Chapter 3

Building a Calibration Set for Genomic Prediction, Characteristics to Be Considered, and Optimization Approaches

Simon Rio, Alain Charcosset, Tristan Mary-Huard, Laurence Moreau, and Renaud Rincent

Abstract

The efficiency of genomic selection strongly depends on the prediction accuracy of the genetic merit of candidates. Numerous papers have shown that the composition of the calibration set is a key contributor to prediction accuracy. A poorly defined calibration set can result in low accuracies, whereas an optimized one can considerably increase accuracy compared to random sampling, for a same size. Alternatively, optimizing the calibration set can be a way of decreasing the costs of phenotyping by enabling similar levels of accuracy compared to random sampling but with fewer phenotypic units. We present here the different factors that have to be considered when designing a calibration set, and review the different criteria proposed in the literature. We classified these criteria into two groups: model-free criteria based on relatedness, and criteria derived from the linear mixed model. We introduce criteria targeting specific prediction objectives including the prediction of highly diverse panels, biparental families, or hybrids. We also review different ways of updating the calibration set, and different procedures for optimizing phenotyping experimental designs.

Key words Calibration population, Optimization, Prediction accuracy, Genomic selection, CDmean, PEVmean

1 Introduction

Several factors affect the accuracy of genomic prediction including (1) trait-specific characteristics like the heritability and the genetic architecture of the trait, (2) population-specific characteristics like the level of linkage disequilibrium (LD) between markers and quantitative trait loci (QTLs) and the number of effective chromosome segments (Me) segregating in the population, (3) the statistical method used to make predictions, and (4) experiment-specific

Laurence Moreau, and Renaud Rincent are co-last authors.

Nourollah Ahmadi and Jérôme Bartholomé (eds.), *Genomic Prediction of Complex Traits: Methods and Protocols*, Methods in Molecular Biology, vol. 2467, https://doi.org/10.1007/978-1-0716-2205-6_3,

characteristics such as the marker density, the size of the calibration set (CS) and the degree of genetic relationship between the CS and the predicted set (PS). The choice of the CS, i.e., reference individuals and their genotypic and phenotypic data, used to calibrate the prediction model is therefore crucial, especially when predicted traits are difficult or expensive to phenotype. In animal breeding, the pedigree-based BLUP model has been used in routine for several generations to predict the genetic value of candidates since the pioneer work of Henderson [1]. The use of genomic selection has modified the way the relationship between individuals was estimated by adding marker genotypes to pedigree information. In dairy cattle, it had little impact on the phenotypic data used to calibrate predictions for traits previously addressed in routine, but has opened the way to considering many new traits such as disease resistance, which cannot be evaluated directly for all animals [2]. In major crop species, the main focus is to select the best inbred lines that can be produced after selfing generations within large biparental families, either for their direct use as varieties or as parents of single-cross hybrid varieties. The number of candidate lines per family is often larger than the phenotyping capacities and only a small set of them can be evaluated in different environments to evaluate their adaptation to various field conditions. In this context, pedigree information is not useful to identify the best individuals [3] within a given family and pedigree-BLUP based on historical data has therefore not been broadly used in crop breeding. With genomic selection, the differences in genetic covariance between pairs of individuals from a biparental family can be accounted for in the model, unlike with pedigree data. Phenotypes are no longer used exclusively as proxies of the genetic values of candidate individuals but also to train a predictive model involving molecular markers as predictors, which potentially modifies phenotyping strategies. The advent of genomic selection clearly opens new possibilities for improving the breeding efficiency of both animal and plant species but raises the key question of how to define the best CS, especially in plants. In the first part of this chapter, we provide some general guidelines to be considered for this purpose. These guidelines are illustrated with examples and their biological bases are discussed. In a second part, we present the different approaches that have been proposed to optimize the reference population. In the last part we show some applications of reference population optimization, depending on the prediction objective. Even if genomic predictions are also used in human genetics, this chapter focuses on the application of genomic predictions for breeding objectives in animal and plant species, with more emphasis on plants where the issue of the optimization of the reference population has been the most extensively considered.

2 Impact of the Composition of the Calibration Set on the Accuracy of Genomic Prediction

2.1 Calibration and Predicted Individuals Ideally Originate from the Same Population

Most genomic prediction models such as GBLUP or in the declination of Bayesian models [4–6] assume that CS and PS individuals are drawn from the same population. As described in the examples below, this hypothesis is often violated and comes at the cost of a reduced accuracy.

The first case is the application of genomic prediction in a population stratified into genetic groups. This scenario has been the subject of several studies to investigate (1) to which extent a generic CS can be efficient to predict over a wide range of genetic diversity, (2) or to which extent genetic groups with limited resources can benefit from the data originating from other genetic groups with more resources. In general, when a genomic prediction model is trained on one genetic group to perform predictions in another genetic group, the accuracy tends to be lower than what can be achieved within each group. This phenomenon has been illustrated in many animal and plant species including dairy cattle [7–9], sheep [10], maize [11, 12], soybean [13], barley [14], oat [15], or rice [16]. This case can also be extended to the prediction between families, as shown in mice [17], maize [18, 19], wheat [20], barley [21], or triticale [22]. In the worst case, the addition of individuals to the CS that are genetically distant from the PS can lead to a deterioration in the accuracy, as shown for instance in barley [23] and maize [24].

Beyond genetic groups and families, differences in type of genetic material (e.g., purebred and crossbred) between the CS and PS can lead to reduced accuracies compared to what can be achieved when the CS and the PS are of the same type of genetic material. Examples are the prediction of crossbred individuals using a purebred parental CS, as shown in pig [25], the prediction of maize admixed population between two heterotic groups using one of the parental population as CS [26], or the prediction of interspecific hybrids, as shown in Miscanthus [27].

Finally, the CS and PS may be drawn from the same population but different breeding generations. This scenario is very common when cycles of selection are done solely based on predicted genetic values to shorten breeding cycles or when candidates are preselected to reduce their number and limit phenotyping costs. Decrease in accuracies can generally be observed over cycles when prediction models are not updated with data from the selected generation, as shown using simulations [28], or real data in pig [29], sugar beet [30], alfalfa [31], maize [32], wheat [33], barley [34], and rye [35]. Note that genotype-by-environment interactions can also contribute to the drop in accuracy when the CS and PS are not evaluated in the same environments.

As illustrated with these examples, a decrease in accuracies can be expected when CS and PS individuals come from different populations or breeding generations, and this decrease may result from different factors that are presented in the following subsections.

2.1.1 LD Between Markers and QTLs Can Be Different Between Populations

Since the early development of genomic prediction, the LD between molecular markers and QTLs has been identified as a major factor of accuracy [4]. It can be defined as the nonindependence between alleles at different loci on the same gamete. In general, LD between markers and QTLs is assumed to be homogeneous within a population, but it may vary when the population is stratified, which affects the success of across-breed genomic prediction [7, 36]. Differences in LD between markers and QTLs may indeed lead to differences in effects estimated at markers, impacting the accuracy when one population is used to predict another. LD between two loci is a function of recombination rate, minor allele frequency (MAF) at both loci and effective population size [37]. Differences in MAF and effective population size are very common whenever a population is stratified into groups [38], but differences in recombination rate can also be observed between genetic groups, as shown in maize [39]. Differences in LD extent estimated with markers have been observed among populations in dairy and beef cattle [40, 41], pig [42], chicken [43], maize [44, 45], or wheat [46]. Differences in the sign of the correlation between the allelic state of loci pairs can also be found and are referred to as differences in the linkage phase [40, 45]. In presence of dense genotyping, the effect of differences in LD between populations on the accuracy is expected to be minimized [40, 47], as most QTLs are expected to be in high LD with at least one shared marker in both populations. The ideal situation is that only causal loci are captured by the genotyping.

2.1.2 QTL Allele Frequencies Can Be Different Between Populations

In addition to differences in LD, genomic prediction accuracy across populations can be affected by differences in QTL allele frequencies. The most extreme scenario consists of a QTL for which an allele is fixed in the CS but is segregating in the PS. The effect of such a QTL cannot be estimated using the CS, and the genetic variance that it explains will not be accounted for in the prediction [48, 49]. In the context of prediction across biparental populations, Schopp et al. [50] proposed to adapt the formula of Daetwyler et al. [48], used to forecast the accuracy, by including a new term: the proportion of markers that segregates in both the CS and PS relative to the total number of markers segregating in the PS. Based on simulations, they showed that this criterion computed using markers is a good approximation of the equivalent criterion based on QTLs when the marker density is sufficiently large, and is critical for the accuracy of predictions across families. In more

complex populations, one can estimate the genetic differentiation (F_{ST}) using markers, as an indication of how QTL alleles frequency differ between the CS and PS, which was shown to be negatively related to the accuracy of genomic prediction [51].

The consistency of alleles frequency between the CS and PS can be extended to the consistency of the frequencies of genotypic states at QTLs (i.e., the two homozygous states and the heterozygous state for a biallelic QTL). One example consists of dominance effects at QTLs that can only be accounted for if some individuals present a heterozygous state in the CS [52]. This phenomenon can explain the decrease in accuracy observed when predicting crossbred individuals using purebred individuals for traits with substantial dominance effects [25].

2.1.3 QTL Allele Effects Can Be Different Between Populations

In most genomic prediction models, the effects of QTLs are assumed to be consistent between the CS and PS. But this assumption can be violated when the CS and the PS are drawn from different populations. "Statistical" additive effects reflect the average effect of substituting an allele with the alternative allele in the population and are implicitly or explicitly taken into account in most models, including GBLUP. "Functional" dominance and epistatic effects at QTLs contribute to the statistical additive effect along with the functional additive effect, but, unlike the latter, their contributions depend on allele frequencies [52–54]. From this phenomenon emerges the concept of genetic correlation between populations that aims at quantifying this difference in statistical additive effects [55, 56]. Practically, note that genetic correlations are often estimated using markers and then also include the heterogeneity generated by differences in LD between markers and QTLs.

2.2 Genetic Relationships Between Calibration and Predicted Individuals Are Needed

In the present chapter, we define genetic relationships as standardized covariances between individuals relative to the genetic components of traits. In this context, they are defined at QTLs (i.e., at causal loci) level and reflect the sharing of alleles at these loci. Genetic relationships notably include additive genetic relationships (AGRs) that describe relationships between individuals for additive allele effects. As causal loci are generally unknown, AGRs must be estimated based on the pedigree or by using markers. From a pedigree perspective, the sharing of alleles at QTLs is considered to result from their inheritance from a common ancestor. Those alleles are characterized as identical-by-descent (IBD, see Thompson [57] for a review), with IBD being defined relative to a founder population as a reference starting point of the pedigree. In this context, the coefficients of the pedigree relationship matrix (PRM) consist of expected AGRs conditional on pedigree information. Since the advent of molecular markers, AGRs can be estimated using the genomic relationship matrix (GRM) in GBLUP, often allowing better estimates of additive genetic variances compared to those obtained using the PRM (see Speed and balding [58] for a review).

In the early developments of genomic prediction, the ancestral LD between markers and QTLs was suspected of contributing alone to the genomic prediction accuracy [4]. Ancestral LD can be defined as statistical dependencies between loci that already existed within the population founders of the pedigree, which were generated by ancestral evolutionary forces. Thus, it does not characterize additional dependencies between loci that arise from the pedigree relationships between individuals. When individuals are not related by pedigree, the GRM can only capture AGRs through ancestral LD, as illustrated in Fig. 1A. However, when CS and PS individuals are related by pedigree, the GRM captures AGRs even in absence of ancestral LD between markers and QTLs [5, 59–61], as illustrated in Fig. 1B. In this scenario, the GRM describes IBD at markers and can be considered as a proxy for the PRM. It explains why nonnull accuracies can be obtained when applying genomic predictions with only few markers. The contribution of ancestral LD and pedigree relationships to the accuracy depends on the genomic prediction model. Variable selection approaches like LASSO, or Bayesian approaches like Bayes-B tend to better exploit ancestral LD than GBLUP to make predictions [60, 62, 63]. In addition, for a given pedigree structure, the relative contribution of ancestral LD and pedigree relationships to the prediction accuracy depends on population size: pedigree relationships tend to have a greater effect than ancestral LD on the accuracy for CS of small size, and conversely for CS of large size [63–65]. This contribution of ancestral LD relative to pedigree relationships is an important parameter to consider when applying genomic prediction between genetically distant CS and PS, as the accuracy due to pedigree relationships will drop more quickly than that due to ancestral LD with decreasing relatedness [59, 62].

In addition to pedigree relationships and ancestral LD between markers and QTLs, Habier et al. [61] also demonstrated that the GRM captures cosegregations between markers and QTLs. Cosegregations characterize the nonrandom association of alleles between linked loci that can be observed within the individuals of a given family of the pedigree. It can be distinguished from ancestral LD that characterizes the nonrandom association between alleles of different loci that were already established in the founders of the pedigree. In the absence of ancestral LD between markers and QTLs, marker alleles will, nevertheless, cosegregate with QTL alleles to which they are physically linked when new individuals are generated, as illustrated in Fig. 1C. This information will be accounted for in the GRM and will contribute to the genomic prediction accuracy. Note that cosegregations help to describe genetic covariances between individuals of the same family. When several families are pooled into a common CS, differences in linkage phase can be observed and may considerably limit the contribution of cosegregations to the accuracy [24].

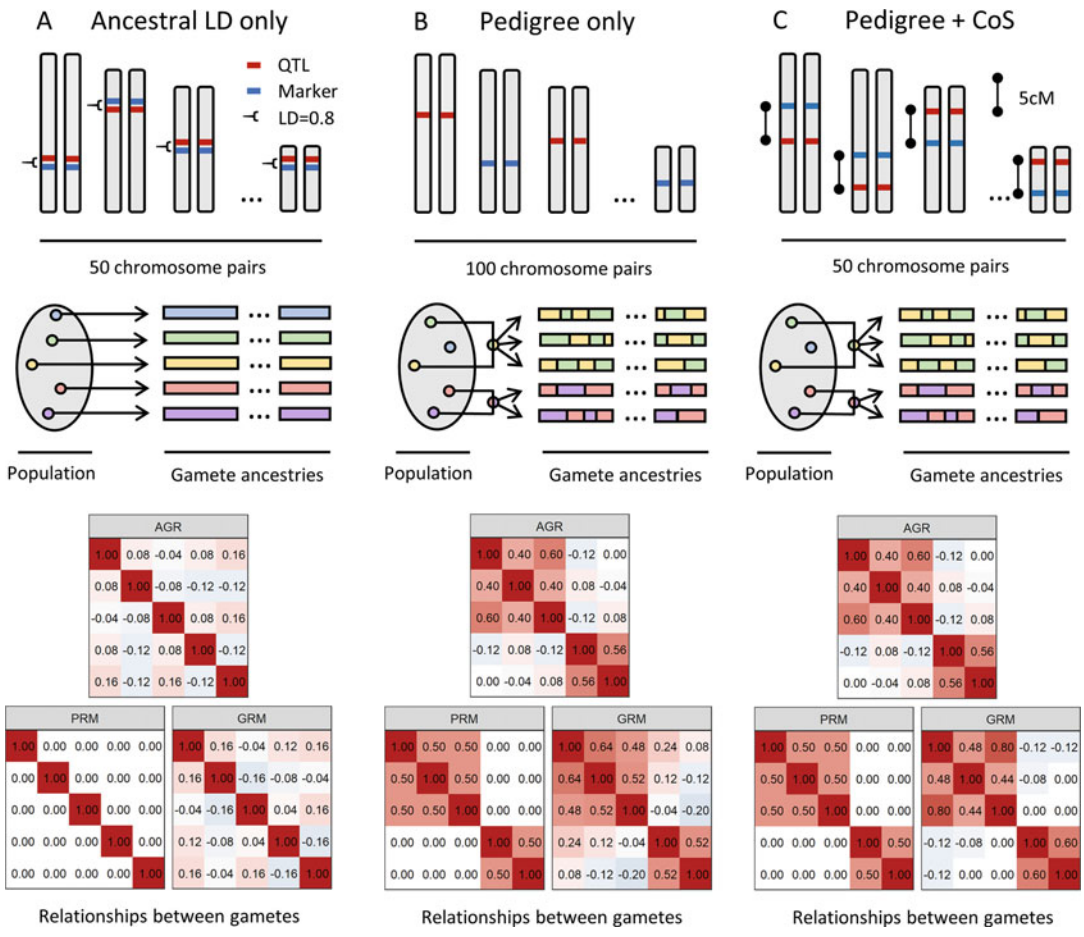

Fig. 1 Hypothetical scenarios adapted from Habier et al. [61] illustrating different types of information captured by the genomic relationship matrix (GRM): (**A**) Ancestral LD only, (**B**) Pedigree only, and (**C**) Pedigree + cosegregations (CoS). For each scenario, 50 QTLs and 50 markers are considered with minor allele frequency of 0.5. In scenario (**A**), QTLs and markers are assigned in pairs to chromosomes (a single pair per chromosome with each loci pair being in LD = 0.8). In scenario (**B**), QTLs and markers are assigned to different chromosomes (a single locus per chromosome). In scenario (**C**), QTLs and markers are assigned in pairs to chromosomes (a single pair per chromosome with each loci pair being genetically distant by 5 cM but not in LD). In scenario (**A**), gametes are generated independently from different founders of the population, while they are generated from individuals resulting from the crossing of founders for scenarios (**B**) and (**C**). The additive genetic relationship (AGR) between gametes is computed by applying the formula of [124] to QTLs using simulated allele frequencies. The pedigree relationship matrix (PRM) between gametes is constructed by assigning a coefficient of 0.5 between gametes originating from the same individuals, and 0 otherwise. The GRM is calculated like the AGR but using markers. In scenario (**A**), the GRM can estimate the AGR using ancestral LD, even in absence of pedigree relationship between gametes. In scenario (**B**), the GRM can estimate the AGR by tracing pedigree relationships between gametes (like the PRM), even in absence of ancestral LD between markers and QTLs. In scenario (**C**), the GRM can better estimate the AGR within a family of gametes compared to the GRM in scenario (**A**) and the PRM, thanks to cosegregations between QTLs and marker alleles that are physically linked on chromosomes. Note that we considered haploid gametes to simplify the schematic representation but those concepts can be generalized to relationships between diploid individuals

Several studies have shown that the accuracy of genomic prediction is linked to the AGRs between CS and PS individuals. Based on simulation, Pszczola et al. [66] established the link between the deterministic reliability of genomic prediction and the average squared genomic relationship coefficient between CS and PS individuals. Habier et al. [60] illustrated that the accuracy of genomic prediction increased with increasing a-max in dairy cattle, where a-max is defined as the maximum pedigree relationship coefficient between CS and PS individuals. This result was confirmed in maize [67] and oil palm [68]. More generally, the need for close pedigree relationships between CS and PS individuals has been illustrated in several species like in mice [17].

In addition to AGRs, other types of genetic relationships can be modeled to improve genomic prediction accuracies such as dominance and epistatic genetic relationships. Like AGRs, these other relationships directly reflect the sharing of alleles at QTLs and can be estimated using the pedigree [57] or using markers [52, 69]. However, they are often not accounted for in genomic prediction models, as they generally have a limited contribution to the overall genetic variance, except for specific applications such as the prediction of hybrids (*see* Subheading 4.3).

2.3 Calibration Set Should Be As Large as Possible

When building a CS, increasing the number of individuals is generally beneficial. The importance of CS size has been shown theoretically using deterministic equations of the accuracy of genomic prediction [48, 70–74]. They showed that the population size should be large enough to properly estimate the effect of each of the effective chromosome segments that segregate in the population (quantified by their number Me), in particular for low heritability traits. The effect of the CS size on genomic prediction accuracy has been illustrated experimentally in plants [62, 75–77] and animal species [78, 79], as well as in Human [80]. However, one should keep in mind that increasing the number of individuals should be done with caution if additional individuals are genetically distant from the PS individuals, as mentioned in the previous subsection. There is also a compromise to be found between the number of phenotyped individuals and the accuracy of the phenotyping that can be increased in plants by increasing the number of observations per individual (*see* discussion in Subheading 4.4).

2.4 Genetic Relationships Between CS Individuals Should Be Limited

Finally, it is generally admitted that genetic relationships among individuals should be limited within the CS. This idea is related to the common assumption that, in genomic prediction, experimental designs should aim at replicating alleles rather than individuals [81]. Because individuals with high degrees of genetic relationship can be considered as partial replicates and somewhat redundant, including them all may not be the best allocation of resources regarding genomic prediction accuracy. Based on simulations,

Pszczola et al. [28, 66] have shown that average reliabilities of genomic prediction decreased with increasing genetic relationships within the CS. These results were confirmed in dairy cattle [82]. However, limiting the genetic relationships between CS individuals is not sufficient to maximize the accuracy, as maximizing genetic relationships between CS and PS individuals is also important.

3 Methods to Optimize the Composition of the Calibration Set

Considering all the factors affecting predictive ability mentioned above, optimizing the composition of CS is not a simple task. We can, however, suppose that when CS and PS cover the same genetic space, they will have similar LD patterns, same segregating QTLs and a high genetic relationship, which are the main drivers of predictability. Numerous criteria have been proposed in the literature to optimize the composition of the CS. They can be grouped into two classes. The first class of approaches consists in identifying optimized CS based on model-free relatedness criteria. The second class of approaches is directly based on the genomic selection (GS) statistical model. They mostly rely on GBLUP, which is one of the reference GS models. They consist in defining CS by optimizing some criteria derived from the linear mixed model: the (generalized) Prediction Error Variance (PEV) and Coefficient of Determination (CD), or the expected Pearson correlation between predicted and observed values (r). Each criterion has advantages and drawbacks related to their efficiency to maximize predictive ability, their computational demand, and their ability to optimize the CS prior to phenotyping. A brief description of the methods/criteria (including references and, if available, scripts or tools to implement them) is presented in Table 1. In this part, we will review the two classes of criteria. The specific questions of predicting biparental families, optimizing or updating the CS when phenotypes are available, optimizing CS for hybrids, and optimizing experimental designs will be reviewed in Subheading 4.

3.1 Model-Free Optimization Criteria Based on Genetic Distances Between Individuals

As mentioned above, one of the main objectives when designing a CS is to ensure that the genetic space covered by the PS is well captured by the CS. Relatedness being a key contributor to prediction accuracy, different relatedness-based criteria were proposed to optimize the composition of the CS.

3.1.1 Optimization Based on Genetic Diversity Within the CS

A first way of optimizing the composition of the CS using relatedness is to minimize genetic similarity within the CS. The underlying idea is that genetic similarity within the CS can be seen as partial redundancy, and so CS including related individuals would be less

Table 1
Main methods and criteria to optimize the composition of the calibration set in genomic selection. Methods are grouped in two categories (Type) depending on whether they rely on a statistical model ("model based") or not ("model free") and depending on the target objective (partition a genotyped population into CS/PS (A), optimize a CS for a given PS (B), or subsample historic data to predict a given PS (C))

Name/criterion	Description	Type	Objective	Specificities	Main references	Scripts/packages
Amean and Amax	Minimize the average or the maximum of the relationship coefficients within the CS	Model-free	A		[83]	Upon request
Uniform coverage of the target genetic space	A geometric approach ensuring that no close relatives are included in the CS	Model-free	A		[84]	
Fast and unique representative subset selection (FURS)	Method based on graphical networks, which principle is to identify nodes (individuals) with highest degree centrality	Model-free	A		[85]	https://github.com/TingtingGuo0722/OptimalDesign
Partitioning around Medoids (PAM)	Individuals are grouped into clusters and representatives of each group are identified	Model-free	A		[85]	https://github.com/TingtingGuo0722/OptimalDesign
Gmean, Avg_GRM, and Crit_Kin	Maximize a criterion based on the genetic similarity between the CS and the PS	Model-free	A or B		[23, 86, 87]	
Gmax	The objective is that each predicted individual has a close relative in the CS	Model-free	A or B		Unpublished	
Stratified sampling	It ensures that each group of structure is well represented in the CS	Model-free	A	Considers population structure	[84, 87, 84, 90–93]	

PEVmean	Minimize the average generalized prediction error variance (PEV) of the PS	Model-based	A or B		[83, 99, 145, 122][b]	Upon request [83], STPGA [99]
CDmean	Maximize the average generalized expected reliability (CD) of the PS	Model-based	A or B		[83, 91, 99, 145, 98][b]	Upon request [83]
CDmin	Maximize the minimal expected reliability (CD) of the PS	Model-based	A or B		[121]	TrainSel[a], https://github.com/TheRocinante-lab/TrainSel
CDpop	Maximize the within population generalized expected reliability	Model-based	A or B	Considers population structure/multiparental populations	[87]	Upon request
CDmulti	Maximize the generalized expected reliability for multitrait or multienvironment models	Model-based	A or B	Multitrait, multienvironment	[116]	Upon request
Expected predictive ability or accuracy (r)	Maximize the expected predictive ability or accuracy	Model-based	A or B		[101]	TSDFGS
EthAcc	Maximize the expected accuracy for a given genetic architecture	Model-based	C	Takes QTL information into account	[118, 119]	Supplementary materials
PEVmean1	A CS is specifically designed for each PS individual by minimizing its individual PEV	Model-based	C	Defines a CS specific to each predicted individual	[103]	
Sparse selection index	Optimization based on a sparse selection index	Model-based	C	Defines a CS specific to each predicted individual	[136]	https://github.com/MarcooLopez/SFSI

[a] TrainSel allows different types of criteria to be implemented (optimization function can be provided by users) beyond the built-in CDmin criterion presented in the publication
[b] Individual CD or PEV considered instead of the generalized ones

informative than more diverse CS (*see* Subheading 1). This can be for instance done by minimizing the average or the maximum of the relationship coefficients within the CS [83]. A similar approach was proposed by Bustos-Korts et al. [84] to design a CS leading to a uniform coverage of the target genetic space. This is based on a geometric approach that ensures that no close relatives are included in the CS. Guo et al. [85] proposed a method called Partitioning around Medoids (PAM), in which individuals are grouped into clusters and representatives of each group are identified. They also proposed Fast and Unique Representative Subset Selection (FURS), which is a sampling method based on graphical networks. The principle of FURS is to identify nodes (individuals) with highest degree of centrality. These criteria are best adapted to identify a CS among a set of candidates that will be used to predict the performance of the remaining candidates (like the scenario in Fig. 2A). They cannot be used to optimize a CS for an independent PS.

CS optimization based on these criteria has led to a higher prediction accuracy than random sampling [84, 85], but they do not directly consider the genetic relatedness between the CS and the PS. If most of the PS individuals are present in a small part of the genetic space, it is important to have many CS individuals in this part, even if it leads to a low diversity in the CS. In other words, it is important to weigh the different parts of the genetic space according to the distribution of the PS individuals, the optimal CS being not necessarily the one with the highest genetic diversity.

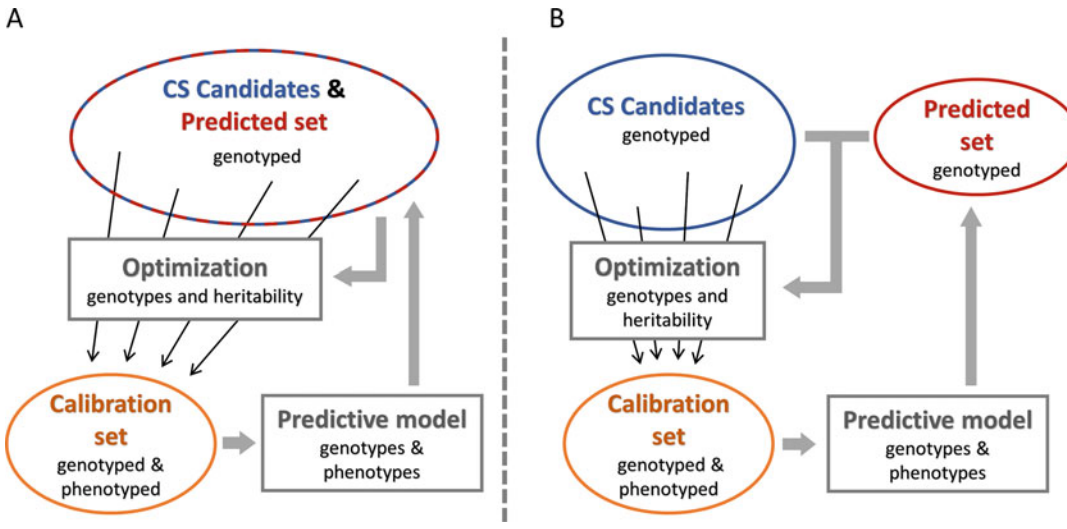

Fig. 2 Comparison of two standard scenarios that can be considered for the optimization of the calibration set (CS) based on PEVmean and CDmean. In (**A**) a single population is split into a CS and a predicted set (PS) with the objective of optimizing the CS for best predicting non phenotyped individuals (PS), while in (**B**) the set of CS candidates and the PS (both genotyped) are distinct. In both scenarios, the PEVmean and the CDmean criteria are computed directly for the PS individuals

3.1.2 Optimization Based on Genetic Relatedness Between the CS and the PS

The genetic relatedness between the CS and the PS is taken into account by criteria Gmean [23], Avg_GRM [86], and Crit_Kin [87]. For Gmean and Avg_GRM, candidates to the CS are ranked according to their average genetic relatedness to the PS. The individuals with the highest average are included in the CS. Roth et al. [88] and Berro et al. [89] proposed similar approaches in which individuals are ranked according to their maximum or median genetic relatedness to the PS. For Crit_Kin the average of the relationship coefficients between the CS and the PS is maximized. These criteria generally resulted in higher predictive ability than random sampling [23, 86–88, 90]. Contrary to the previous criteria (Subheading 3.1.1 above), Gmean, Avg_GRM and Crit_Kin take into account genetic relatedness between CS and PS, but do not consider redundancy within the CS. To design an optimal CS, it seems important to balance the two aspects.

Another criterion that has, to our knowledge, not yet been tested in the literature and that could help reaching this balance, is Gmax, the maximum relatedness coefficient between a given PS individual and the CS individuals. As prediction accuracy decreases with the genetic distance between the CS and the PS (see above), it seems interesting to ensure that each PS individual has a close relative in the CS, as illustrated in Clark et al. [63] and Habier et al. [59]. One criterion could be to identify a CS maximizing the average of the Gmax of all PS individuals. This criterion seems promising, as its maximization will result in similar genetic dispersion for CS and PS. We can indeed think that the optimal CS is the PS itself, and Gmax will help identify a CS as close as possible to this optimum.

3.1.3 Taking Population Structure into Account

In case of population structure, all the abovementioned criteria derived from relatedness coefficients could also be useful as population structure is partially captured by genetic relatedness. But in the case of strong structure, it may be necessary to directly take it into account. It was proposed in numerous studies to define the CS by stratified sampling. These algorithms ensure that each group is well represented in the CS, possibly taking the size of each group into account [84, 87, 90–93]. The efficiency of these approaches to increase predictive ability was disappointing, as they were often not better (sometimes worse) than random sampling, and most of the time not as good as relatedness-based criteria not taking structure into account. This is probably due to the fact that they only rely on population structure and do not consider relatedness within the groups. Their efficiency, however, increases with the importance of the structuration [91]. Optimizing CS by combining information on population structure and relatedness seems an interesting alternative strategy to achieve higher accuracy. The specific and extreme case of predicting a population stratified into biparental populations will be discussed below.

3.2 Optimization Using "Model-Based" Criteria Derived from the Mixed Model Theory (PEV, CD, r)

The abovementioned criteria are model-free in the sense that they do not rely on any genomic prediction model. We can, nevertheless, think that if a GS model results in accurate predictions, working with theoretical criteria derived from the same model could be valuable. One of the reference and most efficient GS model is the GBLUP mixed model. It is particularly adapted to polygenic traits because it relies on the infinitesimal model. Analytical developments were made within the mixed model theory to derive criteria related to the expected predictive ability of the model before any phenotyping. In this section we introduce these criteria and how they were used to optimize CS.

3.2.1 CS Optimization Using the Prediction Error Variance (PEV) or the Coefficient of Determination (CD)

Rincent et al. [83] proposed to use the generalized Prediction Error Variance (PEV(c)) or the generalized expected reliability (generalized Coefficient of Determination, CD(c)) of contrast **c** to optimize the composition of CS in genomic prediction.

In animal breeding, PEV(c) or CD(c) were first proposed to track disconnectedness in experimental designs [94, 95]. The contrast **c** indicates in which comparison we are interested in. If one wants to consider the comparison between the prediction of individual 1 and the prediction of individual 2 in a set of four individuals, then $\boldsymbol{c}' = [1 \quad -1 \quad 0 \quad 0]$. If one wants to compare a group of individuals (1 and 2) with another group of individuals (3 and 4), then $\boldsymbol{c}' = [1 \quad 1 \quad -1 \quad -1]$. The sum of the contrast elements always has to be null. Contrary to plants, animals cannot be replicated in different environments, and so the comparison of animals of different years or different herds can be a problem. The genetic relatedness between individuals obtained from the pedigree can be used to connect the different management units. Taking into account this connectivity is important to ensure that the comparison between animals is reliable. PEV(c) and CD(c) were initially used to optimize experimental designs (repartition of animals in different herds) to make the comparisons reliable [94, 95]. More recently, it was applied to models relying on realized relationship matrices based on marker information [96, 97], possibly in the presence of nonadditive effects [98].

The generalized PEV and CD are derived from the GBLUP model, with: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$, where $\boldsymbol{y}$ is a vector of phenotypes, $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{u}$ is a vector of random genetic values (polygenic effect) and $\boldsymbol{e}$ is the vector of errors. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices. The variance of the random effect $\boldsymbol{u}$ is: $\text{var}(\boldsymbol{u}) = \boldsymbol{A}\sigma^2_g$, where $\boldsymbol{A}$ is the relationship matrix (realized relationship matrix in the context of genomic prediction) and σ^2_g is the additive genetic variance in the population. The variance of the errors $\boldsymbol{e}$ is: $\text{var}(\boldsymbol{e}) = \boldsymbol{I}\sigma^2_e$, where $\boldsymbol{I}$ is the identity matrix.

The PEV of any contrast $\boldsymbol{c}$ of predicted genetic values can be equivalently calculated as:

$$PEV(c) = \frac{\mathrm{var}(c'u - c'u)}{c'c},$$

$$PEV(c) = \frac{c'\left(Z'MZ + \lambda A^{-1}\right)^{-1} c * \sigma_e^2}{c'c},$$

$$PEV(c) = \frac{c'(A - AZ'M_2ZA)c * \sigma_g^2}{c'c},$$

where c is a contrast, $\hat{u}$ is the BLUP of u, M is an orthogonal projector on the subspace spanned by the columns of X: $M = I - X(X'X)^{-}X'$ and $(X'X)^{-}$ is a generalized inverse of $X'X$ [94], $M_2 = \tilde{\Sigma}^{-1} - \tilde{\Sigma}^{-1}X(X'\tilde{\Sigma}^{-1}X)^{-1}X'\tilde{\Sigma}^{-1}$, $\tilde{\Sigma} = ZAZ' + \lambda I$ is the phenotypic covariance matrix scaled to some variance ratio and $\lambda = \sigma_e^2/\sigma_g^2$. The last expression of PEV(c) has the advantage of being computationally much more efficient when the size of the CS is small in comparison to the total number of individuals considered. PEV(c) is influenced by the genetic distance between the compared individuals and by the expected amount of information brought by the experiment on the compared individuals. A low PEV of a contrast between two individuals can be due to their close genetic similarity, or to the important amount of information brought by the experiment on the given comparison (e.g., the two individuals are related to many CS individuals meaning that their predictions will be precise).

The generalized CD [94] is defined as the squared correlation between the true and the predicted contrast of genetic values, and is computed as:

$$CD(c) = \mathrm{cor}(c'u, c'u)^2,$$

$$CD(c) = \frac{c'\left(A - \lambda\left(Z'MZ + \lambda A^{-1}\right)^{-1}\right)c}{c'Ac},$$

$$CD(c) = \frac{c'(AZ'M_2ZA)c}{c'Ac}.$$

As for PEV(c) the last expression of CD(c) is computationally more efficient, because of the reduced size of the matrix to be inverted when the number of observations is smaller than the total number of individuals. The CD(c) is equivalent to the expected reliability of the contrast. It takes values between 0 and 1, a CD(c) close to 0 means that the prediction of the contrast is not reliable, whereas a CD(c) close to 1 means that the prediction is highly reliable. The generalized CD(c) is equal to $CD(c) = 1 - \frac{PEV(c)}{c'Ac\sigma_g^2}$. As a result, the CD(c) increases with diminishing PEV (c) and with increasing genetic distance between individuals involved in the targeted contrast. An increase of the genetic distance will indeed increase the genetic variance of the contrast. Note

that if c is replaced by a vector of 0 and a single 1, the resulting CD is no longer the generalized CD of a contrast but the individual CD of the corresponding individual.

Rincent et al. [83] first proposed to use PEV(c) and CD(c) to optimize the composition of the CS in genomic prediction. As CD (c) is the expected reliability of a given contrast **c**, it is a criterion of choice to maximize prediction accuracy by optimizing the composition of the CS. The main aim of genomic selection is indeed to discriminate between individuals based on their predicted breeding values. As shown above, the computation of these criteria only requires the kinship matrix and the ratio of the error and genetic variances (λ) that can be chosen based on prior knowledge. No phenotypic information is required, so the optimization of the CS can be done prior to any phenotyping. The optimization was only marginally affected by λ in Rincent et al. [83] and Akdemir et al. [99], which means that the CS optimizing PEV(c) or CD(c) is supposed to be efficient for any polygenic trait.

Rincent et al. [83] have proposed the criteria $PEVmean = \frac{1}{N_{PS}}\sum_{i=1}^{N_{PS}} PEV(c_i)$ and $CDmean = \frac{1}{N_{PS}}\sum_{i=1}^{N_{PS}} CD(c_i)$, where c_i is the contrast between PS individual i and the mean of the population, and N_{PS} is the size of the PS. CDmean (PEVmean) is the average of the CD(c) (PEV(c)) of the individuals in the PS considering a given CS. CDmean is expected to be better than PEVmean for improving GS accuracy, as illustrated in Rincent et al. [83] and Isidro et al. [91], since the CD(c) is related to the ability to discriminate individuals. By maximizing CDmean of the PS, we define a CS able to discriminate each predicted individual from the average population, so that we are able to reliably identify the best (or the worst) individuals. Using two maize diversity panels, Rincent et al. [83] considered a case when only part of a population could be phenotyped so the CS was optimized in order to predict the non phenotyped individuals (PS), and a case when the CS was optimized in order to predict a predetermined PS (Fig. 2). They showed that a considerable increase of prediction accuracy could be reached by optimizing the CS with PEVmean and even more with CDmean in comparison to randomly sampled CS. From another perspective, PEVmean and CDmean based CS enabled the same prediction accuracy as random CS with twice as less phenotyped individuals. One key point with these criteria is that they take into account kinship between all individuals (CS and PS), and therefore result in the sampling of an optimized CS specific to a given PS. As a result, it is highly recommended to optimize the PEVmean or CDmean of the predicted individuals [83, 87, 99, 100] rather than those of the individuals composing the CS [91, 101]. These criteria have been tested and validated in different species such as maize [83, 86, 87, 93], palm tree [68], wheat [102–104], barley [90], oat [15], cassava [105, 106], miscanthus [27], Arabidopsis

[99], apple tree [88], and peas [107] in populations of various levels of relatedness. CDmean led to prediction accuracies at least as good as those obtained with model-free criteria [83, 86, 87, 91, 93] with some exceptions [88–90, 108]. Note that the contrasts are flexible and can be adapted to address specific prediction objectives. For instance, in the context of biparental families, different contrasts have to be defined if one is interested in comparing families or individuals within families (see criterion CDpop below). In case of strong population structure, it can be necessary to adapt these criteria [87, 91, 101]. Isidro et al. [91] have proposed the stratified CDmean maximizing the CDmean within each group. This criterion did not improve prediction accuracy in comparison to CDmean. This may be explained by the fact that CDmean takes population structure into account as long as it is captured by the kinship matrix. One of the strengths of PEV(c) and CD(c) is that they can be adapted to address specific prediction objectives (e.g., scenarios a and b in Fig. 2) by adapting the contrasts. It can be used to optimize CS for a given PS (Fig. 2A), or to select the best CS within a population that can only be partially phenotyped, the remaining individuals being predicted (Fig. 2B). Rincent et al. [87] proposed to adapt the contrasts to take population structure into account. In this study based on connected biparental populations, new criteria were proposed to maximize prediction accuracy within each population (CDpop), or the global accuracy not taking population structure into account (CDmean). They showed that the definition of the contrasts could be adapted to specifically address each prediction objective (see below). Examples of CS optimized with CDmean or CDpop are presented in Fig. 3.

3.2.2 Multitrait CS Optimization with CDmulti

Genomic prediction models can be adapted to take into account multitraits and multienvironments in a same statistical model. This was shown to increase prediction accuracy in particular when a low-cost secondary trait is measured on the PS, i.e. trait-assisted prediction [109–111], or when all PS individuals are phenotyped in at least one environment in a multienvironment trial, i.e., sparse testing [112–116]. In these situations, the partition between CS and PS is not as clear as in the previous paragraphs, as some of the PS individuals are partially observed (phenotyped for a secondary trait, and/or in some of the environments). The optimization is more complex, as the experimental design involves more than one trait or environment. The underlying model is $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$, in which $\boldsymbol{y}$ is a vector of phenotypes concatenating the different traits, $\boldsymbol{u}$ is the corresponding vector of multitrait polygenic effects, and $\boldsymbol{e}$ is the vector of errors, $var(\boldsymbol{u}) = \boldsymbol{\Sigma_a} \otimes \boldsymbol{A}$ and $var(\boldsymbol{e}) = \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{I}$, with $\boldsymbol{\Sigma_a}$ the matrix of genetic variance/covariance between traits, and $\boldsymbol{\Sigma_\epsilon}$ the matrix of error variance/covariance between traits. Generalized CD can be derived from this model [117] to compute the expected

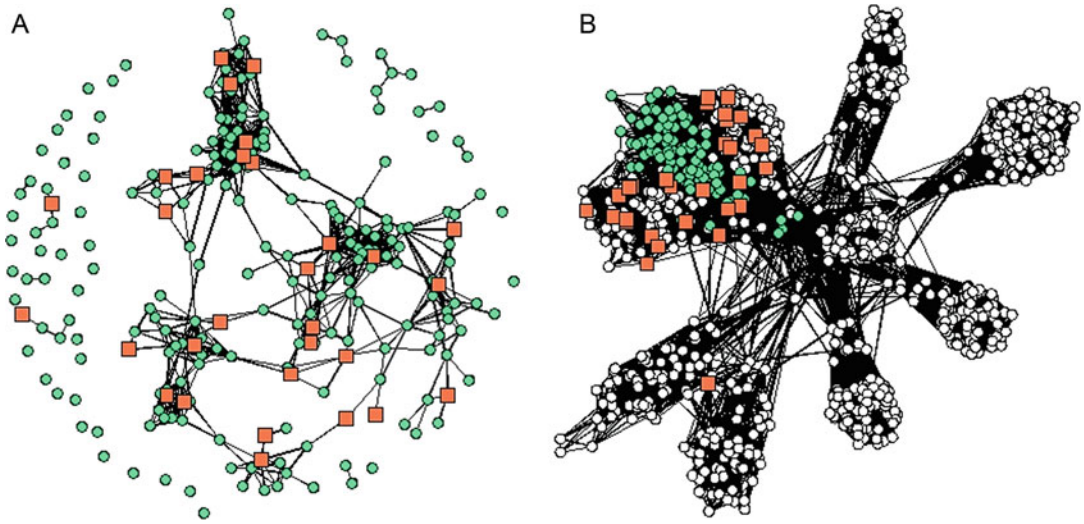

Fig. 3 Networks representing examples of calibration set (CS) optimized with generalized CD criteria (**A**: CDmean and **B**: CDpop). The green dots indicate the individuals to be predicted (PS), the red squares indicate the 30 individuals composing the CS optimized with CD criteria. Individuals are connected with an edge when their genetic relationship is above a given threshold. In (**A**) we considered a highly diverse panel in which the objective was to sample a CS optimal for the prediction of the remaining individuals. In (**B**), we considered different biparental populations of a Nested Association Mapping (NAM) design, with the objective of predicting one given biparental family by sampling an optimal CS from the other biparental families. The contrasts were adapted to answer these two prediction objectives and correspond to the criteria CDmean (**A**) and CDpop (**B**), *see* Rincent et al. [83, 87]. In (**A**), the network indicates that CDmean selects key individuals related to many others. In (**A**) the network illustrates that CDpop samples individuals the most representative of the PS, mostly belonging to biparental populations strongly related to the PS

reliability for each individual–trait combination. This is a generalization of the single trait CD, in which the genetic and error covariances are adapted to the multitrait context. The computation of this criterion (CDmulti) is as follows.

$\text{CDmulti}(\boldsymbol{c}) = \frac{\boldsymbol{c}'((\boldsymbol{\Sigma}_a \otimes \boldsymbol{A}) - (\boldsymbol{Z}'\boldsymbol{M}_3\boldsymbol{Z} + (\boldsymbol{\Sigma}_a^{-1} \otimes \boldsymbol{A}^{-1}))^{-1})\boldsymbol{c}}{\boldsymbol{c}'(\boldsymbol{\Sigma}_a \otimes \boldsymbol{A})\boldsymbol{c}}$, with $\otimes$ the Kronecker product,

$$\boldsymbol{M}_3 = (\boldsymbol{\Sigma}_\varepsilon^{-1} \otimes \boldsymbol{I}) - (\boldsymbol{\Sigma}_\varepsilon^{-1} \otimes \boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}'(\boldsymbol{\Sigma}_\varepsilon^{-1} \otimes \boldsymbol{I})\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}_\varepsilon^{-1} \otimes \boldsymbol{I}).$$

Computing CDmulti requires prior knowledge on genetic covariances between traits (genetic and error covariance matrices between traits), and so the optimized multitrait design is specific to a set of traits or environments. In CDmulti, each individual–trait combination is characterized by a CD value (using the corresponding contrast). Ben-Sadoun et al. [117] considered a trait-assisted prediction scenario with a target trait and a secondary trait easy and inexpensive to phenotype and correlated to the target trait. The goal was to identify which individuals should be phenotyped for the target trait, for the secondary trait or for both, to maximize prediction accuracy of the PS for the target trait with

budget constraints. They showed that phenotyping strategies optimized with CDmulti resulted in a slight but systematic increase of prediction accuracy in comparison to random sampling. In a multi-environment context, one can expect some levels of GxE and different phenotyping costs associated to each environment. In this situation, CDmulti could help determine which individuals should be phenotyped in each environment.

3.2.3 CS Optimization Using the Expected Predictive Ability Or Accuracy (r)

More recently, Ou and Liao [101] proposed to derive the expected Pearson correlation between phenotypes and predicted breeding values (r) in the PS, often referred to as predictive ability. This optimization criterion is also derived from the GBLUP model and can be computed without any phenotypic data. This criterion is interesting as it directly targets the predictive ability, which is related to genetic progress. The authors showed that it resulted in higher predictive ability than other criteria derived from the GBLUP model (PEV, CD) and stratified sampling. This conclusion could, however, partly be due to the fact that the CD criteria were computed within the CS (the genotypes of the PS individuals were not considered).

The main limitation common to all aforementioned criteria is that they rely on genome-wide relatedness through the use of a GBLUP model, which means that they are only adapted to polygenic traits. This is not a problem for most productivity traits, but they are not adapted to traits influenced by major genes such as some disease resistances or phenology. Theoretical developments could be proposed in the future to adapt these criteria to trait specific genetic architecture, in particular to the presence of major genes. A new criterion (EthAcc) targeting the expected prediction accuracy ($r(u, u)$) was proposed to better take genetic architecture into account using the results of genome wide association studies obtained with historic data and genotypic information of the PS [118, 119]. The objective here is to determine an optimal CS from existing phenotyped and genotyped individuals. This is a common situation in plant breeding, as breeders accumulate such data year after year. This criterion was efficient to determine the optimal size and composition of the CS, but the search algorithms were unable to identify the optimal CS without using phenotypic information from the PS. This approach implies that the CS is specific to a given trait and requires the identification of QTLs prior to CS optimization.

3.3 Search Algorithms for Optimal CS and Corresponding Packages

For most of the abovementioned criteria, it is not possible to directly determine the CS with the optimal value for the chosen criterion. For instance, there is no analytical way to determine the CS with the best CD, PEV or r value. Different iterative optimization algorithms were proposed based on exchanges of individuals between the CS and the remaining individuals to improve step by

step the criterion computed for a combination CS/PS. These algorithms can be simple exchange algorithms [83, 87, 117], genetic algorithms [99, 101, 120, 121] or differential exchange algorithms [121] (*see* Table 1 for the list of scripts available implementing different algorithms). Such iterative algorithms do not guarantee convergence toward the global optimum, and have to be run with different starting values and with sufficient iterations to reach a better CS than the initial one. One of the main limits of these criteria is that the search algorithm is computationally demanding for large datasets composed of thousands of individuals or beyond [99]. Approaches based on approximation of the PEV [123–125], including principal component analysis on the genotypic data [99], can reduce computational time. It would be interesting to include contrasts in this approach to optimize more specific prediction objectives.

4 Focus on Some Specific Applications of CS Optimization

4.1 CS Optimization for Predicting Biparental Populations

Plant breeders mainly work with full-sib families, which is a specific case of population structure. Optimizing a CS is particularly challenging in this case because of the different LD phases and QTLs segregating between families. Considering a single family, the optimization of the CS can be done with the criteria based on genetic relatedness presented above. However, in Marulanda et al. [126] where CS optimization was applied within each family, all the tested criteria failed to optimize the CS. In this scenario, due to strong relatedness between full sibs, the improvement associated with CS optimization is expected to be limited in comparison to what can be observed with more diverse material. Apart from these simple within-family scenarios or the situations in which the parents involved in the different crosses are genetically close, the identification of families highly predictive of a target family is challenging [87, 127] even when the phenotypic variance and heritability of each family is known [128]. It is common that unrelated families result in negative prediction accuracy [19, 127], and so it is important to remove such families from the calibration set.

To identify the best predictive families Schopp et al. [50] proposed criteria such as the proportion of shared segregating SNP in the CS and the PS families (θ), the linkage phase similarity [40], or the simple matching coefficient [129]. θ was efficient to predict the accuracy when averaged over many traits, but was much less efficient when considering a given trait because of trait specific genetic architecture. Brauner et al. [127] concluded that it was too risky to add unrelated families to the CS with regard to the potential gain in predictive ability, and so recommended to include only full and half sibs.

Rincent et al. [87] proposed a criterion (CDpop) derived from the generalized CD to predict the prediction accuracy within a given family when using as CS individuals sampled from one or several other families. This criterion was able to predict the observed prediction accuracy quite accurately, and was efficient to optimize CS specifically designed to predict a given family. The prediction accuracies were on average much higher with CDpop than with random sampling. However, this study was based on families of half-sibs (NAM, [39]) and CDpop has not been tested yet on unrelated families.

4.2 CS Optimization or Update When Phenotypes Are Already Available

The criteria introduced in the previous parts were mostly proposed to optimize the composition of the CS prior to any phenotyping. Breeders are, however, facing situations in which some individuals have already been phenotyped, for instance when the CS has to be selected from previous breeding cycles. In these situations, the information provided by the phenotypes may be used to improve the composition of the CS. This would be valuable in two situations: the regular update of the CS along breeding cycles, or the selection of phenotypes from historical data.

4.2.1 Updating the CS

Prediction accuracy decreases over time in successive breeding cycles because of the lower genetic similarity and increased discrepancy of segregating QTLs between the CS and the PS [28–35]. This makes it necessary to regularly update the CS by phenotyping additional individuals. The selection of the new individuals to include in the CS, can be done with the abovementioned criteria, but we can think that the phenotyping data collected in the previous cycles could help updating the CS. Neyhart et al. [130] and Brandariz and Bernardo [131] have proposed to update the CS with the individuals with the best and worst GEBV in the previous generation(s). Simulations showed that it resulted in higher prediction accuracy than random sampling, PEVmean or CDmean. The efficiency of this approach was illustrated in various experimental studies [132–135]. We can suppose that the efficiency of this strategy is due to the maximization of the number of segregating QTLs in the CS.

4.2.2 Subsampling Historical Phenotypic Records

Breeders have access to important phenotypic data collected year after year that can be used to calibrate the GS model. It was, however, shown that subsampling part of the available phenotypic data can improve the predictive ability in comparison to using the full dataset. The presence of genetically distant individuals can indeed decrease predictive ability [23]. This subsampling can be done with classical criteria such as PEVmean, CDmean, or r derived from the GBLUP, but they cannot be used to determine the optimal CS size as they always improve when adding additional individuals. They can, however, be used to determine the

composition and size of the CS after which the criterion only marginally improves [101]. Criterion such as EthAcc [119] does not present the same limitation, but its use in practice is hindered by the poor ability of the search algorithm to identify the optimal CS without including the PS phenotypes. Another option is to determine a CS specific to each predicted individual by selecting its most related individuals [23] or by optimizing criteria based on PEV(c) or CD(c) (PEVmean1, [103]). With PEVmean1, a CS is specifically designed for each PS individual by minimizing its individual PEV. Predictive abilities obtained with PEVmean1 were generally similar to those obtained with PEVmean, but higher for small CS. De los Campos and Lopez-Cruz [136] have formalized an approach in which a penalty is used to set to zero the contribution of some individuals to the prediction. They showed that it could significantly increase predictive ability when the penalty coefficient is well determined.

4.2.3 Optimizing the Choice of Individuals to Be Genotyped

In all the optimization approaches presented above, it was supposed that genotypic information was available for all CS individuals. It can, however, happen that only part of the individuals with historical phenotypic data have been genotyped, and in this case it could be valuable to genotype some additional key individuals to improve the predictions. This selection can be guided by the phenotypic data or the pedigree. Boligon et al. [133] and Michel et al. [134] have proposed to apply the "best and worst individuals" sampling strategy to identify the individuals that should be preferentially genotyped. Maenhout et al. [137] have used the generalized CD (computed with pedigree) to improve the subsampling of historical data by taking into account the balance (number of replicates of each variety) and the connectedness between individuals (disconnectedness can be present when unrelated individual are evaluated in distinct trials). Bartholomé et al. [138] proposed a two-step strategy involving pedigree information and simulations.

4.3 Optimization of the Calibration Set in the Context of Hybrid Breeding

For many plant and animal species, commercial products are hybrids between individuals from different genetic groups (different breeds or heterotic groups). In animal species such as pigs or poultry, even if the commercial products are hybrids, the conventional selection is often done at the purebred level and hybrid performances are seldom considered. With the advent of GS, several studies investigated the interest of accounting for crossbred performances in CS in addition or instead of purebred performances. Recently, Wientjes et al. [139] explored how to optimize CS in this context using simulations but focused mainly on the crossing design used to generate the crossbred individuals from the purebred and not on the composition of the crossbred CS itself. For allogamous plant species such as maize or sunflower, the breeding objective is to produce single-cross hybrid varieties from two

inbred lines, each selected in complementary groups. In this context, the total number of potential single-cross hybrids is very large ($N_1 \times N_2$, if N_1 and N_2 are the numbers of inbred lines in group 1 and 2 respectively) and all of them cannot be evaluated. Classically, the genetic value of a hybrid is decomposed as the sum of the general combining abilities (GCAs) of each of its parental lines (i.e., the average performances of the hybrids progeny generated by crossing one parental line to the lines of the other group) and the Specific Combining Ability (SCA) of the cross (i.e., the complementarity between the two parental lines). In 1994, Bernardo [140] proposed to use molecular markers to compute covariances between the GCAs of parental lines in each group and between SCAs of intergroup hybrids to predict performances of nonphenotyped hybrids from phenotyped ones. It was the first application of GS in plants. Genomic selection is particularly interesting in this context since the genotypes of all potential hybrids can be derived from the genotypes of inbred lines. This offers the possibility to use genotypes of inbred lines to (1) predict GCA of each candidate line evaluated or not as hybrid and (2) to directly predict all potential single-cross hybrid values (GCAs+SCA) to identify the most promising varieties.

First optimization approaches of the CS based on empirical data highlighted that the qualities of prediction of new hybrids were higher when the CS and PS hybrids shared common parental lines, that is when the new hybrids derived from parental lines that contributed to the CS hybrids [141–143]. However, there is a trade-off between the number of hybrid contributions per candidate line and the total number of lines that can contribute to the CS [142]. This trade-off depends on the proportion of SCA relative to the GCA, the total number of hybrids that can be evaluated and on the prediction objective: the prediction of new hybrid combinations between new lines (T0 hybrids) or the prediction of new hybrids between lines that contributed to the CS (T1 or T2 hybrids when respectively one or two of the parental lines are parents of some CS hybrids) [144]. Studies based on real [142], and simulated data [144] showed that increasing the number of lines contributing to the CS at the expense of the number of hybrids evaluated per line is beneficial for better predicting T0 hybrids. However, doing so decreases the total number of T0 hybrids among the whole set of potential hybrids, so the optimal solution over all categories of hybrids depends on the percentage of hybrids that can be phenotyped. This advantage is also reduced when the percentage of SCA is high since the accuracy of SCA prediction decreases when inbred lines are only evaluated in one single CS hybrid. When the objective is to predict the hybrid values in the next generations, increasing the number of lines in the CS at the expense of their contribution is generally the best solution (unless a large percentage of the variance is due to SCA). Recently, Guo et al.

[85] proposed a strategy called MaxCD (Maximization of Connectedness and Diversity). In this strategy, a representative subset of parental lines is first selected from patterns detected in the inbred line genomic relationship matrix. From these lines, a set of hybrids with nonoverlapping parental lines is defined and combined with a set of hybrids issued from pairs of inbred lines most distant from each other. The idea is to represent in the CS the expected diversity of the whole set of hybrids.

Besides these empirical optimizations, other criteria such as those based on PEV and CD were proposed recently. Momen and Morota [98] extended the CD and PEV to include nonadditive effects. In a model including additive and dominance effects they proposed to use a multikernel approach for the predictions and to use as K matrix in the CD and PEV, a linear combination of the additive and dominance relationship matrices (G and D) each weighted by the proportion of variance associated with these variance components, that is,

$$\boldsymbol{K} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_D^2}\boldsymbol{G} + \frac{\sigma_D^2}{\sigma_A^2 + \sigma_D^2}\boldsymbol{D}$$

They evaluated the link between the CD and the genomic prediction accuracies in an animal breeding context using simulations and real pig data. Based on their results they proposed to use the CD for optimizing the CS. Note, however, that they did not consider a hybrid design between unrelated populations and therefore assumed in their prediction model that there was only a single additive variance component and a dominance variance component, which does not correspond to the decomposition of hybrid value in terms of GCA and SCA commonly used for factorial designs. Fritsche Neto et al. [145] used the same formalism to evaluate the interest of genomic selection in different maize hybrid designs and optimized the CS using PEV. They used historical data of variance component estimation to weigh the proportion of additive and dominance variance in PEV computation and also considered, as a benchmark, PEV based on additivity only. Their results showed the interest in using PEV to optimize the hybrid CS, but not the interest of considering dominance for its computation. In agreement with empirical optimization, they found that an optimal hybrid CS should involve as many parental lines as possible. More recently, Heslot and Feoktistov [122] also confirmed on sunflower data the interest of optimizing the hybrid CS using PEV based on a single additive variance. Kadam et al. [93] used an individual CD to identify among all potential hybrids that could be produced from segregation families those to be phenotyped to be included in the CS. They confirm the interest in using these criteria (individual CD or PEV) for optimizing the CS compared to the use of stratified sampling. Akdemir et al. [121] proposed to

choose wheat hybrids to be included in the CS to best predict the remaining hybrids by maximizing the worst individual CD of the PS (CDmin) and showed its interest relative to random sampling. To our knowledge, no optimization study has been based so far on CD or PEV of contrasts (CDmean and PEVmean) and questions remain on the extension of these criteria using a GCA / SCA formalism.

4.4 Optimization of the Phenotypic Evaluation of the Calibration Set

In terms of optimization of the CS, beyond its composition, a key question is the optimization of the experimental design for its evaluation in next to come experiments or, if the CS is based on historical data, the choice of the phenotypic data that should be included in the model calibration process.

Optimization of the phenotyping design is a classical question in plant breeding as a compromise must be found between the number of individuals to phenotype, which has a direct impact on the selection intensity, and the phenotyping effort: number of traits measured, number of replicates within each field trial, and the number of field trials [146]. Marker-assisted selection (MAS), allowing selecting nonphenotyped individuals using marker-based predictions, leads to a different optimal resource allocation compared to phenotypic selection. In MAS, phenotypes are mostly used to estimate marker effects and detect QTLs. As population size plays a major role in determining the power of QTL detection, optimal resources allocation strategies for QTL-based MAS are to phenotype a larger number of individuals but with a lower number of replications per individual compared to phenotypic selection [147].

The first attempts to optimize the experimental design for phenotyping the CS, focused on selection within a given biparental population. Those approaches were based on simulations [81] and/or deterministic formula of the expected accuracy of GS adapted from Daetwyler et al. [48]:

$$r(g,\widehat{g}) = \sqrt{\frac{Nh^2}{Nh^2 + M_e}}$$

where N is the size of CS, h^2 is the trait heritability at the design level (which depends on the individual plot heritability, the number of plots and the GxE variance component) and M_e corresponds to the number of independent loci segregating in the population. This formula assumes that the accuracy of prediction does not depend on CS composition. When considering a segregating population where the LD is only due to cosegregation, M_e can be approximated from the number of chromosomes and the expected number of recombination events along chromosomes. Both Lorenz [81] and Riedelsheimer and Melchinger [148] therefore considered an M_e value around 30 for a single biparental segregating family of

maize. Endelman et al. [149] estimated M_e on two real data sets of barley and maize and used this estimate to derive expected accuracies for the optimization process. In GS, phenotypic data are used to calibrate prediction equations with little concern on the accuracy of each marker effect estimation compared to MAS. So even if the prediction accuracy of untested individuals increases with the CS size, it plateaus more quickly than for MAS giving more flexibility in terms of design in the trade-off between the number of individuals evaluated and the number of replicates. Riedelsheimer and Melchinger [148] extended the approach by (1) considering the prediction accuracy of untested individuals but also of the tested individuals included in the CS to predict the genetic gain and (2) by taking into account GxE when optimizing the number of environments in which the CS is evaluated. Endelman et al. [149] showed that an efficient strategy is to combine GS and sparse designs in which different subsets of CS individuals are phenotyped in each trial, reducing the total number of plots needed without reducing the number of phenotyped individuals nor the number of locations. Other optimization approaches [150, 151] also studied optimal resource allocation for the phenotyping of the CS using deterministic simulations but instead of studying the impact of the resource allocation on the GS accuracy, they considered it as an entry parameter. Jarquin et al. [115] using maize experimental data confirmed the interest in using genomic prediction models including GxE effects with sparse designs in which most genotypes are evaluated in only one trial. They, nevertheless, recommended having a small percentage of individuals common to the different trials.

All the abovementioned approaches aim at optimizing the phenotyping for a next to come population of candidates considering that part of them will be phenotyped to predict the remaining ones. They did not consider the genotypic information of the candidates when choosing among them which individuals should be included in the CS at a fixed CS size. More recently, Atanda et al. [86] extended the use of the CDmean proposed by Rincent et al. [83] to this purpose in a maize data set composed of segregating families. They considered two different phenotypic designs: sparse testing (ST) design where all candidates of the targeted family are evaluated but each in only one trial and another strategy where only half of the candidates of the targeted family (HFS) are evaluated in all field trials. In both cases, they showed that CDmean efficiently selects the subset of individuals to be evaluated in each trial in ST designs and which individuals should be evaluated in the targeted family to predict the remaining ones in the HFS design. Extensions of this approach, considering phenotypes in different trials as correlated traits, showed the interest of using multitrait CD to optimize the allocation of CS individuals to different field trials [116, 121]. This opens the way to combine optimization of CS with optimal resource allocation.

A step forward into the optimization would be to fully integrate the optimization of the CS with the optimization of the experimental design up to the plot allocation of individuals in each field trial. Recently, Cullis et al. [152] showed by simulation that partially repeated field trial designs, optimized using "model-based design" and considering genetic relatedness between genotypes based on pedigree, increased the prediction accuracy of their genetic values. The optimization was based on a sum of the PEV of all pairwise contrasts between the genetic values of the individuals which ensured an efficient comparison between all of them. Ideally, it would be interesting to extend this approach for optimizing experimental design and CS composition for the prediction of individuals in the PS. This would require efficient optimization processes to jointly address these two issues.

5 Conclusion and Prospects

The practical implementation of a new tool in breeding mainly depends on the balance between costs and benefits. In this regard, the optimization of the experimental designs and in particular the optimization of the calibration set in genomic prediction is essential because it can reduce costs and increase benefits [153]. CS composition optimized with the criteria presented here most of the time resulted in higher prediction accuracy than random CS. The choice of the appropriate criterion depends on many factors including the prediction objectives, the population structure, the genetic architecture of the trait and the type of data available (e.g., PS individuals genotyped or not). In any case, there is no universal CS that would be optimal for any genetic material and any trait. We emphasize that it is fundamental to take the genotypic information of the PS into account when available to optimize the CS.

Criteria such as CD, PEV, or r should be further investigated to address other questions such as the optimization of the CS for predicting hybrids or crosses that have not been produced yet [93, 122]. Another application in a plant breeding context would be to optimize jointly the CS size, its composition as well as the phenotypic design for each individual (we can suppose that it might be beneficial to phenotype more deeply key individuals).

Another issue that should be taken care of, is the effect of the composition of the CS on the loss of diversity in the breeding population. Eynard et al. [154] have indeed shown that the way of updating the CS affected the genetic diversity of the breeding population along cycles, maybe because reducing the diversity within the CS can result in fixing some of the QTLs. The effect of CS optimized with the abovementioned criteria on this potential loss of diversity has not been studied yet. A CS constrained optimization procedure that combines both objectives by maximizing

predictive ability while controlling the loss of diversity would be valuable, this was not addressed yet in literature.

References

1. Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447. https://doi.org/10.2307/2529430
2. Tsairidou S, Woolliams JA, Allen AR, Skuce RA, McBride SH, Wright DM, Bermingham ML, Pong-Wong R, Matika O, McDowell SWJ, Glass EJ, Bishop SC (2014) Genomic prediction for tuberculosis resistance in dairy cattle. PLoS One 9:e96728. https://doi.org/10.1371/journal.pone.0096728
3. Daetwyler HD, Villanueva B, Bijma P, Woolliams JA (2007) Inbreeding in genome-wide selection. J Anim Breed Genet 124:369–376. https://doi.org/10.1111/j.1439-0388.2007.00693.x
4. Meuwissen THE, Hayes BJ, Goddard ME (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
5. Gianola D, des los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363. https://doi.org/10.1534/genetics.109.103952
6. Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194:573–596. https://doi.org/10.1534/genetics.113.151753
7. Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME (2009) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41:51. https://doi.org/10.1186/1297-9686-41-51
8. Pryce JE, Gredler B, Bolormaa S, Bowman PJ, Egger-Danner C, Fuerst C, Emmerling R, Sölkner J, Goddard ME, Hayes BJ (2011) Short communication: genomic selection using a multi-breed, across-country reference population. J Dairy Sci 94:2625–2630. https://doi.org/10.3168/jds.2010-3719
9. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME (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
10. Daetwyler HD, Kemper KE, van der Werf JHJ, Hayes BJ (2012) Components of the accuracy of genomic prediction in a multi-breed sheep population. J Anim Sci 90:3375–3384. https://doi.org/10.2527/jas.2011-4557
11. Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink J-L, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, Melchinger AE (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
12. Rio S, Mary-Huard T, Moreau L, 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
13. Duhnen A, Gras A, Teyssèdre S, Romestant M, Claustres B, Daydé J, Mangin B (2017) Genomic selection for yield and seed protein content in soybean: a study of breeding program data and assessment of prediction accuracy. Crop Sci 57:1325–1337. https://doi.org/10.2135/cropsci2016.06.0496
14. Lorenz AJ, Smith KP, 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
15. Rio S, Gallego-Sánchez L, Montilla-Bascón G, Canales FJ, Isidro y Sánchez J, 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
16. Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, 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
17. Legarra A, Robert-Granié C, Manfredi E, Elsen J-M (2008) Performance of genomic

selection in mice. Genetics 180:611–618. https://doi.org/10.1534/genetics.108.088575

18. Albrecht T, Wimmer V, Auinger H-J, Erbe M, Knaak C, Ouzunova M, Simianer H, Schön C-C (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339. https://doi.org/10.1007/s00122-011-1587-7
19. Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, Flament P, Melchinger AE, Menz M, Meyer N, Moreau L, Moreno-González J, Ouzunova M, Pausch H, Ranc N, Schipprack W, Schönleben M, Walter H, Charcosset A, Schön C-C (2014) Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. Genetics 198:3–16. https://doi.org/10.1534/genetics.114.161943
20. Herter CP, Ebmeyer E, Kollers S, Korzun V, Würschum T, Miedaner T (2019) Accuracy of within- and among-family genomic prediction for fusarium head blight and Septoria tritici blotch in winter wheat. Theor Appl Genet 132:1121–1135. https://doi.org/10.1007/s00122-018-3264-6
21. Nielsen NH, Jahoor A, Jensen JD, Orabi J, Cericola F, Edriss V, Jensen J (2016) Genomic prediction of seed quality traits using advanced barley breeding lines. PLoS One 11:e0164494. https://doi.org/10.1371/journal.pone.0164494
22. Würschum T, Maurer HP, Weissmann S, Hahn V, Leiser WL (2017) Accuracy of within- and among-family genomic prediction in triticale. Plant Breed 136:230–236. https://doi.org/10.1111/pbr.12465
23. Lorenz AJ, Smith KP (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
24. Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink J-L, Melchinger AE (2013) Genomic predictability of interconnected Biparental maize populations. Genetics 194:493–503. https://doi.org/10.1534/genetics.113.150227
25. Hidalgo AM, Bastiaansen JWM, Lopes MS, Harlizius B, Groenen MAM, de Koning D-J (2015) Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3 (Bethesda) 5:1575–1583. https://doi.org/10.1534/g3.115.018119
26. Rio S, Moreau L, Charcosset A, Mary-Huard T (2020) Accounting for group-specific allele effects and admixture in genomic predictions: theory and experimental evaluation in maize. Genetics 216:27–41. https://doi.org/10.1534/genetics.120.303278
27. Olatoye MO, Clark LV, Labonte NR, Dong H, Dwiyanti MS, Anzoua KG, Brummer JE, Ghimire BK, Dzyubenko E, Dzyubenko N, Bagmet L, Sabitov A, Chebukin P, Głowacka K, Heo K, Jin X, Nagano H, Peng J, Yu CY, Yoo JH, Zhao H, Long SP, Yamada T, Sacks EJ, Lipka AE (2020) Training population optimization for genomic selection in Miscanthus. G3 (Bethesda) 10:2465–2476. https://doi.org/10.1534/g3.120.401402
28. Pszczola M, Calus MPL (2016) Updating the reference population to achieve constant genomic prediction reliability across generations. Animal 10:1018–1024. https://doi.org/10.1017/S1751731115002785
29. Castro Dias Cuyabano B, Wackel H, Shin D, Gondro C (2019) A study of genomic prediction across generations of two Korean pig populations. Animals 9:672. https://doi.org/10.3390/ani9090672
30. Hofheinz N, Borchardt D, Weissleder K, Frisch M (2012) Genome-based prediction of test cross performance in two subsequent breeding cycles. Theor Appl Genet 125:1639–1645. https://doi.org/10.1007/s00122-012-1940-5
31. Li X, Wei Y, Acharya A, Hansen JL, Crawford JL, Viands DR, Michaud R, Claessens A, Brummer EC (2015) Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. Plant Genome 8:plantgenome2014.12.0090. https://doi.org/10.3835/plantgenome2014.12.0090
32. Wang N, Wang H, Zhang A, Liu Y, Yu D, Hao Z, Ilut D, Glaubitz JC, Gao Y, Jones E, Olsen M, Li X, San Vicente F, Prasanna BM, Crossa J, Pérez-Rodríguez P, Zhang X (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
33. Michel S, Ametz C, Gungor H, Epure D, Grausgruber H, Löschenberger F, Buerstmayr H (2016) Genomic selection across multiple breeding cycles in applied bread wheat breeding. Theor Appl Genet 129:1179–1189. https://doi.org/10.1007/s00122-016-2694-2
34. Sallam AH, Endelman JB, Jannink J-L, Smith KP (2015) Assessing genomic selection prediction accuracy in a dynamic barley breeding population. Plant Genome 8:

plantgenome2014.05.0020. https://doi.org/10.3835/plantgenome2014.05.0020

35. Auinger H-J, Schönleben M, Lehermeier C, Schmidt M, Korzun V, Geiger HH, Piepho H-P, Gordillo A, Wilde P, Bauer E, Schön C-C (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.). Theor Appl Genet 129:2043–2053. https://doi.org/10.1007/s00122-016-2756-5
36. de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. Genetics 183: 1545–1553. https://doi.org/10.1534/genetics.109.104935
37. Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231. https://doi.org/10.1007/BF01245622
38. Wright S (1949) The Genetical structure of populations. Ann Eugenics 15:323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x
39. Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Rincent R, Schipprack W, Altmann T, Flament P, Melchinger AE, Menz M, Moreno-González J, Ouzunova M, Revilla P, Charcosset A, Martin OC, Schön C-C (2013) Intraspecific variation of recombination rate in maize. Genome Biol 14:R103. https://doi.org/10.1186/gb-2013-14-9-r103
40. de Roos APW, Hayes BJ, Spelman RJ, Goddard ME (2008) Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179: 1503–1512. https://doi.org/10.1534/genetics.107.084301
41. Porto-Neto LR, Kijas JW, Reverter A (2014) The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genet Sel Evol 46:22. https://doi.org/10.1186/1297-9686-46-22
42. Badke YM, Bates RO, Ernst CW, Schwab C, Steibel JP (2012) Estimation of linkage disequilibrium in four US pig breeds. BMC Genomics 13:24. https://doi.org/10.1186/1471-2164-13-24
43. Heifetz EM, Fulton JE, O'Sullivan N, Zhao H, Dekkers JCM, Soller M (2005) Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics 171: 1173–1181. https://doi.org/10.1534/genetics.105.040782
44. Van Inghelandt D, Reif JC, Dhillon BS, Flament P, Melchinger AE (2011) Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. Theor Appl Genet 123:11–20. https://doi.org/10.1007/s00122-011-1562-3
45. Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125: 1181–1194. https://doi.org/10.1007/s00122-012-1905-8
46. Hao C, Wang L, Ge H, Dong Y, Zhang X (2011) Genetic diversity and linkage disequilibrium in Chinese bread wheat (Triticum aestivum L.) revealed by SSR markers. PLoS One 6:e17279. https://doi.org/10.1371/journal.pone.0017279
47. Ibáñez-Escriche N, Fernando RL, Toosi A, Dekkers JC (2009) Genomic selection of purebreds for crossbred performance. Genet Sel Evol 41:12. https://doi.org/10.1186/1297-9686-41-12
48. Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:e3395. https://doi.org/10.1371/journal.pone.0003395
49. Wientjes YC, Calus MP, Goddard ME, Hayes BJ (2015) Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genet Sel Evol 47:42. https://doi.org/10.1186/s12711-015-0124-6
50. Schopp P, Müller D, Wientjes YCJ, Melchinger AE (2017) Genomic prediction within and across Biparental families: means and variances of prediction accuracy and usefulness of deterministic equations. G3 (Bethesda) 7: 3571–3586. https://doi.org/10.1534/g3.117.300076
51. Scutari M, Mackay I, Balding D (2016) Using genetic distance to infer the accuracy of genomic prediction. PLoS Genet 12:e1006288. https://doi.org/10.1371/journal.pgen.1006288
52. Varona L, Legarra A, Toro MA, Vitezica ZG (2018) Non-additive effects in genomic selection. Front Genet 9:78. https://doi.org/10.3389/fgene.2018.00078
53. Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4:e1000008. https://doi.org/10.1371/journal.pgen.1000008
54. Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominant variance and

covariance of individuals within the genomic selection scope. Genetics 195:1223–1230. https://doi.org/10.1534/genetics.113.155176

55. Wientjes YCJ, Bijma P, Vandenplas J, Calus MPL (2017) Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. Genetics 207:503–515. https://doi.org/10.1534/genetics.117.300152
56. Wientjes YCJ, Calus MPL, Duenk P, Bijma P (2018) Required properties for markers used to calculate unbiased estimates of the genetic correlation between populations. Genet Sel Evol 50:65. https://doi.org/10.1186/s12711-018-0434-6
57. Thompson EA (2013) Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194:301–326. https://doi.org/10.1534/genetics.112.148825
58. Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? Nat Rev Genet 16:33–44. https://doi.org/10.1038/nrg3821
59. Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397. https://doi.org/10.1534/genetics.107.081190
60. Habier D, Tetens J, Seefried F-R, Lichtner P, 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
61. Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194:597–607. https://doi.org/10.1534/genetics.113.152207
62. Zhong S, Dekkers JCM, Fernando RL, 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
63. Jannink J-L, Lorenz AJ, 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
64. Clark SA, Hickey JM, Daetwyler HD, van der Werf JH (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
65. Wientjes YCJ, Veerkamp RF, Calus MPL (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
66. Pszczola M, Strabel T, Mulder HA, Calus MPL (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
67. Albrecht T, Auinger H-J, Wimmer V, Ogutu JO, Knaak C, Ouzunova M, Piepho H-P, 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
68. Cros D, Denis M, Sánchez L, Cochard B, Flori A, Durand-Gasselin T, Nouy B, Omoré A, Pomiès V, Riou V, Suryana E, Bouvet J-M (2015) Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). Theor Appl Genet 128:397–410. https://doi.org/10.1007/s00122-014-2439-z
69. Vitezica ZG, Legarra A, Toro MA, Varona L (2017) Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. Genetics 206:1297–1307. https://doi.org/10.1534/genetics.116.199406
70. Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91:47–60. https://doi.org/10.1017/S0016672308009981
71. 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
72. Goddard ME, Hayes BJ, Meuwissen THE (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 128:409–421. https://doi.org/10.1111/j.1439-0388.2011.00964.x
73. Elsen J-M (2016) Approximated prediction of genomic selection accuracy when reference and candidate populations are related. Genet Sel Evol 48:18. https://doi.org/10.1186/s12711-016-0183-3

74. Elsen J-M (2017) An analytical framework to derive the expected precision of genomic selection. Genet Sel Evol 49:95. https://doi.org/10.1186/s12711-017-0366-6
75. Heffner EL, Jannink J-L, Iwata H, Souza E, Sorrells ME (2011) Genomic selection accuracy for grain quality traits in Biparental wheat populations. Crop Sci 51:2597–2606. https://doi.org/10.2135/cropsci2011.05.0253
76. Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112:48–60. https://doi.org/10.1038/hdy.2013.16
77. Norman A, Taylor J, Edwards J, 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
78. Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186. https://doi.org/10.1186/1471-2105-12-186
79. Dehnavi E, Mahyari SA, Schenkel FS, Sargolzaei M (2018) The effect of using cow genomic information on accuracy and bias of genomic breeding values in a simulated Holstein dairy cattle population. J Dairy Sci 101:5166–5176. https://doi.org/10.3168/jds.2017-12999
80. Lello L, Raben TG, Yong SY, Tellier LCAM, Hsu SDH (2019) Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. Sci Rep 9:15286. https://doi.org/10.1038/s41598-019-51258-x
81. Lorenz AJ (2013) Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. G3 Bethesda 3:481–491. https://doi.org/10.1534/g3.112.004911
82. Wu X, Lund MS, Sun D, Zhang Q, Su G (2015) Impact of relationships between test and training animals and among training animals on reliability of genomic prediction. J Anim Breed Genet 132:366–375. https://doi.org/10.1111/jbg.12165
83. Rincent R, Laloe D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodriguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E, Schoen C-C, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L (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
84. Bustos-Korts D, Malosetti M, Chapman S, Biddulph B, van Eeuwijk F (2016) Improvement of predictive ability by uniform coverage of the target genetic space. G3 (Bethesda) 6(11):3733–3747. https://doi.org/10.1534/g3.116.035410
85. Guo T, Yu X, Li X, Zhang H, Zhu C, Flint-Garcia S, McMullen MD, Holland JB, Szalma SJ, Wisser RJ, 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
86. Atanda SA, Olsen M, Burgueño J, Crossa J, Dzidzienyo D, Beyene Y, Gowda M, Dreher K, Zhang X, Prasanna BM, Tongoona P, Danquah EY, Olaoye G, Robbins KR (2021) 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
87. Rincent R, Charcosset A, 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
88. Roth M, Muranty H, Di Guardo M, Guerra W, Patocchi A, Costa F (2020) Genomic prediction of fruit texture and training population optimization towards the application of genomic selection in apple. Hortic Res 7:1–14. https://doi.org/10.1038/s41438-020-00370-5
89. Berro I, Lado B, Nalin RS, Quincke M, Gutiérrez L (2019) Training population optimization for genomic selection. Plant Genome 12:190028. https://doi.org/10.3835/plantgenome2019.04.0028
90. Tiede T, Smith KP (2018) Evaluation and retrospective optimization of genomic selection for yield and disease resistance in spring barley. Mol Breed 38:55. https://doi.org/10.1007/s11032-018-0820-3
91. Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME (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

92. Adeyemo E, Bajgain P, Conley E, Sallam AH, Anderson JA (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
93. Kadam DC, Rodriguez OR, Lorenz AJ (2021) Optimization of training sets for genomic prediction of early-stage single crosses in maize. Theor Appl Genet 134(2): 687–699. https://doi.org/10.1007/s00122-020-03722-w
94. Laloë D (1993) Precision and information in linear models of genetic evaluation. Genet Sel Evol 25:557. https://doi.org/10.1186/1297-9686-25-6-557
95. Laloë D, Phocas F, Ménissier F (1996) Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. Genet Sel Evol 28:359. https://doi.org/10.1186/1297-9686-28-4-359
96. Yu H, Spangler ML, Lewis RM, Morota G (2018) Do stronger measures of genomic connectedness enhance prediction accuracies across management units?1. J Anim Sci 96: 4490–4500. https://doi.org/10.1093/jas/sky316
97. Zhang S-Y, Olasege BS, Liu D-Y, Wang Q-S, Pan Y-C, Ma P-P (2018) The genetic connectedness calculated from genomic information and its effect on the accuracy of genomic prediction. PLoS One 13:e0201400. https://doi.org/10.1371/journal.pone.0201400
98. Momen M, 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
99. Akdemir D, Sanchez JI, 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
100. Akdemir D, 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
101. Ou J-H, 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
102. Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL, Sorrells ME (2015) Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. Plant Genome 8: eplantgenome2014.09.0046. https://doi.org/10.3835/plantgenome2014.09.0046
103. Sarinelli JM, Murphy JP, Tyagi P, Holland JB, Johnson JW, Mergoum M, Mason RE, Babar A, Harrison S, Sutton R, Griffey CA, Brown-Guedira G (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
104. Charmet G, Tran L-G, Auzanneau J, Rincent R, Bouchet S (2020) BWGS: a R package for genomic selection and its application to a wheat breeding programme. PLoS One 15:e0222733. https://doi.org/10.1371/journal.pone.0222733
105. Wolfe MD, Del Carpio DP, Alabi O, Ezenwaka LC, Ikeogu UN, Kayondo IS, Lozano R, Okeke UG, Ozimati AA, Williams E, Egesi C, Kawuki RS, Kulakow P, Rabbi IY, Jannink J-L (2017) Prospects for genomic selection in cassava breeding. Plant Genome 10. https://doi.org/10.3835/plantgenome2017.03.0015
106. Ozimati A, Kawuki R, Esuma W, Kayondo IS, Wolfe M, Lozano R, Rabbi I, Kulakow P, Jannink J-L (2018) Training population optimization for prediction of cassava Brown streak disease resistance in west African clones. G3 (Bethesda) 8:3903–3913. https://doi.org/10.1534/g3.118.200710
107. 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, 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
108. Keep T, Sampoux J-P, Blanco-Pastor JL, Dehmer KJ, Hegarty MJ, Ledauphin T, Litrico I, Muylle H, Roldán-Ruiz I, Roschanski AM, Ruttink T, Surault F, Willner E, Barre P (2020) High-throughput genome-wide genotyping to optimize the use of natural genetic resources in the grassland species perennial ryegrass (Lolium perenne L.). G3 (Bethesda) 10:3347–3364. https://doi.org/10.1534/g3.120.401491
109. Calus MP, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol 43:26. https://doi.org/10.1186/1297-9686-43-26
110. Jia Y, Jannink J-L (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:

1513–1522. https://doi.org/10.1534/genetics.112.144246

111. Robert P, Le Gouis J, Consortium TB, Rincent R (2020) Combining crop growth modeling with trait-assisted prediction improved the prediction of genotype by environment interactions. Front Plant Sci 11:827. https://doi.org/10.3389/fpls.2020.00827
112. Saint Pierre C, Burgueño J, Crossa J, Fuentes Dávila G, Figueroa López P, Solís Moya E, Ireta Moreno J, Hernández Muela VM, Zamora Villa VM, Vikram P, Mathews K, Sansaloni C, Sehgal D, Jarquin D, Wenzl P, Singh S (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
113. Ly D, Chenu K, Gauffreteau A, Rincent R, Huet S, Gouache D, Martre P, Bordes J, Charmet G (2017) Nitrogen nutrition index predicted by a crop model improves the genomic prediction of grain number for a bread wheat core collection. Field Crops Res 214: 331–340. https://doi.org/10.1016/j.fcr.2017.09.024
114. Rincent R, Malosetti M, Ababaei B, Touzy G, Mini A, Bogard M, Martre P, Le Gouis J, van Eeuwijk F (2019) Using crop growth model stress covariates and AMMI decomposition to better predict genotype-by-environment interactions. Theor Appl Genet 132: 3399–3411. https://doi.org/10.1007/s00122-019-03432-y
115. Jarquin D, Howard R, Crossa J, Beyene Y, Gowda M, Martini JWR, Covarrubias Pazaran G, Burgueño J, Pacheco A, Grondona M, Wimmer V, Prasanna BM (2020) Genomic prediction enhanced sparse testing for multi-environment trials. G3 (Bethesda) 10:2725–2739. https://doi.org/10.1534/g3.120.401349
116. Rio S, Akdemir D, Carvalho T, Isidro y Sánchez J (2021) Assessment of genomic prediction reliability and optimization of experimental designs in multi-environment trials. Theor Appl Genet. https://doi.org/10.1007/s00122-021-03972-2
117. Ben-Sadoun S, Rincent R, Auzanneau J, Oury FX, Rolland B, Heumez E, Ravel C, Charmet G, 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
118. Rabier C-E, Barre P, Asp T, Charmet G, Mangin B (2016) On the accuracy of genomic selection. PLoS One 11:e0156086. https://doi.org/10.1371/journal.pone.0156086
119. Mangin B, Rincent R, Rabier C-E, Moreau L, 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
120. Akdemir D (2017) Selection of training populations (and other subset selection problems) with an accelerated genetic algorithm (STPGA: an R-package for selection of training populations with a genetic algorithm). ArXiv170208088 Cs Q-bio stat
121. Akdemir D, Rio S, 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
122. Heslot N, 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
123. Misztal I, Wiggans GR (1988) Approximation of prediction error variance in large-scale animal models. J Dairy Sci 71:27–32. https://doi.org/10.1016/S0022-0302(88)79976-2
124. VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980
125. Hickey JM, Veerkamp RF, Calus MP, Mulder HA, Thompson R (2009) Estimation of prediction error variances via Monte Carlo sampling methods using different formulations of the prediction error variance. Genet Sel Evol 41:23. https://doi.org/10.1186/1297-9686-41-23
126. Marulanda JJ, Melchinger AE, Würschum T (2015) Genomic selection in biparental populations: assessment of parameters for optimum estimation set design. Plant Breed 134: 623–630. https://doi.org/10.1111/pbr.12317
127. Brauner PC, Müller D, Molenaar WS, Melchinger AE (2020) Genomic prediction with multiple biparental families. Theor Appl Genet 133:133–147. https://doi.org/10.1007/s00122-019-03445-7
128. Edwards SM, Buntjer JB, Jackson R, Bentley AR, Lage J, Byrne E, Burt C, Jack P, Berry S, Flatman E, Poupard B, Smith S, Hayes C, Gaynor RC, Gorjanc G, Howell P, Ober E, Mackay IJ, Hickey JM (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

129. Sneath PHA, Sneath PHA, Sokal RR, Sokal URR (1973) Numerical taxonomy: the principles and practice of numerical classification. W. H. Freeman, New York
130. Neyhart JL, Tiede T, Lorenz AJ, Smith KP (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
131. Brandariz SP, Bernardo R (2018) Maintaining the accuracy of Genomewide predictions when selection has occurred in the training population. Crop Sci 58:1226–1231. https://doi.org/10.2135/cropsci2017.11.0682
132. Jimenez-Montero JA, Gonzalez-Recio O, Alenda R (2012) Genotyping strategies for genomic selection in small dairy cattle populations. Animal 6:1216–1224. https://doi.org/10.1017/S1751731112000341
133. Boligon AA, Long N, Albuquerque LG, Weigel KA, Gianola D, Rosa GJM (2012) Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection. J Anim Sci 90:4716–4722. https://doi.org/10.2527/jas.2012-4857
134. Michel S, Ametz C, Gungor H, Akgöl B, Epure D, Grausgruber H, Löschenberger F, Buerstmayr H (2017) Genomic assisted selection for enhancing line breeding: merging genomic and phenotypic selection in winter wheat breeding programs with preliminary yield trials. Theor Appl Genet 130:363–376. https://doi.org/10.1007/s00122-016-2818-8
135. Hu X, Carver BF, Powers C, Yan L, Zhu L, Chen C (2019) Effectiveness of genomic selection by response to selection for winter wheat variety improvement. Plant Genome 12:180090. https://doi.org/10.3835/plantgenome2018.11.0090
136. Lopez-Cruz M, de los Campos G (2021) Optimal breeding-value prediction using a sparse selection index. Genetics 218:iyab030. https://doi.org/10.1093/genetics/iyab030
137. Maenhout S, De Baets B, Haesaert G (2010) Graph-based data selection for the construction of genomic prediction models. Genetics 185:1463–1475. https://doi.org/10.1534/genetics.110.116426
138. Bartholomé J, Van Heerwaarden J, Isik F, Boury C, Vidal M, Plomion C, Bouffier L (2016) Performance of genomic prediction within and across generations in maritime pine. BMC Genomics 17:604. https://doi.org/10.1186/s12864-016-2879-8
139. Wientjes YCJ, Bijma P, Calus MPL (2020) Optimizing genomic reference populations to improve crossbred performance. Genet Sel Evol 52:65. https://doi.org/10.1186/s12711-020-00573-3
140. Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34:cropsci19940011183X003400010003x. https://doi.org/10.2135/cropsci1994.0011183X003400010003x
141. Massman JM, Gordillo A, Lorenzana RE, 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
142. Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197:1343–1355. https://doi.org/10.1534/genetics.114.165860
143. Kadam DC, Potts SM, Bohn MO, Lipka AE, Lorenz AJ (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
144. Seye AI, Bauland C, Charcosset A, 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. https://doi.org/10.1007/s00122-020-03573-5
145. Fristche-Neto R, Akdemir D, Jannink J-L (2018) Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. Theor Appl Genet 131:1153–1162. https://doi.org/10.1007/s00122-018-3068-8
146. Gauch HG, Zobel RW (1996) Optimal replication in selection experiments. Crop Sci 36:cropsci1996.0011183X003600040002x. https://doi.org/10.2135/cropsci1996.0011183X003600040002x
147. Moreau L, Lemarié S, Charcosset A, Gallais A (2000) Economic efficiency of one cycle of marker-assisted selection. Crop Sci 40:329–337. https://doi.org/10.2135/cropsci2000.402329x

148. Riedelsheimer C, Melchinger AE (2013) Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor Appl Genet 126:2835–2848. https://doi.org/10.1007/s00122-013-2175-9
149. Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, Sorrells ME, Jannink J-L (2014) Optimal Design of Preliminary Yield Trials with genome-wide markers. Crop Sci 54:48–59. https://doi.org/10.2135/cropsci2013.03.0154
150. Longin CFH, Mi X, Melchinger AE, Reif JC, Würschum T (2014) Optimum allocation of test resources and comparison of breeding strategies for hybrid wheat. Theor Appl Genet 127:2117–2126. https://doi.org/10.1007/s00122-014-2365-0
151. Longin CFH, Mi X, 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. https://doi.org/10.1007/s00122-015-2505-1
152. Cullis BR, Smith AB, Cocks NA, Butler DG (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
153. Lorenz A, Nice L (2017) Training population design and resource allocation for genomic selection in plant breeding. In: Varshney RK, Roorkiwal M, Sorrells ME (eds) Genomic selection for crop improvement: new molecular breeding strategies for crop improvement. Springer International Publishing, Cham, pp 7–22
154. Eynard SE, Croiseau P, Laloë D, Fritz S, Calus MPL, Restoux G (2018) Which individuals to choose to update the reference population? minimizing the loss of genetic diversity in animal genomic selection programs. G3 (Bethesda) 8:113–121. https://doi.org/10.1534/g3.117.1117