frontiers in Plant Science

ORIGINAL RESEARCH
published: 09 September 2021
doi: 10.3389/fpls.2021.715910

Training Set Optimization for Sparse Phenotyping in Genomic Selection: A Conceptual Overview

Julio Isidro y Sánchez[1]* *and Deniz Akdemir*[2]*

[1] Centro de Biotecnologia y Genómica de Plantas, Instituto Nacional de Investigación y Tecnologia Agraria y Alimentaria, Universidad Politécnica de Madrid, Campus de Montegancedo, Madrid, Spain, [2] Animal and Crop Science Division, Agriculture and Food Science Centre, University College Dublin, Dublin, Ireland

Genomic selection (GS) is becoming an essential tool in breeding programs due to its role in increasing genetic gain per unit time. The design of the training set (TRS) in GS is one of the key steps in the implementation of GS in plant and animal breeding programs mainly because (i) TRS optimization is critical for the efficiency and effectiveness of GS, (ii) breeders test genotypes in multi-year and multi-location trials to select the best-performing ones. In this framework, TRS optimization can help to decrease the number of genotypes to be tested and, therefore, reduce phenotyping cost and time, and (iii) we can obtain better prediction accuracies from optimally selected TRS than an arbitrary TRS. Here, we concentrate the efforts on reviewing the lessons learned from TRS optimization studies and their impact on crop breeding and discuss important features for the success of TRS optimization under different scenarios. In this article, we review the lessons learned from training population optimization in plants and the major challenges associated with the optimization of GS including population size, the relationship between training and test set (TS), update of TRS, and the use of different packages and algorithms for TRS implementation in GS. Finally, we describe general guidelines to improving the rate of genetic improvement by maximizing the use of the TRS optimization in the GS framework.

Keywords: training set optimization, genomic selection, genome-wide markers, statistical design, sparse phenotyping, genomic prediction, mixed models

OPEN ACCESS

Edited by:
Suchismita Mondal, International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Paulino Pérez-Rodríguez, Colegio de Postgraduados (COLPOS), Mexico
Gilles Charmet, INRAE Clermont-Auvergne-Rhône-Alpes, France

***Correspondence:**
Julio Isidro y Sánchez
j.isidro@upm.es
Deniz Akdemir
akdemir.work@gmail.com

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *27 May 2021*
Accepted: *10 August 2021*
Published: *09 September 2021*

Citation:
Isidro y Sánchez J and Akdemir D (2021) Training Set Optimization for Sparse Phenotyping in Genomic Selection: A Conceptual Overview. Front. Plant Sci. 12:715910. doi: 10.3389/fpls.2021.715910

1. INTRODUCTION

The rate of genetic gain in plant breeding must be enhanced to meet the demand of humanity for agricultural products in the next few decades (Xu et al., 2020). Tools, such as genomic assisted breeding (GAB), that improve the understanding of structural and functional aspects of plant genomes are key in modern breeding methods. GAB can be defined as the set of breeding tools (next-generation sequencing, omics information, and statistics) that study whole genomes by integrating multiple disciplines with new technology from informatics and robotic systems to improve selection and mating in plant breeding programs (Varshney et al., 2005, 2021). In GAB, other tools such as genetic transformation and genome editing are currently playing a key role to select better-adapted genotypes while pursuing faster genetic gains (Zhang et al., 2018). One of the emergent methodologies within GAB that have revolutionized plant and animal breeding is genomic selection (GS). GS is considered the most promising tool for genetic improvement of

the complex traits controlled by many genes, each with minor effects because (i) GS can increase the rates of genetic gain through increased accuracy of estimated breeding values (Heffner et al., 2009), (ii) significantly shorter breeding cycles (Crossa et al., 2017), and (iii) the better utilization of available genetic resources through genome-guided mate selection (Akdemir and Sánchez, 2016).

Breeders test candidate genotypes in multi-year and multi-location trials to select superior genotypes with high performance. This approach limits the number of variety candidates to be tested, and it is the main cause of the fact that plant breeding programs are time and cost-intensive. A breeding tool that combines the power of GS and the potential of an extensive collection of germplasm, assisted by new technologies, will offer promise in crop breeding to contribute to global food security (Xu et al., 2020) because it can accelerate the generation interval by reducing the generation time in plant breeding programs (Falconer and Mackay, 1996).

Bernardo (1994) was the first who proposed the use of genomic information as covariates for predicting untested genotypes but it Meuwissen et al. (2001) who came through with a new methodology to deal with the challenge of fitting prediction models when the number of genomic covariates (markers, p) is larger than the number of data points (n). Since then, simulations and empirical studies have demonstrated that GS could greatly accelerate the breeding cycle (Heffner et al., 2009), maintain genetic diversity within the breeding programs, and increase genetic gain beyond what is possible with phenotypic selection or quantitative trait loci (QTL) mapping approaches (Crossa et al., 2017). Genomic selection is a breeding tool that uses supervised machine learning approach with a training set (TRS) to predict genomic estimated breeding values (GEBVs) of an un-phenotyped test set (TS). (Isidro et al., 2016) of genotypes. The prediction of GEBVs involves a whole-genome regression model in which the known phenotypes are regressed on the markers. The GS models are trained on data that consists of both phenotypic and genome-wide markers data that is used to estimate marker (or lines) effects de los Campos et al. (2013). The combination of the marker effect estimates and the marker data from the TS is used to calculate GEBVs for the TS. The selection of individuals is based on the GEBVs as the selection criterion. The performance of the GS model is determined by calculating the correlation between GEBVs (genomic predictions) and the unknown true breeding value. As the true breeding values are never known, the available phenotypic records in the TRS are used by cross-validation values to evaluate GS. This is called prediction ability and should not be confused with prediction accuracy. The latter provides an estimate of the genotypic correlation and is estimated as the prediction ability divided by the square root of the heritability for the trait being predicted (Dekkers, 2007; Lee et al., 2008; Lorenzana and Bernardo, 2009; Riedelsheimer et al., 2012). Enhancing GS accuracy is very important for the success of GS breeding programs since the expected genetic gain from GS is directly proportional to the accuracy of GS models (Crossa et al., 2010; de los Campos et al., 2013).

There are many factors affecting the accuracy in GS by interacting in a complex network relationship (Zhong et al., 2009; Isidro et al., 2016; Liu et al., 2018; Zhang et al., 2019). Within these factors, there is one that is key to the accuracy of the prediction models in GS, and it is the design of the TRS since the predictability of a model is critical for the success of GS. In this study, the aim is to shed some light on the different TRS optimization criteria by covering the fundamentals of TRS optimization and its uses in GS, including selection strategies for long-term gains. We focus on reviewing the TRS methods from the literature that can be used as tools for designing a TRS and constructed an example to compare the TRS optimization strategies.

2. POPULATIONS IN GS

Genomic selection requires training of statistical models on available genotypic and phenotypic data from a TRS to make predictions about new genotypes. The selection of TRS involves different populations (**Figure 1**):

1. A calibration set (CS): is the group of genotypes available for the breeders from which the TRS is selected. If the individuals in this CS are phenotyped and genotyped, the populations for GS will be CS (TRS) and TS, and in theory, no need for optimization of the TRS (branch a in **Figure 1**). Nevertheless, a subset of the CS might be preferable, i.e., if very distant individuals (Lorenz and Smith, 2015) are present, to include or exclude extreme phenotypes (Lopez-Cruz and de Los Campos, 2021), or to remove irrelevant individuals (Brandariz and Bernardo, 2018). If only genotypic information is available and just a subset of them can be used for phenotyping due to budget restrictions, then a TRS will be carefully identified from the CS (branch b in **Figure 1**).
2. Training set (TRS): is where the prediction equation will be built. The TRS individuals present genotypic and phenotypic information. Under budget constraints, the aim is to select the minimum number of genotypes to phenotype, but that will assure an optimal accuracy on the TS population. The selection of the best genotypes to select from the CS to create the TRS is called optimization of the TRS. In TRS, the true response values are known (phenotypes). In this study, we used both the genotype and phenotype information from the TRS to obtain a prediction equation, which predicts the effect of each marker (or line) on the trait.
3. Remaining set population (RS): is the remaining genotypes in the CS that are used in the process of optimization. It could be also reserved for evaluating the performance of the statistical model before making predictions if the phenotypic information is available.
4. Test or Target set (TS): is the set of genotypes to predict. Only genotypic information is available in this population.

Therefore, the different populations in GS depend on whether or not the phenotypic information is available within the CS. **Figure 1** shows the distinction between the two major

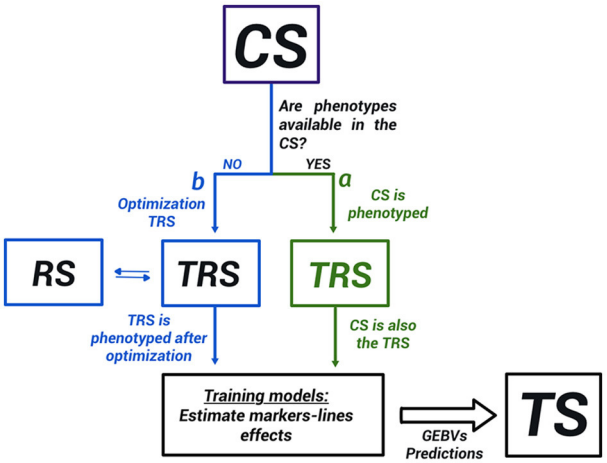

FIGURE 1 | Populations in genomic selection (GS). CS, calibration set, TRS, training set, RS, remaining set, and TS is the test set. When CS has all the phenotypic and genotypic information, CS and TRS are the same populations. Otherwise, we could have up to four different populations in the GS scheme. There are two different types of TRS selection problems: in one of these (a), the CS is already phenotyped and genotyped, and a subset of the CS is used as a TRS in the modeling stage. In the other (b), the CS is only genotyped and a TRS is constructed by phenotyping a subset of the CS.

groups of TRS optimization methods found in the literature. The first group of methods addresses the situation where the phenotypic information is already available in the CS (Neyhart et al., 2017; Brandariz and Bernardo, 2018; Lopez-Cruz and de Los Campos, 2021). They aim to use only a part of the CS when building a GS model excluding irrelevant genotypic and phenotypic information. For instance, constructing a TRS from only the individuals with high or low values of the phenotypes (Neyhart et al., 2017; Brandariz and Bernardo, 2018), or the more recently proposed sparse modeling approach Lopez-Cruz and de Los Campos (2021). The second group of methods, which is the main focus of discussion in this study, assumes that the phenotypic information is not available in the CS, and will be obtained after selecting a TRS. In this case, the resources of the breeding program are limited and just a subset of the individuals can be phenotype. In this situation, the TRS must be carefully built within the CS through an optimization process, and distinguish four different populations (CS, TRS, RS, and TS; **Figure 1**). In both groups of methods, the model validation is usually accomplished by cross-validation within the TRS (Heffner et al., 2009; Luan et al., 2009).

In general, within the TRS optimization framework, when the objective is to select a TRS to predict the remaining individuals from the same population we talk about *Un-targeted TRS*. Likewise, when a TS is first defined and genotyped, and then the TRS is optimized specifically around the TS then we define a *targeted TRS*. It is important to note, that not all optimization criteria are sensitive to this distinction, (i.e., refer next section, PAM, A-OPT, D-OPT), nevertheless, when it is so, this is reflected in how the optimization criteria are calculated (Lorenz and Smith, 2015; Akdemir and Isidro-Sánchez, 2019).

In addition, when there is heterogeneity within the environment such as row/column effects in the field, the optimal TRS of the phenotypic experiment involves not only the selection of the TRS but also the placement of genotypes in the environment (Heslot and Feoktistov, 2020). The experimental design might need blocking structure and environmental covariates and in these cases, the order in which the individuals are positioned in the environment will be important. We refer to this kind of optimization as the "ordered" optimization as opposed to the "unordered" optimization (Akdemir et al., 2021).

3. DESIGN OPTIMIZATION CRITERIA

The TRS optimization process is an optimal experimental design problem, and many aspects of GS implementation captured the attention of statisticians in the past (Smith, 1918; Kiefer, 1959; Fisher, 1960; Fedorov, 1972; Atkinson and Donev, 1992; Pukelsheim and Rosenberger, 1993; Fedorov and Hackl, 2012; Silvey, 2013). The design of the concept of the experiment should be more used to plan experimental designs in plant breeding programs and perform sets of well-selected optimization TRS to get the most informative combination out of the given factors.

The most common design optimization criteria method is indisputably the classical simple random or stratified sampling, mainly because of its simplicity and generality (Gentle, 2006), but also because of the difficulty to sample more efficiently when the number of candidate solutions is large. We classified the different design optimization criteria in to three major groups.

- Parametric design criteria are based on the assumption that the experimenter has specified a model before collecting the training data. These criteria usually depend on a scalar function of the information matrix for the model parameters which indicates the sampling variances and covariances of the estimated parameters or inferences of the model made from these models such as predictions for new individuals. Many popular designs such as the $A-$, $D-$, $E-$ criteria (Kiefer et al., 1985) are derived using a linear model as the underlying model. A linear model is a regression model where a response variable is modeled as a linear function of features that are functions of the explanatory variables plus some residual error:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is the n dimensional vector for independent realizations of the response variable, $\boldsymbol{X}$ is the $n \times p$ design matrix for the corresponding explanatory variables and $\boldsymbol{X}$ is the $n \times q$ feature matrix, $\boldsymbol{\epsilon}$ is the n dimensional vector of independent residual terms which we assume to have mean zero and fixed variance σ_e^2 and finally, $\boldsymbol{\beta}$ is the q dimensional vector of regression coefficients. The least-squares estimator for the regression coefficients is given by $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and for this estimator of the coefficients we can write the variance-covariance matrix as

$$Cov(\widehat{\boldsymbol{\beta}}) = \sigma_\epsilon^2((\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Now, suppose we have a certain design we want to evaluate which is expressed in a specific design matrix $\boldsymbol{X}_{TRS}$. Since

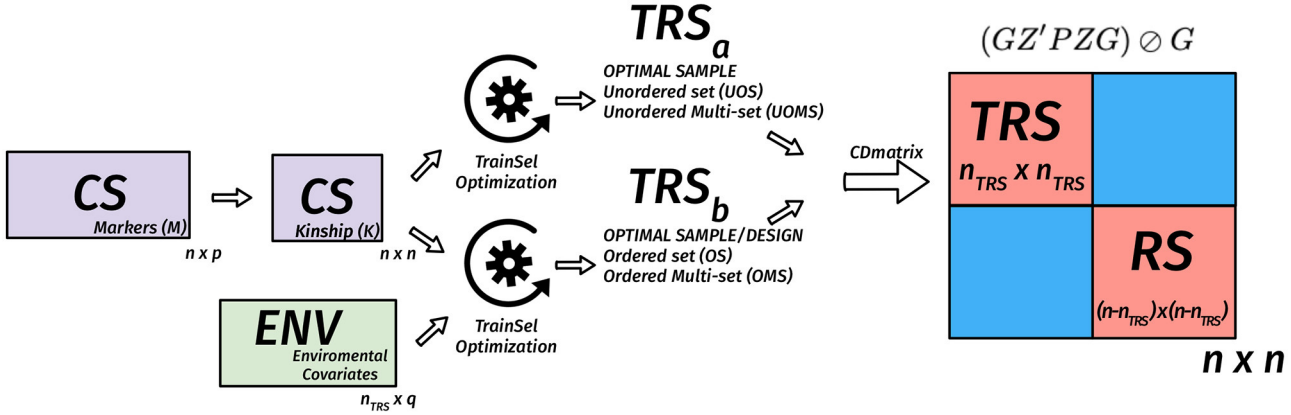

FIGURE 2 | Diagram of the matrix of the coefficient of determination (CD) criterion in the TRS optimization for unordered and ordered experiments for the un-targeted case. We assume that we have n genotypes in the candidate set, n_{TRS} genotypes are selected to the TRS. In this case, the $n \times n$ kinship matrix is used to calculate the $n \times n$ CD matrix for a given TRS, then the mean CD is calculated based on certain diagonals that correspond to TRS or remaining set (RS). The optimization algorithm is used to find the best TRS. When provided with a design matrix that has heterogeneous rows, then we are also looking for a design in addition to the selection of a TRS. In this case, both the kinship matrix and the $n_{TRS} \times q$ environmental covariates matrix are used to calculate the $n \times n$ CD matrix for a given TRS, then the mean CD is calculated based on certain diagonals that correspond to TRS or RS. The optimization algorithm is used to find the best TRS and best design with this TRS.

we can write the covariance of the estimated coefficients as $(\boldsymbol{X}'_{TRS}\boldsymbol{X}_{TRS})^{-1}$ up to a proportionality constant (which is the same for all other possible designs), we can use a function of this matrix to compare it with other designs. In general, a scalar function of this matrix is used to order the different designs. D-optimality criterion, for instance, can be expressed as $|(\boldsymbol{X}'_{TRS}\boldsymbol{X}_{TRS})|$, and designs with higher values are considered better. A-optimality criterion is expressed as $trace\left[(\boldsymbol{X}_{TRS})'\boldsymbol{X}_{TRS})^{-1}\right]$, and designs with lower values are considered better.

Some other criteria such as *CDmean*, *PEVmean*, (Laloë, 1993; Rincent et al., 2012; Isidro et al., 2015) rely on a mixed model as the underlying model: In the linear mixed-effects model of interest, the observations are assumed to result from a hierarchical linear model:

$$\boldsymbol{y} = \boldsymbol{E}\boldsymbol{\beta}_{env} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

with $\boldsymbol{E}$ is the $n \times p$ design matrix for the environmental covariates, $\boldsymbol{\beta}_{env}$ is the p vector of the effects of the environmental covariates, $\boldsymbol{Z}$ is the $n \times N$ design matrix for the N genotypes in the candidate set, $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \boldsymbol{R})$ is independent of $\boldsymbol{u} \sim N_q(\boldsymbol{0}; \boldsymbol{G})$. When using this mixed model in genomic prediction for a single environment, we use $\boldsymbol{G} = \sigma^2_k\boldsymbol{K}$ and $\boldsymbol{R} = \sigma^2_e\boldsymbol{I}$, where $\boldsymbol{K}$ is the relationship matrix of the genotypes (CS and if available the TS). When we use this mixed model with a multi-environmental genomic prediction, we assume $\boldsymbol{G} = \boldsymbol{V}_k \otimes \boldsymbol{K}$ and $R = \boldsymbol{V}_e \otimes \boldsymbol{I}$.

For this model, the CD matrix of $\widehat{\boldsymbol{u}}$ for predicting $\boldsymbol{u}$ is given by

$$(\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}) \oslash \boldsymbol{G}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{E}(\boldsymbol{E}'\boldsymbol{V}^{-1}\boldsymbol{E})^{-1}\boldsymbol{E}'\boldsymbol{V}^{-1}$ is the projection matrix and $\oslash$ expresses the element-wise division. Usually, the mean of certain diagonal elements of the CD matrix is used to measure the quality of a sample. For instance, in a targeted design, the mean of the diagonal elements that correspond to the TS genotypes are used. When the design is un-targeted, we can use the mean over the diagonals that correspond to the remaining set. Another approach involves the calculation of the CD matrix for a given set of contrasts then taking the mean of the diagonals of this matrix (Rincent et al., 2012, 2017). In **Figures 2–4**, we diagrammatically illustrate the different populations, input matrices, the different parts of the CD matrix, and the process of optimization.

- Non-parametric designs criteria are model-free, i.e., they do not rely on models we intend to use with the resulting data. Some nonparametric designs are based on distance or similarity measures and aim to spread the TRS over the design space (space-filling design). Different measures or metrics quantify how a set of points is spread out. Some examples are: (i) partition around medoids (PAM) where the objective is to find a sequence of objects called medoids that are centrally located in clusters for a given distance measure, (ii) the maximin criteria are such that the minimum distance among the TRS is maximized, (iii) the minimax design (Johnson et al., 1990) where the TRS is such that the maximum of the minimum distances from the TRS to the rest of the CS or the TS is minimized, (iv) the Latin hypercube sampling divides the design region evenly into cubes and ensuring that the sample contains just one point in each such segment and aims at ensuring that each of the scalar inputs has the whole of its range well scanned, according to a probability distribution, and (v) the minimum spanning tree (MST) (Dussert et al., 1986). An MST is a tree that connects all the candidate design points and whose total edge lengths are minimal. Once a spanning tree of the candidate points is built, the mean and SD of edge lengths can be calculated. The spanning trees with the

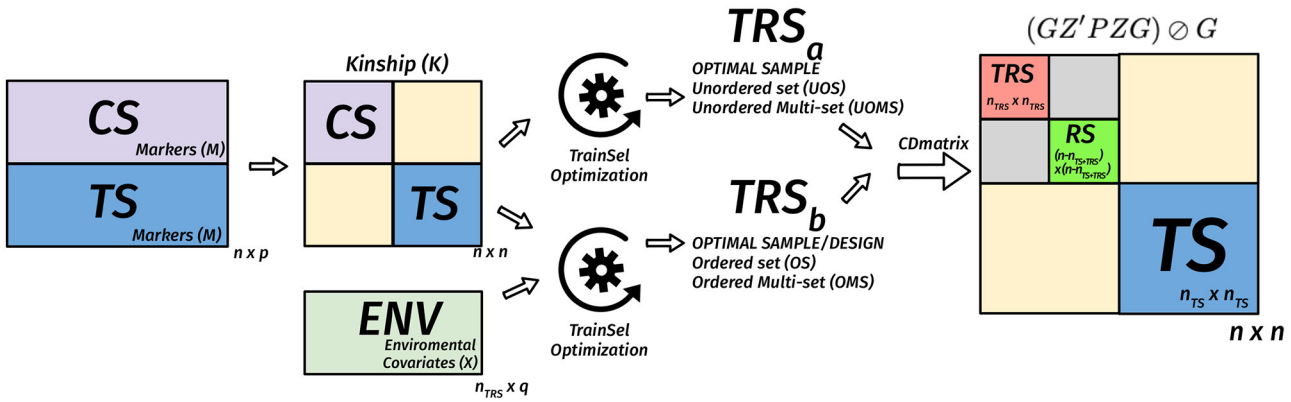

FIGURE 3 | Diagram of the matrix of the CD criterion in the TRS optimization for unordered and ordered experiments for the targeted case. We assume that we have n_{TS} genotypes in the TS, $n - n_{TS}$ genotypes in the CS, n_{TRS} genotypes are selected to the TRS from the CS. In this case, the $n \times n$ kinship matrix is used to calculate the $n \times n$ CD matrix for a given TRS, then the mean CD is calculated based on certain diagonals that correspond to TS. The optimization algorithm is used to find the best TRS. When provided with a design matrix that has heterogeneous rows, then we are also looking for a design in addition to the selection of a TRS. In this case, both the kinship matrix and the $n_{TRS} \times q$ environmental covariates matrix are used to calculate the $n \times n$ CD matrix for a given TRS, then the mean CD is calculated based on diagonals that correspond to TS. The optimization algorithm is used to find the best TRS and best design with this TRS.

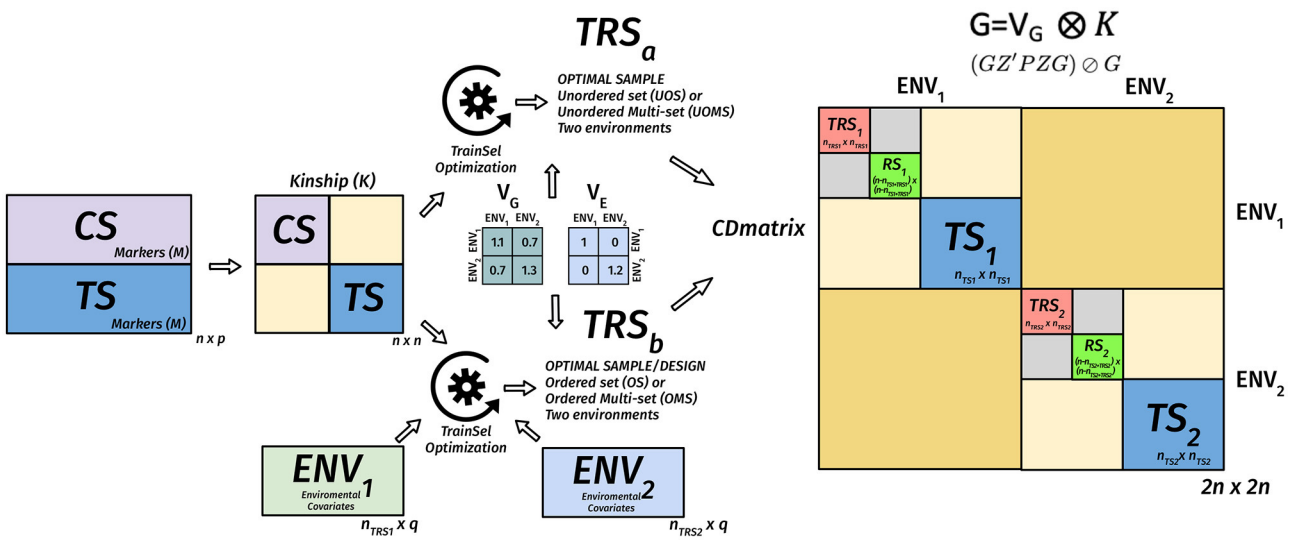

FIGURE 4 | Diagram of the matrix of the CD criterion in the TRS optimization for unordered and ordered experiments for the targeted case in a multi-environmental phenotypic experiment. We assume that we have n genotypes, n_{ts} of them are in the target set (TS), the remaining of in the candidate set, n_{TRS_1} genotypes are selected to the TRS in environment 1, n_{TRS_2} genotypes are selected to the TRS in environment 2. Two environments are assumed to have a positive genetic covariance, and this is expressed in V_k. The residual genetic covariance expressed in VE is diagonal, meaning that errors are uncorrelated between the two environments. These covariance matrices along with the genomic relationship matrix and if provided environmental covariates matrices for the environments are used to calculate the CD matrix ($2n \times 2n$) for a given design. The mean of the diagonals of this matrix that correspond to the TS is used as a criterion for evaluating different designs. The optimization algorithm tries to find the design that maximizes this criterion.

smallest mean are called minimal and among them, the ones with high variance are preferred. A TRS from an MST can be obtained by recursively pruning out, from the candidate set, the candidate points on the leaves of the MST with small edge lengths (Guo et al., 2019).

Non-parametric designs such as space-filling designs are well suited to the initial exploration objective. They can be used to select a smaller candidate set from a bigger candidate set to reduce the computational complexity of optimizing parametric design criteria.

- **Multiple design criteria**. Multiple models optimal experimental design criteria try to overcome the choice issue by combining more than one criteria into one *via* some type of averaging on multiple-objective optimization methods (Pukelsheim, 1993; Akdemir and Sánchez, 2016). In this approach, the Pareto front approach is used to evaluate several

criteria. The Pareto front is a set of non-dominated designs, i.e., as compared to the design points on the frontier, no other design point can be found that does not degrade at least one of these criteria values (as shown in **Figure 5**).

Many GS experiments will be performed in several environments and then the TRS optimization aims to find subsets of genotypes from the candidate set to be tested in each of the environments and perhaps the corresponding designs within some of these environments to address the heterogeneity within environments. The use of CD for this situation is illustrated in a diagram in **Figure 4**.

4. TRS OPTIMIZATION FOR SPARSE PHENOTYPING

The most important current bottleneck in plant breeding programs is the phenotypic evaluation (Crossa et al., 2017). Although genotyping is still costly, next-generation sequencing has decreased genotyping cost more than 100K folds in the last 20 years (National Human Genome Research Institute, 2020), and therefore, phenotyping needs to be optimized within a breeding program. The use of GS in breeding programs is potentially costly without the careful design of populations. When designing the implementation of the GS scheme into the breeding cycle, breeders need to focus first on several aspects: (i) to generate a specific breeding database for GS, (ii) to choose the filial generation to start GS, and (iii) to select the TRS to start GS modeling (Albrecht et al., 2011; Clark et al., 2012). The design of the TRS, also called optimization of the TRS, is the breeding process that uses the information from these aspects to create a TRS to start the GS process.

Training set optimization consists of choosing (within a panel of candidates) a set of training individuals that will better predict un-phenotyped germplasm in a TS. TRS optimization has attracted notable interest in the breeding community for several reasons (**Table 1**). First, the fact that predictions are based on markers or line effects calculated on the TRS raises the question of how to select the TRS to increase the efficiency and effectiveness of GS. Second, currently, the high cost of phenotyping makes the phenotype information the most important constraint in plant breeding programs. Better allocation of resources within plant breeding programs by observing a small size but representative TRS would reduce phenotypic cost and increase the quality of the phenotypic data by focusing on more expensive traits with more sophisticated instruments, or increasing complementary measurements of the same traits (sparse or selective phenotyping). Third, the traditional optimization process based on random sampling as a strategy to create the TRS does not always lead to an increase in predictive ability due to the under or over-representation of the genetic information in the TRS. The TRS optimization aims to enhance the process of sparse phenotyping, to reduce the cost of phenotyping while maintaining high prediction accuracy models.

Two important aspects within the TRS optimization are the fact that the TRS is a dynamic populations that must be updated through the breeding cycle program, and also that the TS needs to be into account when building the TRS (Akdemir et al., 2015).

The design of the TRS was initially started in animal breeding (Habier et al., 2007, 2010; Clark et al., 2012; Pszczola et al., 2012). These studies and others in plants (Windhausen et al., 2012; Wientjes et al., 2013) were focused on the importance of the relatives for the makeup of the TRS and on how to update the TRS to improve genomic prediction across generations. They highlighted how the TRS should be composed in terms of resemblance between TRS and TS, but they did not perform any optimization process, TRS was selected randomly. A random sampling of genotypes from a CS is a risky procedure because could lead to low-quality coverage of the total genetic space especially when the CS contains population structure (Windhausen et al., 2012; Isidro et al., 2015; Bustos-Korts et al., 2016). In the last decade, many studies (**Table 1**) examined the importance of optimization of the TRS by comparing specific selection criteria to random sampling.

The first study highlighting the importance of using statistical approaches to develop an optimal TRS was shown by Rincent et al. (2012) (**Table 1**). In this study, the objective was to define which individuals from a calibration (candidate) set are the optimal ones to predict a selection (TS) candidates. The idea was to use a criterion that could minimize genetic similarity within the TRS, because of the more similar the individuals within the TRS, the more duplication of the alleles, and therefore, more redundancy. Based on concepts from the mixed model equations introduced by Laloë (1993), Rincent et al. (2012) introduced criteria that aimed to maximize the reliability CD, the square correlation between GEBVs and true breeding values or minimized the prediction error variance (PEV) on the CS. In this study, they used a generalized version of CD and PEV (the contrast between breeding values). They showed that the optimization criteria improved prediction accuracy when comparing with random sampling. Rincent et al. (2012) have shown that mean of the coefficient of determination (CDmean) captured more genetic variability when building the TRS than mean of the prediction error variance (PEVmean) and that an optimized set of 100 lines achieved on average the same prediction accuracy as a set of 200 lines selected at random.

Isidro et al. (2015) proposed stratified sampling and stratified CD as alternative algorithms to improve the optimization of TRS under population structure effects. The optimization of the TRS based on genomic relationships resulted in higher prediction accuracies when compared with random sampling. In this study, they concluded that the optimization of the TRS depended on the interaction of trait architecture and population structure and on the ability of the algorithm to capture phenotypic variance. In the same year, Akdemir et al. (2015) derived a computationally efficient approximation to the PEV based on principal components of the genotypes as a criterion for TRS design that showed less computational burden than previous criteria. These studies were the first ones that open the door to other strategies to optimize the TRS. Bustos-Korts et al. (2016) proposed a TRS construction method that uniformly sampled the genetic space comprised by the target population (TS) of genotypes, although, the results were similar to CDmean.

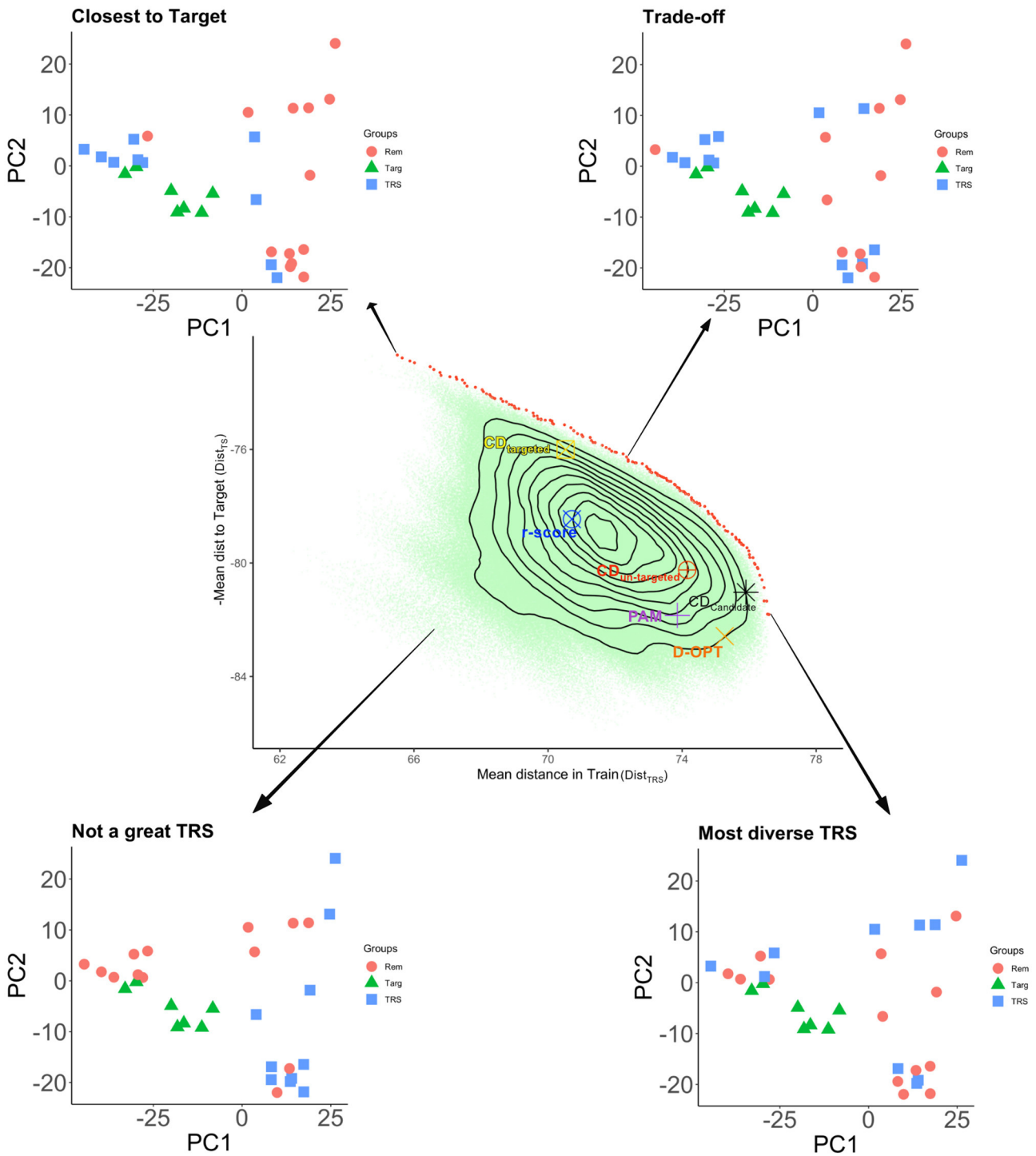

FIGURE 5 | Pareto front for minimizing the mean genomic distance of TRS genotypes to the TS genotypes (i.e., maximizing the negative of this quantity), and maximizing the mean genomic distance among the TRS genotypes. The figure represents a toy example where a sample of size 10 is selected from 23 candidate genotypes to predict 7 TS genotypes (the total number of different solutions is 23 choose 10 which is more than a million). Target genotypes along with selected TRS and the remaining sets are displayed in the genotypic space represented by the first two principal components of the marker matrix. With different symbols and colors, we indicate the optimal CD TRS's for targeted and un-targeted cases, D-optimal TRS, and the TRS selected by PAM. The red dots are the TRS that are on the Pareto front, i.e., no other TRS will be better than any of these for both criteria (non-dominated solutions). All the brown dots are dominated by the same two criteria. We get the most diverse set when the mean genetic distance in the TRS is maximal. We get a TRS closest to the TS when we minimize the mean genetic distance (maximize the negative) of TRS to TS. All of the parametric design criteria and PAM are dominated. Among those, CDmean targeted gives a TRS that is close to the TS. The remaining optimal TRS's are genetically diverse. The most genetically diverse set among the optimization criteria is the CDmean calculated for all genotypes in CS.

TABLE 1 | Key relevant scientific studies on training set (TRS) optimization.

Study	CDmean	PEVmean	Clustering	Other criteria	Package
Rincent et al. (2012)	✗	✗	–	–	Own code
Isidro et al. (2015)	✗	✗	✗	–	Own code
Akdemir et al. (2015)	✗	✗	✗	–	**STPGA**
Lorenz and Smith (2015)	–	–	–	Levels of TRS relationship	Own code
Bustos-Korts et al. (2016)	✗	–	✗	Uniform Sampling	Own code
He et al. (2016)	–	–	–	Random	–
Rincent et al. (2017)	✗	–	✗	CDpop and Crit_Kin	Own code
Neyhart et al. (2017)	✗	✗	–	Top and bottom proportion	Own code
Cericola et al. (2017)	–	–	–	Random sampling	Own code
Momen and Morota (2018)	✗	✗		Additive and Non-additive	Own code
Norman et al. (2018)	–	–	✗	Random	Own code
Akdemir and Isidro-Sánchez (2019)	✗	✗	–	D and A-OPT	**STPGA**
Ou and Liao (2019)	✗	✗	✗	r-score	**TSDFGS**
Mangin et al. (2019)	✗	✗	–	EthAcc	Own code
Guo et al. (2019)	✗	✗	PAM	FURS	**STPGA**
de Bem Oliveira et al. (2020)	–	–	–	Random, Family Random	Own code
Adeyemo et al. (2020)	–	–	✗	–	Own code
Mendonça and Fritsche-Neto (2020)	–	✗		–	**STPGA**
Olatoye et al. (2020)	✗	–	–	Random	Own code
Roth et al. (2020)	–	✗		Maximum and Mean relationship	**STPGA**
Sarinelli et al. (2019)	–	✗	✗	–	Own code
Tayeh et al. (2015)	✗	–	–	–	Own code
Atanda et al. (2021)	✗	–	–	Avg_GRM	Own code
Yu et al. (2020)	–	–	–	Upper Bound reliability	Own code
Ben-Sadoun et al. (2020)	✗	–	–	CDmean-multi	Own code
Heslot and Feoktistov (2020)		–	–	PEVridge	Own code
Akdemir et al. (2021)	✗	✗	✗	–	**TrainSel**
Kadam et al. (2021)	✗	✗	–	–	**STPGA**

CDmean, Mean of the coefficient of determination; PEVmean, Mean of the predictor error variance. A cross in the cell indicates that the criterion has been used for TRS optimization. Criteria different than CD, PEV, and Clustering are shown in the column Other Criteria. The software using R is specified in the Package column.

Other studies also stressed the importance of considering an other way to construct the TRS by random sampling (Lorenz and Smith, 2015; He et al., 2016; Cericola et al., 2017; Neyhart et al., 2017; Norman et al., 2018; de Bem Oliveira et al., 2020; Olatoye et al., 2020), clustering approaches (Akdemir et al., 2015; Isidro et al., 2015; Bustos-Korts et al., 2016; Rincent et al., 2017; Norman et al., 2018; Guo et al., 2019; Sarinelli et al., 2019; Adeyemo et al., 2020), by using different levels of relatedness between TRS and TS (Lorenz and Smith, 2015; Berro et al., 2019; Roth et al., 2020) or by using other alternatives algorithms to CD-mean and PEV-mean such as different design matrix algorithm (Akdemir and Isidro-Sánchez, 2019), estimated theoretical accuracy (EthAcc) (Mangin et al., 2019), upper bound reliability (Yu et al., 2020), or the Fast and Unique Representative Subset Selection (FURS) (Guo et al., 2014). A criterion that is derived directly from Pearson's correlation between GEBVs and phenotypic values of the TS derived from the GBLUP model showed higher predictive ability than CD and PEV (Ou and Liao, 2019). Most aforementioned approaches above, do not use information from the TS while building the TRS, which is detrimental for prediction accuracy (Lorenz and Smith, 2015; Akdemir and Isidro-Sánchez, 2019; Ou and Liao, 2019). The main reason for the decrease in accuracies is because the most informative TRS to predict the TS is the one where individuals are more closely related to the TS. This is because when pairs of individuals are closely related, they tend to inherit QTL blocks in the same linkage phase (Andreescu et al., 2007; Habier et al., 2010). This is especially critical when there is low marker density coverage because the assumption in GS of getting at least one marker in QTL with the trait of interest will not be perfectly met. The genetic relatedness between TRS and TS was addressed by Lorenz and Smith (2015), Rincent et al. (2017), and Akdemir and Isidro-Sánchez (2019). Recently, Atanda et al. (2021) used the average genomic relationship (Avg_GRM in **Table 1**) between a specific line in the TRS and all lines in the TS, and they statistically significant increase in the accuracies when compared with CD in some bi-parental populations. Nevertheless, this approach as in Rincent et al. (2017) did not consider the possible alleles duplication within the TRS.

Training optimization selection also has been used for pre-breeding discovery. Tanaka and Iwata (2018) proposed a strategy that used genomic prediction in pre-breeding for discovering

the best genotypes from a large number of candidates. They demonstrated by simulation that their Bayesian optimization could reduce the number of phenotyped accessions needed to find the best accession among a large number of candidates. Their strategy was based on predict uncertainty of the prediction rather than based only on high predicted values. Following this strategy, Tsai et al. (2021) used an augmented expected improvement for sequential phenotyping to identify the best individual from the CS. It is important to note that these studies are not focusing on building a TRS for GP, but on identifying the best candidate to be used for commercial or mating purposes. These approaches could be used when phenotyping is very expensive and not very time-consuming.

In the area of hybrid breeding, the optimization of the TRS is even more critical than in other breeding systems, since the selection of superior F1 hybrids (single crosses between fully inbred lines) implies developing first inbred lines and then identifying the best hybrid combinations between them. To facilitate this process, breeders typically split germplasm into complementary heterotic groups and select lines within each group for their ability to produce good hybrids when crossed to lines from a complementary group. The fullest assessment of single-cross performances would be a complete factorial mating design achieved by making all possible single crosses. However, the high number of lines to be evaluated per heterotic group makes this approach prohibitive (i.e., for 1,000 lines in each heterotic group, there would be 1 million possible crosses). Genomic models have been applied to hybrid prediction mainly in maize (Bernardo, 1994; Schrag et al., 2009; Technow et al., 2014; Kadam et al., 2016; Marulanda et al., 2016; Fristche-Neto et al., 2018; Seye et al., 2020), and wheat (Zhao et al., 2013, 2014, 2015; Longin et al., 2015; Marulanda et al., 2016; Schulthess et al., 2017), and less in other species such as rye (Wang et al., 2014) or sunflower (Reif et al., 2013; Mangin et al., 2017; Dimitrijevic and Horn, 2018; Heslot and Feoktistov, 2020). These studies have emphasized the interest in using TRS optimization compared to the traditional crossing designs.

In general, most of the TRS studies have used model-based parametric criteria (CDmean, PEVmean, and r-score), followed by non-parametric (i.e., PAM, FURS), and just a few studies used their own criteria (i.e., AvgGRM, U score) (**Table 1**). All these studies show that there is not a universal criterion to create a TRS. It will mainly depend on linkage disequilibrium between markers on TRS vs. TS, the relationship between TRS and TS (Habier et al., 2007; Goddard, 2009), the genetic architecture of the trait (McClellan et al., 2007; Jannink, 2010; Burstin et al., 2015), trait heritability (Hayes et al., 2009), and population structure effects (Isidro et al., 2015; Rincent et al., 2017).

To shed some light on the different TRS optimization criteria, we constructed a toy example where we compared several design criteria (CD, PAM, D-OPT, and r.score) with each other (**Figure 5**). In this example, there were 30 genotypes in total, seven of these genotypes were selected as the TS. The remaining 23 genotypes were used as the CS. We set the TRS size to 10, giving 23 choose 10 (1144066) different TRS possibilities. For each of these designs, we calculated the value of the mean genetic distance among the TRS ($Dist_{TRS}$), and the negative of the mean genomic distance from TRS to the TS ($Dist_{TS}$). In the Figure, the red dots are the TRS that are on the Pareto front, i.e., no other TRS will be better than any of these for both criteria (non-dominated solutions). Balancing the $Dist_{TRS}$ and $Dist_{TS}$ in the Pareto front gives you different TRS. For instance, when we minimize the mean genetic distance (maximize the negative) of TRS to TS, we obtained a TRS closest to the TS (top left graph). We get the most diverse TRS when the $Dist_{TRS}$ in the TRS is maximal (bottom right graph). If you balance both distances, then we get a TRS where there is a trade-off between $Dist_{TRS}$ and $Dist_{TS}$. The remaining TRS on the same plot is dominated with respect to the same two criteria. A TRS is dominated if we can find another TRS that improves at least one of these criteria without deteriorating the other criterion value. All of the design criteria and PAM are dominated with respect to $Dist_{TRS}$ and $Dist_{TS}$. Among those, CDmean targeted gives a TRS that is close to the TS, where CDmean calculated over the candidate set (CDMEAN-Cand) comes very close to the most diverse design. The contours of the density of $Dist_{TRS}$ and $Dist_{TS}$ over 1144066 different TRS possibilities show that a random design on average would be dominated by all of the optimal samples and would fall far away from the Pareto frontier. It is important to understand the different trade-offs involved in choosing a good TRS since this will help the experimenter to choose a suitable TRS or a TRS selection criterion among the alternatives.

Breeding programs usually deal CS's with 1,000's or 10,000's of genotypes. Although direct enumeration of all the possible TRS's is not possible in these cases, multi-objective optimization techniques can be utilized to approximate the frontier curves and single-objective optimization tools can be used to find optimal TRS's according to several single criteria. Then a plot similar to the one presented in **Figure 5** can be produced to evaluate the trade-offs among different designs. When the number of genotypes in the CS is so large that computationally intensive methods are prohibitive, we recommend using a less intensive method such as PAM or stratified sampling (Isidro et al., 2015; Guo et al., 2019), or one of the space-filling designs to reduce the number of CS to a manageable size ahead of comprehensive analysis. A practical overview of the statistical analysis needed to optimize the TRS using R and issues associated with the analysis have been addressed along with the R code in the study by Isidro y Sánchez et al. (2022). In addition, extra information can be found in the extensive vignette (https://github.com/TheRocinante-lab/TrainSel/blob/main/inst/TrainSelUsage.pdf).

5. SOFTWARE TOOLS FOR TRS OPTIMIZATION

While the practical use of TRS optimization in GS is supported by the literature, as shown above, the number of software tools for implementation is limited. As far as we are concerned, just three software have been developed and available for public use. The package STPGA Akdemir (2017) is an R package that uses a modified GA for solving subset selection problems but also allows users to chose from many predefined or user-defined criteria. Similarly, the package TSDFGS Ou and Liao (2019) is

an R package that focuses on optimization of the TRS by a genetic algorithm (GA) and can be used for TRS optimization based on three built-in design criteria [CDscore, PEVscore, and Pearson correlation (r-score)]. Recently, Akdemir et al. (2021) designed a new package called TrainSel to provide many more options than previous software. For example, TrainSel can select multiple sets from multiple candidate sets, users can specify whether or not the resulting set needs to be ordered, or the power to perform multi-objective optimization. In addition, TrainSel can be used for searching for solutions to a variety of TRS and experimental design problems, such as randomized complete block design, and lattice design, etc. Furthermore, it can be also used in combinatorial optimization problems for supervised and also unsupervised learning. The strength of TrainSel is that it combines TRS optimization with a particular experimental design, which has not been implemented in both of the above alternatives by Akdemir et al. (2021).

6. GENERAL GUIDELINES FOR A GOOD TRS

In this study, we highlight some of the guidelines learned from the literature when building an optimal TRS:

- When building the first TRS is key to keep, within the TRS, the historical germplasm used to generate the breeding populations. This will allow capturing the allelic diversity within the breeding program.
- The larger the TRS size the better predictions (Daetwyler et al., 2008; Zhong et al., 2009), since most characters are quantitative with a large number of loci and a very small effect size. The number of loci affecting quantitative characters likely ranges from 2,000 to 4,000 (MacLeod et al., 2016). Although adding genetically distant individuals might decrease accuracy (Lorenz and Smith, 2015), this is not a general rule. In addition, large TRS are needed to capture rare alleles at high frequencies to obtain a reliable estimate of their effects (MacLeod et al., 2016), even for highly quantitative traits if the rare allele is present in the sequencing or the genotyping is done from coding and regulatory regions.
- Markers can capture genetic relationships among genotypes, thereby affecting the accuracies of GEBVs (Habier et al., 2007). Therefore, a genetic relationship between TRS and TS is needed to obtain high accuracies. In general, a TRS should maximize the relationship with the TS (Albrecht et al., 2011; Pszczola et al., 2012; Akdemir and Isidro-Sánchez, 2019), but should minimize the relationship within the TRS (Clark et al., 2011; Lorenz, 2013; Bustos-Korts et al., 2016; Pszczola and Calus, 2016). That is to say, if TRS and TS come from different populations or breeding generations, a drop in accuracy is expected. The main reasons for the drop in accuracy are because LD between markers and QTL, or that QTL allele frequencies and/or effects can differ among populations (Hayes et al., 2009; Wientjes et al., 2015, 2017). The difference in allele frequencies between TRS and TS can affect prediction accuracy because allele frequencies can affect the estimated genomic relationship matrix when GBLUP models are implemented.
- The TRS must be updated with new genotyped and phenotyped individuals to assure the accuracy of GEBVs, is maintained over generations. Otherwise, recombination events will decrease LD between markers and QTL (Auinger et al., 2016). As phenotypes are the current bottleneck in plant breeding programs, the quality of the phenotypes is critical to the TRS optimization.
- The design of the TRS highly depends on the TS population. For example, if your TS is highly diverse, your TRS must be built to capture that diversity, otherwise, a significant drop in accuracy might occur. That is why targeted optimization approaches are chosen when building TRS (Akdemir and Isidro-Sánchez, 2019; Akdemir et al., 2021). From **Figure 5** we can observe that we get a TRS closest to the TS when we minimize the mean genetic distance (maximize the negative) of TRS to TS. Among the different TRS selection criteria, CDmean targeted gives a TRS that is close to the TS. The remaining optimal TRS's are genetically diverse but the most genetically diverse set among the optimization criteria is the CDmean calculated for all genotypes in CS. This type of evaluation of different design criteria together along with a frontier curve should shed some light on the selection of a particular TRS.
- If certain QTL with large effects for traits of interest exists, then these QTL can be given more influence while selecting the TRS. This could be done, for example in the mixed modeling framework by using the QTL as fixed effects (Spindel et al., 2016). In the non-parametric approach, more weights can be given when calculating the genetic distance matrix.
- In general, optimization criteria from mixed model theory (CDmean, PEVmean) performs better than random sampling under most scenarios, except for scenarios with a large population structure where these criteria might not be optimal (Isidro et al., 2015).

7. PERSPECTIVES FOR TRS OPTIMIZATION

Genomic selection is an emergent methodology that revolutionized plant and animal breeding, by using a statistical framework that uses genome-wide markers to predict breeding values for key breeding traits. In this framework, one critical step is how to select the best individuals to train the statistical models. As shown above, there has been quite a great research in this area, but there are still some questions to be answered. Following the literature, there is no "best" strategy to optimize the TRS, and therefore, a comparison between algorithms focusing on the different factors affecting the TRS on different populations would be helpful to answers some questions regarding TRS optimization.

We envision a substantial benefit applying TRS optimization methods to hybrid prediction, and also sparse testing in multi-environment, and multi-trait experiments (Jarquín et al., 2014; Akdemir et al., 2021; Crossa et al., 2021). For instance, in hybrid

prediction, TRS are traditionally constructed by methods such as top crosses, North Caroline design, etc. It has been shown that the TRS optimization methods improve hybrid prediction accuracies when comparing with the traditional design methods (Zhao et al., 2015, 2021; Fristche-Neto et al., 2018; Heslot and Feoktistov, 2020; Yu et al., 2020; Technow et al., 2021).

It is also expected that TRS selection methods will be used more commonly in multi-environmental phenotypic experiment design (Montesinos-López et al., 2019; McGowan et al., 2020) as more flexible and powerful tools such as the package R TrainSel becomes available for researchers. The use of genomic information in designing these experiments shifts the attention from replication of individuals to replication and representation of alleles in different environments.

In addition, more studies using haplotypes rather than just markers are needed, since accuracies are greater if TRS and TS share long-range haplotypes (Akdemir et al., 2015; Meuwissen et al., 2016; Scott et al., 2021). The decrease of whole genomic sequencing is allowing us to develop pan-genomes studies of many crops, which will allow us to switch from SNPs to longer more important haplotypes in the design of TRS populations. The development of haplotype-informed DNA markers will enable the selection of new haplotype combinations, which will increase the opportunity to attain optimized genetic combinations for improved performance and disrupt linkage drag (Varshney et al., 2021).

An unresolved issue in TRS optimization is the determination of the size of TRS. The size of TRS is usually dictated by the budget for the experiment, however, a breeder might need guidance for selecting a TRS size to avoid redundancy of individuals. For example, even though a breeder might have the resources to do 20 individuals, the breeder should know what is the optimal size to experiment. The optimal size of the TRS can be obtained from the multi-objective optimization framework Akdemir et al. (2019). The solutions on the Pareto front of an optimization problem Markowitz (1968), where one or more design criteria along with the TRS size are optimized, will provide the experimenter with a scenery of the optimal design space at each sample size. The usual methods of selecting a solution on a frontier can guide the determination of the TRS size.

Finally, a comparison of criteria with different populations, different genetic architectures, heritability values, and relationships among TRS and TS is needed, especially to evaluate if some previous claims in the TRS optimization area are true under the same population scenarios.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

JIS and DA: conception and design of the article, drafting the article, and critical revision of the article. Both authors contributed to the article and approved the submitted version.

FUNDING

Results have been achieved within the framework of the first transnational joint call for research projects in the SusCrop ERA-Net Cofound on Sustainable Crop production, with funding from the Department of Agriculture, Food and the Marine grant no. 2017EN104. This project has also received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 818144. JIS was supported by the Beatriz Galindo Program (BEAGAL18/00115) from the Ministerio de Educación y Formación Profesional of Spain and the Severo Ochoa Program for Centres of Excellence in R&D from the Agencia Estatal de Investigación of Spain, grant SEV-2016-0672 (2017-2021) to the CBGP.

ACKNOWLEDGMENTS

The authors thank all their funders as well as Breicon Genomics LTD for its expertise in genomic-assisted breeding. JIS thanks his family, Francisco Isidro Muñoz, Maria de Gracia Sánchez Sánchez, Antolin y María IS, his dearest friends "El oso" and Bradley, and "Mi collares" "JU&MA" (You are all the cheese of my macaroni). DA is indebted for all the support and help given to him by his dear parents, family, and friends, without which, none of this would be possible.

REFERENCES

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

Akdemir, D. (2017). *STPGA: Selection of Training Populations by Genetic Algorithm*. R package version 5.2.1.

Akdemir, D., Beavis, W., Fritsche-Neto, R., Singh, A. K., and Isidro-Sánchez, J. (2019). Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity* 122, 672–683. doi: 10.1038/s41437-018-0147-1

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci. Rep.* 9, 1–15. doi: 10.1038/s41598-018-38081-6

Akdemir, D., Rio, S., and y Sánchez Julio, I. (2021). Trainsel: an r package for selection of training populations. *Front. Genet.* 12:607. doi: 10.3389/fgene.2021.655287

Akdemir, D., and Sánchez, J. I. (2016). Efficient breeding by genomic mating. *Front. Genet.* 7:210. doi: 10.3389/fgene.2016.00210

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. doi: 10.1186/s12711-015-0116-6

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

Andreescu, C., Avendano, S., Brown, S. R., Hassen, A., Lamont, S. J., and Dekkers, J. C. (2007). Linkage disequilibrium in related breeding lines of chickens. *Genetics* 177, 2161–2169. doi: 10.1534/genetics.107.082206

Atanda, S. A., Olsen, M., Burgue no, J., Crossa, J., Dzidzienyo, D., Beyene, Y., et al. (2021). Maximizing efficiency of genomic selection in cimmyt's tropical maize breeding program. *Theor. Appl. Genet*. 134, 279–294. doi: 10.1007/s00122-020-03696-9

Atkinson, A., and Donev, A. (1992). *Optimum Experimental Designs*. Clarendon. Oxford.

Auinger, H.-J., Schönleben, M., Lehermeier, C., Schmidt, M., Korzun, V., Geiger, H. H., et al. (2016). Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (secale cereale l.). *Theor. Appl. Genet*. 129, 2043–2053. doi: 10.1007/s00122-016-2756-5

Ben-Sadoun, S., Rincent, R., Auzanneau, J., Oury, F., Rolland, B., Heumez, E., et al. (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. doi: 10.1007/s00122-020-03590-4

Bernardo, R. (1994). Prediction of maize single-cross performance using rflps and information from related hybrids. *Crop. Sci*. 34, 20–25. doi: 10.2135/cropsci1994.0011183X003400010003x

Berro, I., Lado, B., Nalin, R. S., Quincke, M., and Gutiérrez, L. (2019). Training population optimization for genomic selection. *Plant Genome* 12, 190028. doi: 10.3835/plantgenome2019.04.0028

Brandariz, S. P., and Bernardo, R. (2018). Maintaining the accuracy of genomewide predictions when selection has occurred in the training population. *Crop. Sci*. 58, 1226–1231. doi: 10.2135/cropsci2017.11.0682

Burstin, J., Salloignon, P., Chabert-Martinello, M., Magnin-Robert, J., Siol, M., Jacquin, F., et al. (2015). Genetic diversity and trait genomic prediction in a pea diversity panel. *BMC Genomics* 16:105. doi: 10.1186/s12864-015-1266-1

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* 6, 3733–3747. doi: 10.1534/g3.116.035410

Cericola, F., Jahoor, A., Orabi, J., Andersen, J. R., Janss, L. L., and Jensen, J. (2017). Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. a case of study in advanced wheat breeding lines. *PLoS ONE* 12:e0169606. doi: 10.1371/journal.pone.0169606

Clark, S. A., Hickey, J. M., Daetwyler, H. D., and van der Werf, J. H. (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, 10–1186. doi: 10.1186/1297-9686-44-4

Clark, S. A., Hickey, J. M., and Van der Werf, J. H. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genet. Sel Evol*. 43, 10–1186. doi: 10.1186/1297-9686-43-18

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgue no, J., Araus, J., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*. 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A., et al. (2021). The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. *Front. Plant Sci*. 12:651480 doi: 10.3389/fpls.2021.651480

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci*. 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* 3:e3395. doi: 10.1371/journal.pone.0003395

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

de los Campos, G., Hickey, J., Pong-Wong, R., Daetwyler, H., and Calus, M. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

Dekkers, J. (2007). Prediction of response to marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet*. 124, 331–341. doi: 10.1111/j.1439-0388.2007.00701.x

Dimitrijevic, A., and Horn, R. (2018). Sunflower hybrid breeding: from markers to genomic selection. *Front. Plant Sci*. 8:2238. doi: 10.3389/fpls.2017.02238

Dussert, C., Rasigni, G., Rasigni, M., Palmari, J., and Llebaria, A. (1986). Minimal spanning tree: a new approach for studying order and disorder. *Phys. Rev. B* 34:3528. doi: 10.1103/PhysRevB.34.3528

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

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Academic Press.

Fedorov, V. V., and Hackl, P. (2012). *Model-Oriented Design of Experiments, Vol. 125*. Springer Science Business Media.

Fisher, R. A. (1960). *The Design of Experiments*. New York, NY: Hafner.

Fristche-Neto, R., Akdemir, D., and Jannink, J.-L. (2018). Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. *Theor. Appl. Genet*. 131, 1153–1162. doi: 10.1007/s00122-018-3068-8

Gentle, J. E. (2006). *Random Number Generation and Monte Carlo Methods*. Springer Science Business Media.

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

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., et al. (2019). Optimal designs for genomic selection in hybrid crops. *Mol. Plant* 12, 390–401. doi: 10.1016/j.molp.2018.12.022

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

Habier, D., Fernando, R., and Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

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. doi: 10.1186/1297-9686-42-5

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

He, S., Schulthess, A. W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., et al. (2016). Genomic selection in a commercial winter wheat population. *Theor. Appl. Genet*. 129, 641–651. doi: 10.1007/s00122-015-2655-1

Heffner, E., Sorrells, M., and Jannink, J. (2009). Genomic selection for crop improvement. *Crop. Sci*. 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Heslot, N., and Feoktistov, V. (2020). Optimization of selective phenotyping and population design for genomic prediction. *J. Agric. Biol. Environ. Stat*. 25, 579–600. doi: 10.1007/s13253-020-00415-1

Isidro y Sánchez, J., Akdemir, D., and Rio, S. (2022). *Hands on Training Optimization in Genomic Selection*. Springer.

Isidro, J., Akdemir, D., and Burke, J. (2016). "Genomic selection," in *The World Wheat Book: A History of Wheat Breeding, Vol. 3, Chapter 32*, eds A. William, B. Alain, and V. G. Maarten (Paris: Lavoisier), 1001–1023.

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. doi: 10.1007/s00122-014-2418-4

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genet. Sel. Evol*. 42, 35. doi: 10.1186/1297-9686-42-35

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

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

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* 6, 3443–3453. doi: 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. doi: 10.1007/s00122-020-03722-w

Kiefer, J. (1959). Optimum experimental designs. *J. R. Stat. Soc. B* 21, 272–319.

Kiefer, J. C., Brown, L., Olkin, I., and Sacks, J. (1985). *Jack Carl Kiefer Collected Papers: Design of Experiments*. Springer.

Laloë, D. (1993). Precision and information in linear models of genetic evaluation. *Genet. Sel. Evol*. 25, 557–576. doi: 10.1186/1297-9686-25-6-557

Lee, S. H., Van Der Werf, J. H., Hayes, B. J., Goddard, M. E., and Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome snp data. *PLoS Genet*. 4:e1000231. doi: 10.1371/journal.pgen.1000231

Liu, X., Wang, H., Wang, H., Guo, Z., Xu, X., Liu, J., et al. (2018). Factors affecting genomic selection revealed by empirical evidence in maize. *Crop J*. 6, 341–352. doi: 10.1016/j.cj.2018.03.005

Longin, C. F. H., Mi, X., and Würschum, T. (2015). Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. *Theor. Appl. Genet*. 128, 1297–1306. doi: 10.1007/s00122-015-2505-1

Lopez-Cruz, M., and de Los Campos, G. (2021). Optimal breeding-value prediction using a sparse selection index. *Genetics* 210:iyab030. doi: 10.1093/genetics/iyab030

Lorenz, A. J. (2013). Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. *G3* 3, 481–491. doi: 10.1534/g3.112.004911

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. doi: 10.2135/cropsci2014.12.0827

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. doi: 10.1007/s00122-009-1166-3

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T. H. (2009). The accuracy of genomic selection in norwegian red cattle assessed by cross-validation. *Genetics* 183, 1119–1126. doi: 10.1534/genetics.109.107391

MacLeod, I., Bowman, P., Vander Jagt, C., Haile-Mariam, M., Kemper, K., Chamberlain, A., et al. (2016). Exploiting biological priors and sequence variants enhances qtl discovery and genomic prediction of complex traits. *BMC Genomics*. 17:144. doi: 10.1186/s12864-016-2443-6

Mangin, B., Bonnafous, F., Blanchet, N., Boniface, M.-C., Bret-Mestries, E., Carrère, S., et al. (2017). Genomic prediction of sunflower hybrids oil content. *Front. Plant Sci*. 8:1633. doi: 10.3389/fpls.2017.01633

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. doi: 10.1371/journal.pone.0205629

Markowitz, H. M. (1968). *Portfolio Selection: Efficient Diversification of Investments, Vol. 16*. Yale university press.

Marulanda, J. J., Mi, X., Melchinger, A. E., Xu, J.-L., Würschum, T., and Longin, C. F. H. (2016). Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. *Theor. Appl. Genet*. 129, 1901–1913. doi: 10.1007/s00122-016-2748-5

McClellan, J., Susser, E., and King, M. (2007). Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry* 190, 194–199. doi: 10.1192/bjp.bp.106.025585

McGowan, M., Wang, J., Dong, H., Liu, X., Jia, Y., Wang, X., et al. (2020). Ideas in genomic selection with the potential to transform plant molecular breeding: a review. [Epub ahead of print].

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. doi: 10.1002/csc2.20267

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Meuwissen, T., Hayes, B., and Goddard, M. (2016). Genomic selection: a paradigm shift in animal breeding. *Anim. Front*. 6, 6–14. doi: 10.2527/af.2016-0002

Momen, M., and Morota, G. (2018). Quantifying genomic connectedness and prediction accuracy from additive and non-additive gene actions. *Genet. Sel. Evolution* 50, 1–10. doi: 10.1186/s12711-018-0415-9

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

National Human Genome Research Institute (2020). *DNA Sequencing Costs: Data*. Available online at: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (accessed May 6, 2021).

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* 7, 1499–1510. doi: 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* 8, 2889–2899. doi: 10.1534/g3.118.200311

Olatoye, M. O., Clark, L. V., Labonte, N. R., Dong, H., Dwiyanti, M. S., Anzoua, K. G., et al. (2020). Training population optimization for genomic selection in miscanthus. *G3* 10, 2465–2476. doi: 10.1534/g3.120.401402

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

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

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

Pukelsheim, F. (1993). *Optimal Design of Experiments, Vol. 50*. siam.

Pukelsheim, F., and Rosenberger, J. (1993). Experimental designs for model discrimination. *J. Am. Stat. Assoc*. 88, 642–649. doi: 10.1080/01621459.1993.10476317

Reif, J. C., Zhao, Y., Würschum, T., Gowda, M., and Hahn, V. (2013). Genomic prediction of sunflower hybrid performance. *Plant Breed*. 132, 107–114. doi: 10.1111/pbr.12007

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., et al. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet*. 44, 217–220. doi: 10.1038/ng.1033

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. doi: 10.1007/s00122-017-2956-7

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., 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. doi: 10.1534/genetics.112.141473

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. *Horticulture Res*. 7, 1–14. doi: 10.1038/s41438-020-00370-5

Sarinelli, J. M., Murphy, J. P., Tyagi, P., Holland, J. B., Johnson, J. W., Mergoum, M., 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. doi: 10.1007/s00122-019-03276-6

Schrag, T., Frish, M., Dhillon, B., and Melchinger, A. (2009). Marker-based prediction of hybrid performance in maize single-crosses involving doubled haploids. *Maydica* 54, 353. doi: 10.1007/s00122-008-0934-9

Schulthess, A. W., Zhao, Y., and Reif, J. C. (2017). "Genomic selection in hybrid breeding," in *Genomic Selection for Crop Improvement* (Springer), 149–183.

Scott, M. F., Fradgley, N., Bentley, A. R., Brabbs, T., Corke, F., Gardner, K. A., et al. (2021). Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. *Genome Biol*. 22, 1–30. doi: 10.1186/s13059-021-02354-7

Seye, A., Bauland, C., Charcosset, A., and Moreau, L. (2020). Revisiting hybrid breeding designs using genomic predictions: simulations highlight the superiority of incomplete factorials between segregating families over topcross designs. *Theor. Appl. Genet*. 133, 1995–2010. doi: 10.1007/s00122-020-03573-5

Silvey, S. (2013). *Optimal Design: An Introduction to the Theory for Parameter Estimation, Vol. 1*. Springer Science Business Media.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika* 12, 1–85.

Spindel, J., Begum, H., Akdemir, D., Collard, B., Redo na, E., Jannink, J., et al. (2016). Genome-wide prediction models that incorporate de novo gwas are a powerful new tool for tropical rice improvement. *Heredity* 116, 395–408. doi: 10.1038/hdy.2015.113

Tanaka, R., and Iwata, H. (2018). Bayesian optimization for genomic selection: a method for discovering the best genotype among a large number of candidates. *Theor. Appl. Genet*. 131, 93–105. doi: 10.1007/s00122-017-2988-z

Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., et al. (2015). Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. *Front. Plant Sci*. 6:941. doi: 10.3389/fpls.2015.00941

Technow, F., Podlich, D., and Cooper, M. (2021). Back to the future: Implications of genetic complexity for hybrid breeding strategies. *G3* 5:jkab153. doi: 10.1093/g3journal/jkab153

Technow, F., Schrag, T. A., Schipprack, W., Bauer, E., Simianer, H., and Melchinger, A. E. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* 197, 1343–1355. doi: 10.1534/genetics.114.165860

Tsai, S.-F., Shen, C.-C., and Liao, C.-T. (2021). Bayesian optimization approaches for identifying the best genotype from a candidate population. *J. Agric. Biol. Environ. Stat*. 1–19. doi: 10.1007/s13253-021-00454-2

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. doi: 10.1016/j.tplants.2021.03.010

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends Plant Sci*. 10, 621–630. doi: 10.1016/j.tplants.2005.10.004

Wang, Y., Mette, M. F., Miedaner, T., Gottwald, M., Wilde, P., Reif, J. C., et al. (2014). The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. *BMC Genomics* 15:556. doi: 10.1186/1471-2164-15-556

Wientjes, Y. C., Bijma, P., Vandenplas, J., and Calus, M. P. (2017). Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. *Genetics* 207, 503–515. doi: 10.1534/genetics.117.300152

Wientjes, Y. C., Calus, M. P., Goddard, M. E., and Hayes, B. J. (2015). Impact of qtl properties on the accuracy of multi-breed genomic prediction. *Genet. Sel. Evol*. 47, 1–16. doi: 10.1186/s12711-015-0124-6

Wientjes, Y. C., Veerkamp, R. F., and Calus, M. P. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193, 621–631. doi: 10.1534/genetics.112.146290

Windhausen, V. S., Atlin, G. N., Hickey, J. M., Crossa, J., Jannink, J.-L., Sorrells, M. E., et al. (2012). Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3* 2, 1427–1436. doi: 10.1534/g3.112.003699

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., et al. (2020). Enhancing genetic gain through genomic selection: from livestock to plants. *Plant Commun*. 1:100005. doi: 10.1016/j.xplc.2019.100005

Yu, X., Leiboff, S., Li, X., Guo, T., Ronning, N., Zhang, X., et al. (2020). Genomic prediction of maize microphenotypes provides insights for optimizing selection and mining diversity. *Plant Biotechnol. J*. 18, 2456–2465. doi: 10.1111/pbi.13420

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. doi: 10.3389/fgene.2019.00189

Zhang, Y., Massel, K., Godwin, I. D., and Gao, C. (2018). Applications and potential of genome editing in crop improvement. *Genome Biol*. 19, 1–11. doi: 10.1186/s13059-018-1586-y

Zhao, Y., Mette, M., Gowda, M., Longin, C., and Reif, J. (2014). Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity* 112, 638–645. doi: 10.1038/hdy.2014.1

Zhao, Y., Mette, M. F., and Reif, J. C. (2015). Genomic selection in hybrid breeding. *Plant Breed*. 134, 1–10. doi: 10.1111/pbr.12231

Zhao, Y., Thorwarth, P., Jiang, Y., Philipp, N., Schulthess, A. W., Gils, M., et al. (2021). Unlocking big data doubled the accuracy in predicting the grain yield in hybrid wheat. *Sci. Adv*. 7:eabf9106. doi: 10.1126/sciadv.abf9106

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Sci*. 53, 802–810. doi: 10.2135/cropsci2012.08.0463

Zhong, S., Dekkers, J. C., 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. doi: 10.1534/genetics.108.098277

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.