	$Aopt^{ridge} = Tr(\left(W'_{[TRS,All]}W_{[TRS,All]} + \lambda I\right)^{-1})$ $\arg \min_{TRS}(Aopt^{ridge})$	a) Parametric b) Targeted or untargeted c) Optimize composition	Akdemir and Isidro-Sánchez (2019)
D-opt	$Dopt = \lvert X'X \rvert$ $\arg \max_{X}(Dopt)$	a) Parametric b) Untargeted c) Optimize composition and distribution. d) Does not require genotypic information of the CS	Mitchell (2000) Edmondson (2020)
D-opt^{ridge}	$Dopt^{ridge} = -\log(\left\lvert\left(W'_{[TRS,All]}W_{[TRS,All]} + \lambda I\right)^{-1}\right\rvert)$ $\arg \min_{TRS}(Dopt^{ridge})$	a) Parametric b) Targeted or untargeted c) Optimize composition	Akdemir and Isidro-Sánchez (2019)
Rscore	$A = W'_{[TRS,All]}\left(W_{[TRS,All]}W'_{[TRS,All]} + \lambda I\right)^{-1}$ $IJ = I - I(1/nTP)$ $q_{12} = \mathrm{Tr}(W'_{[TP,All]}\ IJ\ W_{[TP,All]}AW_{[TRS,All]})$ $q_1 = (nTP - 1) + \mathrm{Tr}(W'_{[TP,All]}\ IJ\ W_{[TP,All]})$ $q_2 = \mathrm{Tr}\left(A'W'_{[TP,All]}\ IJ\ W_{[TP,All]}A\right) + \mathrm{Tr}(W_{[TRS,All]}A'\ IJ\ AW_{[TRSP,All]})$ $Rscore = q_{12}/\sqrt{q_1q_2}$	a) Parametric b) Targeted or untargeted c) Optimize uomposition	Ou and Liao (2019) Fernández-González et al. (2023) Wu et al. (2023)
		a) Used in conjunction with clustering	Ou and Liao (2019)
	Assuming that the best Rscore fitness value is obtained when the entire CS is used as TRS, find the TRS size at which an acceptable, user-defined, fitness reduction happens.	c) Optimize size	Fernández-González et al. (2023) Wu *et al.* (2023)
Upper bound of reliability	$Q = W'_{[TRS,All]}\left(W_{[TRS,All]}W'_{[TRS,All]}\right)^{-}W_{[TRS,All]}$ $UP_i = \frac{(Qw_i)'(Qw_i)}{w'_iw_i}$ $\arg \max_{TRS}(UP_i)$	a) Parametric b) Targeted c) Optimize composition	Karaman et al. (2016) Yu et al. (2020)

Supplementary table 4. Cont…

Method name(s)	Mechanism of action	Observations	References
EthAcc	1) Perform GWAS within the TRS to find causal QTL 2) Estimate theoretical accuracy based on QTL 3) Find TRS that maximizes theoretical accuracy	a) Parametric b) Targeted c) Optimize composition d) Requires CS phenotype	Mangin *et al.* (2019)
SSI	$\tilde{\beta}_{[i,TRS]} = \arg \min_{\beta_{[i,TRS]}} [0.5\beta_{[i,TRS]}\left(G_{[TRS,TRS]} + \lambda I\right)\beta'_{[i,TRS]}$ $-G_{[i,TRS]}\beta'_{[i,TRS]} + \lambda_1 \sum_{j=1}^{nTRS} \lvert\beta_{[i,j]}\rvert]$	a) Parametric b) Targeted or untargeted c) Optimize size and composition d) CS phenotype needed for λ_1 tuning through cross validation; builds a specific TRS for each TP individual	Lopez-Cruz and de los Campos (2021) Lopez-Cruz et al. (2021) Lopez-Cruz et al. (2022)
Uniform coverage of the genetic space	1) Calculate an identity by state matrix among CS genotypes 2) Select for the TRS a random genotype and discard CS genotypes that are genetically close to it 3) Repeat step 2 until all genotypes have either been included in the TRS or discarded	a) Non-parametric b) Untargeted c) Optimize composition d) d) Stand-alone and in conjunction with clustering	Bustos-Korts *et al.* (2016)
Crit_Kin Mean relationship Avg_GRM OPT_MEAN	$\arg \max_{TRS}\left(G_{[TRS,TP]}\right)$	a) Non-parametric b) Targeted c) Optimize composition d) No heuristic needed	Bustos-Korts *et al.* (2016) Roth *et al.* (2020) Atanda *et al.* (2021a) Atanda *et al.* (2021b) Lemeunier *et al.* (2022) Fernández-González et al. (2023)
Max relationship OPT_MAX	$OPT\_MAX_j = \max(G_{[j,TP]})$ 1) Compute OPT_MAX for all genotypes in the CS 2) Select the nTRS genotypes with the highest OPT_MAX values	a) Non-parametric b) Targeted c) Optimize composition d) No heuristic needed	Roth *et al.* (2020) Lemeunier *et al.* (2022)
OPT_MIN Min_GRM	$Min\_GRM_j = \min(G_{[j,TP]})$ 1) Compute *Min_GRM* for all genotypes in the CS 2) Select the *nTRS* genotypes with the lowest *Min_GRM* values	a) Non-parametric b) Targeted c) Optimize composition d) No heuristic needed	Lemeunier *et al.* (2022) Fernández-González et al. (2024)
	1) Optimize a wide range of TRS sizes with *Min_GRM* and fit a sigmoidal function to *Min_GRM* value against *nTRS* 2) Optimal TRS size is the first inflexion point	c) Optimize size	Fernández-González et al. (2024)
OPT_IND	For each TP genotype, train a GS model using only the *nTRS* genotypes with the highest relationship to it	a) Non-parametric b) Targeted c) Optimize composition d) No heuristic needed, specific TRS for each TP individual	Lemeunier *et al.* (2022)
Avg_GRM_self	$Avg\_GRM\_self \ = G_{[TRS,TRS]}$ $\arg \min_{TRS}(Avg\_GRM\_self)$	a) Non-parametric b) Untargeted c) Optimize composition	Fernández-González et al. (2023)
	Assuming that the best Avg_GRM_self fitness value is obtained when the entire CS is used as TRS, find the TRS size at which an acceptable, user-defined, fitness reduction happens.	c) Optimize size	Fernández-González et al. (2023)
Avg_GRM_MinMax	$Avg\_GRM\_MinMax \ = \left(a \cdot G_{[TRS,TP]} - \ b \cdot G_{[TRS,TRS]}\right)$ $\arg \max_{TRS}(Avg\_GRM\_MinMax)$	a) Non-parametric b) Targeted or untargeted c) Optimize composition	Fernández-González et al. (2023)
	1) Optimize TRS composition with Avg_GRM_MinMax for a wide range of TRS sizes 2) Fit the following function (*d*, *p, n* and *m* are parameters estimated in the fitting process): $Avg\_GRM\_MinMax = \frac{\ln(nTRS - d)}{m(nTRS - d)^p} + n$ 3) Optimal size is the one maximizing the value of the fitted function	c) Optimize size	Fernández-González et al. (2023)

Supplementary table 4. Cont…

Method name(s)	Mechanism of action	Observations	References
Stratified sampling	1) Divide CS in clusters 2) The number of randomly selected TRS genotypes from each cluster depends on the cluster sizes in the CS	a) Non-parametric b) Untargeted c) Optimize composition d) No heuristic needed, linear relationship between cluster sizes in TRS and CI	Isidro *et al.* (2015) Norman et al. (2018) Sarinelli *et al.* (2019) de Bem Oliveira et al. (2020) Adeyemo et al. (2020) Fernández-González *et al.* (2023)
		d) No heuristic needed, logarithmic relationship between cluster sizes in TRS and CS	Bustos-Korts *et al.* (2016)
PAM	Clustering approach. Divide the CS in *nTRS* clusters, each centered around a medoid and minimizing the distances between each medoid and the other genotypes in the cluster. The selected TRS is comprised of the *nTRS* medoids.	a) Non-parametric b) Untargeted c) Optimize composition d) No further heuristic needed	Guo *et al.* (2019) (Rio *et al.*, 2021) Fernández-González *et al.* (2023)
MaxCD	For hybrid breeding, select TRS hybrids in such a way that 1) each TP hybrid shares at least one parent with at least one TRS hybrid and 2) according to a hierarchical clustering, TRS individuals are a diverse sampling of the CS	a) Non-parametric b) Targeted c) Optimize size and composition d) No further heuristic needed	Guo *et al.* (2019)
FURS	Build a graphic network where each node is a CS genotype and iteratively select *nTRS* individuals maximizing their degree of centrality	a) Non-parametric b) Untargeted c) Optimize Composition d) No further heuristic needed	Guo *et al.* (2019)
Adversarial Validation	1) Train a binary classifier that predicts if the genotypes belong to the CS or TP based on genotypic information 2) Select the TRS individuals in the CS that the classifier struggles to classify correctly (i.e., they are similar to the TP)	a) Non-parametric b) Targeted c) Optimize size and composition a) No heuristic needed	Montesinos-López and Montesinos-López (2023)
Top	Select the *nTRS* individuals with highest phenotypes or genotypic values	a) Non-parametric b) Untargeted c) Optimize composition d) No heuristic needed, requires CS phenotype	Neyhart *et al.* (2017)
Bottom	Select the *nTRS* individuals with lowest phenotypes or genotypic values	a) Non-parametric b) Untargeted c) Optimize composition a) No heuristic needed, requires CS phenotype	Neyhart *et al.* (2017)
Tails	Select the *nTRS/2* individuals with highest phenotypes or genotypic values and the *nTRS/2* with the lowest values	a) Non-parametric b) Untargeted c) Optimize composition b) No heuristic needed, requires CS phenotype	Neyhart *et al.* (2017) Fernández-González *et al.* (2024)
Multi-objective Optimization	Simultaneously maximize TRS diversity, average relationship to the TP and trial heritability.	a) Non-parametric b) Untargeted c) Optimize size and composition d) Requires CS phenotype	Fernández-González *et al.* (2024)

Supplementary file 1 – Training set optimization algorithms

Note 1: Training set optimization - showcasing key algorithms

In this section, we show the step-by-step calculations involved in an optimization process using a toy dataset. All the code used to make all calculations described here using the R environment is provided in supplementary file 2. We focus on the classical problem of finding an optimal training set for selective phenotyping as described in Rincent et al. (2012). The characteristics of provided examples are described as follows:

- **Candidate set (CS)**: three genotyped but non-phenotyped individuals (GID1, GID2, GID3).
- **Training set (TRS)**: a subset of two genotypes from the CS will be selected to be phenotyped and act as a training set.
- **Remaining set (RS)**: this set is comprised of the CS genotype not included into the TRS.
- **Target population (TP)**: this is the population whose GEBVs are of interest. In this case, we are interested on the GEBVs of the entire CS, both the phenotyped TRS and the non-phenotyped RS. Therefore, we will consider TP = CS = (GID1, GID2, GID3). As the TP is not a distinct population independent from the CS, the problem here is **untargeted optimization**.
- The genotypes in this population have the following additive genomic relationship matrix (G):

$$G = \begin{array}{c c} & \begin{array}{ccc} GID1 & GID2 & GID3 \end{array} \\ \begin{array}{c} GID1 \\ GID2 \\ GID3 \end{array} & \begin{pmatrix} 2.76 & 1.33 & -0.17 \\ 1.33 & 1.49 & -0.02 \\ -0.17 & -0.02 & 1.25 \end{pmatrix} \end{array}$$

The first step in optimization is clearly describing all populations involved as we demonstrated above. The most suitable optimization criterion should be identified once the problem is described. For most optimization scenarios, we recommend either CDmean or Avg_GRM_self. Here we show how to perform optimization with both and later we discuss their strengths and weaknesses.

1.1. CDmean

CDmean is a parametric criterion. Therefore, the first step is describing the mixed model it is based. Here, we present the simplest and by far the most widely used underlying model:

$$y = X\mu + Zu + \epsilon$$

Where y is a vector of phenotypes, μ is the intercept, X is its corresponding design matrix (a vector of ones), $u \sim N(0, G\sigma_g^2)$ is a vector of BLUPs for the random genotypic effect, Z is its corresponding design matrix and $\epsilon \sim N(0, I\sigma_\epsilon^2)$ is a vector of independent, identically distributed (i.i.d.) residuals. This model cannot be fit at the stage of training set optimization because the phenotypes (*y)* are not known yet. Therefore, it is not possible to calculate the values for the variance components for the genotypic effect and the residuals (σ_g^2 and σ_ϵ^2). Instead, they are both assumed to be equal to 1, as Rincent et al. (2012) described that CDmean is robust to the value of the variance components.

Next, the CDmean value for all possible training sets is calculated select the one with the highest CDmean value is selected. As we have 3 genotypes in the CS and want to select 2 for the TRS, there are only 3 possible combinations: TRS1 = (GID1, GID2); TRS2 = (GID1, GID3); TRS3 = (GID2, GID3). Now, all the data is available to describe the matrices involved in CDmean calculation:

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$X = \begin{array}{r} \\ observation1 \\ observation2 \end{array} \begin{array}{c} \mu \\ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \end{array}$$

$$Z_{TRS1} = \begin{array}{c|ccc} & GID1 & GID2 & GID3 \\ observation1 & 1 & 0 & 0 \\ observation2 & 0 & 1 & 0 \end{array}$$

$$Z_{TRS2} = \begin{array}{c|ccc} & GID1 & GID2 & GID3 \\ observation1 & 1 & 0 & 0 \\ observation2 & 0 & 0 & 1 \end{array}$$

$$Z_{TRS3} = \begin{array}{c|ccc} & GID1 & GID2 & GID3 \\ observation1 & 0 & 1 & 0 \\ observation2 & 0 & 0 & 1 \end{array}$$

$$\lambda = \frac{\sigma_\epsilon^2}{\sigma_g^2} = 1$$

It is noteworthy that *Z* is the only input that varies across training sets. Therefore, this matrix is the one that contains the information describing the TRS and is key for the optimization process. When this data is plugged into the CDmean equation, the CDmean value can be evaluated for each training set:

$$M = I - X(X'X)^{-1}X'$$

$$CD = \frac{G - \lambda(Z'MZ + \lambda G^{-1})^{-1}}{G}$$

$$CDmean = mean(diag(CD_{[TP,TP]}))$$

Where mean() indicates the average of a vector, diag() indicates that the diagonal elements of a matrix are taken and the subindex [TP,TP] indicates that a subset of the matrix corresponding only to the TP genotypes is taken.

From the CD matrix, its diagonal elements correspond to the reliability of the predicted GEVB for each genotype, i.e. the squared correlation between the true and estimated breeding values. The CD matrix contains the reliability of all genotypes, but we should only consider the TP reliability values for CDmean calculaiton. However, as in this example TP = CS, the TP encompasses all available genotypes and therefore the reliability of all genotypes is maximized.

The results of the CDmean calculation for each TRS are as follows:

$$CDmean_{TRS1} = 0.21$$

$$CDmean_{TRS2} = 0.49$$

$$CDmean_{TRS3} = 0.17$$

The highest CDmean value is obtained for TRS2, which means the optimal TRS genotypes are GID1 and GID3. This makes sense, as in the G matrix GID1 and GID2 are strongly related (i.e. they are redundant), while GID3 is weakly related to either of them. Therefore, a diverse TRS must include GID3 and either GID1 or GID2.

Finally, in this example we have tested all possible combinations in the TRS to select the optimal one. This guarantees that the optimal solution will be found, but as the number of genotypes in the CS raises, the number of combinations quickly becomes too large. Therefore, in practice, a heuristic is applied instead of testing all combinations, massively reducing the computational time. We suggest using TrainSel R package heuristic as it combines good performance with high flexibility, allowing any user-defined optimization criteria, and has good documentation (https://github.com/TheRocinante-lab/TrainSel). However, many alternative heuristics are available.

1.2. CDmean with contrasts

The above example corresponds to a simplified CDmean equation in which the average reliability of TP genotypes is maximized directly. However, another widely used possibility is using contrasts to maximize the average reliability of the TP genotypes with respect to the population mean (Rincent et al. 2012). These approaches are very similar but not equivalent. Therefore, we will explore the possibility of using contrasts with the same example as before. The CDmean equation with contrasts is as follows:

$$M = I - X(X'X)^{-1}X'$$

$$CD(C) = \frac{C'(G - \lambda(Z'MZ + \lambda G^{-1})^{-1})C}{C'GC}$$

$$CDmean = mean(diag(CD(C)))$$

This equation is very similar to the previous example with the exception of the C matrix. C is a matrix of contrasts of all TP genotypes with respect to the population mean. Each column of the C matrix is a contrast vector for an individual TP genotype. Therefore, by the definition of contrasts, all columns in C must add up to 0. It is important to note that in this implementation, the C matrix has the function of indicating which are the TP individuals, which was previously done with the sub index $[TP, TP]$. In this example, the C matrix is as follow:

$$C = \begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix}$$

Using contrasts, the results of the CDmean calculation for each TRS are:

$$CDmean_{TRS1} = 0.31$$

$$CDmean_{TRS2} = 0.61$$

$$CDmean_{TRS3} = 0.18$$

These results differ slightly from previous results, but yield the same ranking of TRS and the same optimal solution (TRS2). Both CDmean with and without contrasts perform well and both can be applied. Our preference is CDmean without contrasts as it reduces computational time. The contrast matrix is usually large and dense, causing a non-negligible increase in computational cost, which as we will see later is a bottleneck for parametric criteria, such as CDmean.

1.3. Avg_GRM_self

Fernández-González et al. (2023) reported that Avg_GRM_self often outperform the CDmean in untargeted optimization with faster computational performance. This method primarily relies on maximizing the diversity of the TRS, which is one of the primer drivers of TRS quality. The speed of Avg_GRM_self makes it very useful not only for optimization directly but also for any dimensionality reduction problem in which a diverse, representative subset is of interest.

The equation of Avg_GRM_self is:

$$\text{Avg\_GRM\_self} = -mean(G_{[TRS,TRS]})$$

Where mean() refers to the average of all elements in a matrix and $G_{[TRS,TRS]}$ is a subset of the relationship matrix that only includes the genotypes in the TRS. Using the same example as for CDmean, the values for Avg_GRM_self would be the following:

$$\text{Avg\_GRM\_self}_{\ TRS1} = -1.73$$

$$\text{Avg\_GRM\_self}_{\ TRS2} = -0.92$$

$$\text{Avg\_GRM\_self}_{\ TRS3} = -0.68$$

In this case the results are different from CDmean. The TRS with the highest Avg_GRM_self value is TRS3, while for CDmean it was TRS2. Still, TRS3 includes the weakly related GID2 and GID3 making the result sensible. The Avg_GRM_self minimizes both the off diagonals of G (i.e. maximize TRS diversity) and the diagonals of G (i.e., minimize inbreeding). If the latter is undesired, the diagonal elements of G should be removed from Avg_GRM_self calculation.

Note 2: computational time in TRS optimization

Computational time is an important concern in TRS optimization and can impede its application in high dimensional datasets if not managed properly. Therefore, in this section we discuss the computational cost of different criteria.

When performing optimization, numerous training sets have to be evaluated in order to find the best one, usually through a heuristic. Every time a TRS is evaluated, a fitness function has to be calculated, which is the main driver of computational cost. Evaluating at least tens of thousands of training sets is often required in large datasets. Therefore, a small increase in the computational time of the fitness function can cause a major difference in the total time required throughout the optimization process.

2.1. Parametric, GBLUP derived

Parametric criteria based on a GBLUP model (CDmean, PEVmean, A-opt, etc.) are almost universally derived from the prediction error variance-covariance (PEV) matrix of the model. For instance, the CD matrix can be calculated as:

$$CD = \frac{G\sigma_g^2 - PEV}{G\sigma_g^2}$$

The computational bottleneck for these parametric criteria is calculating PEV ($M = I - X(X'X)^{-1}X'$; $PEV = (Z'MZ + \lambda G^{-1})^{-1}\sigma_{\epsilon}^{2}$). Its equation involves the inversion of a matrix of dimensions (n x n), where n is the number of genotypes in the TRS and TP. Therefore, these criteria have roughly a $O(n^3)$ time complexity, i.e. the time needed to calculate them is proportional to the cube of n. This makes them extremely costly for large datasets. With current computing technology, performing optimization for n > 1000 can be problematic.

2.2. Parametric, Ridge regression derived and PCA

These methods ($PEV^{ridge}mean, CD^{ridge}mean, Rscore, etc.$) are derived from a ridge regression, and therefore use the marker matrix (W) instead of the genomic relationship matrix (G). While G has the dimensions (n x n), W has dimensions (n x p), with p being the number of markers. Roughly, these criteria have a computational complexity proportional to $O(n^2p) + O(np^2)$. In most scenarios in a breeding program, p >> n, which makes np^2 >> n^3. Therefore, at first glance these methods seem much slower than the GBLUP derived ones. To solve this problem, the dimensionality of the W matrix is often reduced through principal component analysis (PCA), replacing the markers by a number of their first principal components. This allows to make p << n. In this context, n^2p << n^3, making these methods substantially faster than their GBLUP derived equivalents and suitable for datasets with thousands of genotypes. This, however, comes at the cost of losing some of the genotypic information while performing the PCA, which can degrade the optimization performance. Therefore, we recommend using the GBLUP derived criteria for smaller datasets and the PCA-accelerated, Ridge regression derived methods only if it is strictly necessary due to computational time.

2.3. Non-parametric

There are plenty of very diverse non-parametric methods, which makes it problematic to derive general rules regarding their computational efficiency. However, in general, non-parametric criteria tend to be substantially faster than the parametric, albeit often with reduced performance. For instance, Fernández-González et al. (2023) described that Avg_GRM_self is only $O(n^2)$ and has substantially less overhead than parametric criteria, making it extremely fast. It also outperformed parametric criteria in the untargeted scenario, but these advantages come with the cost of its inability to perform targeted optimization. Conversely, Avg_GRM_MinMax is also $O(n^2)$ and can work in both targeted and untargeted scenarios, but it tends to perform worse than CDmean.

In general, these criteria tend to be useful when the number of genotypes in the dataset is so large that parametric criteria are not viable. Another possibility in high-dimensionality scenarios is making a preselection of a diverse, representative subset of the CS using fast, non-parametric criteria such as

Avg_GRM_self. Subsequently, it is possible to apply a more powerful parametric criteria such as CDmean to the much smaller set of preselected individuals to find the final TRS.

2.4. Ranking method vs heuristic

Instead of the standard methodology of evaluating combinations of genotypes in the TRS (**heuristic** method, large search space), each genotype in the CS can be individually evaluated (**ranking** method, small search space). In the ranking method, genotypes are sorted according to an individually calculated fitness metric and the best ones are selected. This dramatically reduces the number of times the fitness metric has to be calculated, which is the main computational bottleneck in TRS optimization. This was described by Atanda et al. (2021b), using Avg_GRM and CDmean as fitness metrics. While this approach indeed **massively accelerates** the optimization process, it also substantially **degrades its performance**. The reason for it is that, as far as we know, it is impossible to maximize TRS diversity when using a ranking method. When genotypes are evaluated individually, it is not possible to evaluate how they interact with other individuals included in the TRS, e.g., it is impossible to know if one genotype that we decide to include into the TRS is redundant with another genotype already included. In an extreme example, if in our CS we have the best genotype duplicated, the ranking method will include both copies into the TRS (highly redundant and inefficient), while the heuristic method will be able to detect the duplication and select only one copy. Even if there are no duplicated genotypes, nothing precludes the ranking method from selecting strongly related, redundant genotypes. As described in Fernández-González et al. (2023), TRS diversity is probably the most important variable affecting its quality and therefore caution is advised when using the ranking method. In general, if computational time is a strong limitation, we recommend using faster, non-parametric criteria with the heuristic method rather than using the risky ranking method.

2.5. Ordered optimization

Ordered optimization involves not only optimizing which genotypes will be tested in the field, but also in which position they will be located (i.e., the experimental design is optimized). True optimization of the experimental design would involve considering both the genotypic information as well as spatial information (e.g., an autoregressive structure for the plots within a field could be assumed). It is very easy to include this spatial information into parametric criteria if we assume that the spatial covariance is the covariance structure for the model residuals (R). This, however, requires dropping the i.i.d. assumption for the residuals and makes the calculation of the criteria far more time consuming.

For instance, the PEV matrix calculation in this scenario would be as follows:

$$V = \mathrm{ZGZ'} + \mathrm{R}$$

$$M = V^{-1} - V^{-1}X(X'\mathrm{V}^{-1}X)^{-1}X'\mathrm{V}^{-1}$$

$$PEV = (Z'MZ + \lambda G^{-1})^{-1}\sigma_{\epsilon}^{2}$$

It is important to note that in this case the Z matrix contains both the information regarding which genotypes are selected as well as their position (order) in the field. As we explained earlier, GBLUP-derived parametric criteria are usually derived from this PEV matrix, and therefore their computational time is highly dependent on the PEV computational time. The main new problem in this equation is that the V matrix depends on the Z matrix, which itself varies for each TRS. Therefore, the inverse of the V matrix has to be performed every time that the PEV matrix is calculated. That way, for the PEV calculation we need both the inversion of a (n x n) matrix as well as the inversion of the (q x q) V matrix, where q refers to the number of observations (plots) in the experimental design. As a result, computational time would be roughly $O(n^3)$ + $O(q^3)$. In ordered optimization, q > n almost always. Therefore, q^3 >> n^3, resulting in a massively increased computational cost.

From the best of our knowledge, there are only two R packages suitable for full ordered optimization with genomic and spatial data: odw (Butler et al., 2013) and TrainSel (Akdemir et al., 2021). odw tackles the computational time problem by updating the PEV matrix instead of calculating it from scratch every time a new TRS has to be evaluated. However, one problem of odw is that its main heuristic is tabu search, which has a strong tendency of performing only a local search and converging in a local maximum instead of finding the global optimum. On the contrary, TrainSel uses a mixture of a genetic algorithm and simulated annealing to reduce the likelihood of the search being “stuck” in a local maximum. However, TrainSel requires calculating the PEV matrix from scratch for every TRS, which results in an additional computational cost. We are currently working on a $O(n^3)$ implementation of CDmean for ordered optimization, which could easily be integrated in TrainSel, solving the computational time problem.

Supplementary file 2 – R-script with examples of optimizing the TRS

```
#This code is used to calculate the examples described in Supplementary
#File 1, Note 1.
#CS = 3 genotypes
#TRS = 2 genotypes
#Untargeted optimization --> TP = CS (you could also use TP = RS)

#It is important to note that the we have made the CDmean and Avg_GRM_self
#functions as simple as possible to make them easy to read. However, this comes
#at the cost of reduced computational efficiency!


#Describe the G matrix
G <- matrix(c(2.76,1.33,-0.17,
              1.33,1.49,-0.02,
              -0.17,-0.02,1.25), nrow = 3, byrow = TRUE)
rownames(G) <- paste0("GID", 1:nrow(G))
colnames(G) <- paste0("GID", 1:ncol(G))

#Define functions used to calculate CDmean and Avg_GRM_self:
#CDmean
#This function is very computationally inefficient, as making it more efficient
#would also make it more difficult to read.
#It is for showcasing the concept only. We don't advise using it with large-scale
#datasets.
CDmean <- function(TRS, TP, G, X, sigma_e = 1, sigma_g = 1, contrasts = FALSE) {
  lambda <- sigma_e/sigma_g
  TRSfactor <- factor(TRS, levels = rownames(G))
  TPindex <- which(rownames(G)%in%TP)
  Z <- model.matrix(~TRSfactor-1)
  Z #Z matrix indicates which GIDs are observed (present in the TRS) and which
  #ones are not. The unobserved GIDs have zeros in all their rows.
  M <- diag(nrow(X)) - X%*%solve(t(X)%*%X)%*%t(X)
  PEV <- solve(t(Z)%*%M%*%Z + lambda*solve(G))*sigma_e
  if (contrasts) {
    Cmat <- matrix(-1/nrow(G), nrow = nrow(G), ncol = length(TP))
    for (i in 1:length(TP)) {
      Cmat[TPindex[i], i] <-  -1/nrow(G) +1
    }
    CD_C <- t(Cmat)%*%(G-lambda*PEV)%*%Cmat/t(Cmat)%*%G%*%Cmat
    CDmean <- mean(diag(matrix(CD_C)))
  } else {
    CD <- (G-lambda*PEV)/G
    CDmean <- mean(diag(matrix(CD[TPindex,TPindex])))
  }
  return(CDmean)
}
```

```
#Avg_GRM_self function
Avg_GRM_self <- function(TRS, G) {
  return(-mean(G[TRS,TRS]))
}
```