

Chapter 2

Training Population Design and Resource Allocation for Genomic Selection in Plant Breeding

Aaron Lorenz and Liana Nice

2.1 Introduction

Obtaining accurate and inexpensive estimates of genetic value is a fundamental goal for plant breeders. To obtain these estimates and choose new varieties, breeders continue to rely heavily on standard phenotyping practices for their crops and traits of interest. Series of phenotypic testing procedures employed by plant breeders can vary in scale, complexity, and relevance, both within and across breeding programs. Scale can range from early generation, single plant observations to large, prerelease strip trials. Similarly, the complexity of phenotyping traits within a breeding program can range from measuring flowering time, which can be reliably phenotyped in a single environment in many cases, to drought tolerance which can only be measured in a field setting if specific weather conditions occur or if specially designed water stress nurseries are available. While phenotyping followed by selection is the primary means of advancing lines, the time, cost, and environmental error associated with obtaining phenotypic values leave room for improvement. Advancements in phenotyping technologies have resulted in decreased error, fewer inefficiencies in the phenotyping process, or larger quantities of phenotypic data (Araus and Cairns 2014). An alternative yet complementary approach to reducing phenotyping expenditures involves implementing genomic selection using high-throughput molecular markers in breeding programs (Cabrera-Bosquet et al. 2012).

Initially, molecular markers were used in the context of marker-assisted selection (MAS). This approach typically requires identification of tightly linked or causal markers through mapping or cloning of quantitative trait loci (QTL) using mapping populations or discovery panels. These markers are then used in parallel

A. Lorenz (✉) • L. Nice
University of Minnesota, Minneapolis, MN 55455, USA
e-mail: lore0149@umn.edu

with measured phenotypes to make selections in the breeding program (Johnson 2004). The development of genomic selection techniques has altered the relationship between markers and phenotypic data in breeding programs by introducing a new role for phenotyping. Instead of using phenotypic data for direct measurement of the phenotypic or breeding value of lines, phenotypic data in the context of genomic selection is used to estimate marker effects and develop marker-based predictive models. This is accomplished by developing calibration sets, or training populations, that have been both phenotyped and genotyped with dense, genome-wide markers. From this calibration set, a statistical model using all marker information simultaneously is applied to predict the breeding values of individuals that had not been phenotyped, known as the target population or prediction set. With accurate predictive models, breeders can minimize the number of individuals that are phenotyped and continue selection in environments that are not conducive to obtaining quality phenotypes, such as off-season nurseries. Both scenarios can effectively reduce the cost and/or time necessary for achieving the desired genetic gain.

As genotyping costs continue to decrease, genomic selection will play an increasingly important role in plant breeding. Research surrounding the hypothetical and empirical implementation of genomic selection is an active field of study, and the resulting techniques are being adopted by breeders in many crop species. This movement toward an increasingly data-rich breeding process leads to questions surrounding the application of statistics, experimental design, and quantitative genetics, to the selection of progenies for advancement and varietal release. While the implementation of genomic selection may not affect the methods used for phenotyping per se, breeders will need to consider training accurate genomic prediction models when designing field trials, which would involve at least two aspects: (1) selection of genotypes for field testing that are informative for model building (i.e., training population design) in addition to those being advanced toward variety release and (2) the allocation of field plots to genotypes. The objective of this chapter is to review and discuss studies related to these two important topics. It was our aim to provide the reader with a simple and hopefully intuitive introduction to these topics.

2.2 Training Population Design

A critical first step toward the use of genomic selection is the establishment of the training population (Jannink et al. 2010). Training population composition and the way in which it's established varies according to the role of genomic selection, whether it be rapid recurrent selection within a closed population, selection within a single biparental family, or selection among exotic plant accessions comprising a germplasm collection. Approaches to compiling training populations include the collection of new phenotypic data from targeted trials as well as the mining of historical phenotypic data available on genotyped lines. Once again, the choice

between these two basic strategies depends upon the role of genomic selection in a crop improvement program. The most important consideration of training population design is the target population. In other words, the target population should be defined first and foremost, and then the training population is designed around the target population. There are two basic aims of training population design: (1) minimize costs associated with phenotyping by selecting smaller training populations, and (2) maximize prediction accuracy for the set of individuals being predicted. Balancing these goals should help breeders avoid poor prediction accuracies or wasted resources.

To aid in this decision process, we review a range of studies that explore composition and optimization of training population design. Windhausen et al. (2012) laid out four breeding scenarios under which genomic selection may be used: (1) training and target populations are segregating progenies from the same cross, (2) training and target populations include related and unrelated genotypes, (3) training and target sets include lines from a diverse germplasm collection, and (4) recurrent selection within a closed synthetic population. Literature on training population design for the first three scenarios will be reviewed. Literature on training population design for the case of synthetic populations is sparse at the present time; however one recently published study sheds light on this topic (Schopp et al. 2017). Following the discussion of breeding scenarios, we explore methods of training population selection and other considerations for training population design related to population and trait architecture.

2.2.1 Training and Target Populations Are Segregating Progenies from the Same Cross

The most straightforward way to conduct genomic selection is to create family-specific training populations. In this scenario, individuals from the same family, or biparental population, are used as both the training population and target population. This approach has been discussed extensively in the maize breeding literature (Bernardo and Yu 2007; Windhausen et al. 2012; Lorenz 2013; Jacobson et al. 2014), where large biparental families of inbred or doubled haploid lines are common, as well as the wheat breeding literature (Heffner et al. 2011). To perform genomic prediction, the entire family is genotyped, with a subset of these lines serving as the training population to train a model to predict the individuals that were not phenotyped. The genomic prediction model can also be used to predict future selection cycles created by intermating selected individuals within the family (Bernardo and Yu 2007; Combs and Bernardo 2013; Massman et al. 2013; Lorenz 2013). This breeding method is similar to marker-assisted recurrent selection in terms of family structure (Johnson 2004), and the first published studies on genomic selection for plant breeding used this approach (Whittaker et al. 2000; Bernardo and Yu 2007).

Within-family predictions are often accurate, and only modest population sizes and marker numbers are needed to achieve good prediction accuracy. High accuracy is possible because of the extensive linkage disequilibrium (LD) generated by the initial hybridization event (Lorenzana and Bernardo 2009; Zhao et al. 2012). This LD, which provides power for QTL mapping in biparental populations, also leads to accurate predictions in the context of genomic selection. Generally, as training population size increases within families, predictive ability increases until a maximum has been reached. When working with high heritability traits, the maximum prediction accuracy will be reached with a smaller training population size.

In an era when genotyping can be less expensive than phenotyping, selecting a subset of individuals to phenotype based on genotype data in order to reduce population size (and thus cost of phenotyping) while maintaining QTL detection power is a desirable goal. This is known as selective phenotyping. It has been shown that selective phenotyping for QTL detection can enhance mapping power and resolution depending on the number of QTL controlling a trait and their effect sizes (Jannink 2005; Sen et al. 2009). Although the increase in power for QTL mapping was minimal under optimized schemes, researchers have explored whether similar optimizations could be used in genomic prediction. Marulanda et al. (2015) simulated a biparental population with training population sets that varied based on a large number of parameters. The parameters examined included measures of collinearity among markers, LD, allele frequency, genetic relationships among lines, diversity indices, mixed model parameters, and phenotypic variance of the training population sets. While many of these factors varied with training population size, none of the parameters derived from marker data were associated with prediction accuracy. However, they did find that selection for enhanced phenotypic variation of the training set led to greater prediction accuracy in the case of smaller training populations. While marker-based optimization would be ideal, the authors proposed that a first round of phenotyping with little replication could be used for training population selection, followed by more intense phenotyping of the optimized set across multiple locations (Marulanda et al. 2015). Ultimately, the lack of population structure in a biparental cross allows for relatively good prediction from a random sample, as long as marker number and population size are large enough to adequately train the selection model. The use of genomic prediction to select non-phenotyped individuals within a single family, however, needs to be carefully considered as studies on resource allocation have suggested little to no benefit to only phenotyping a subset of a single family in order to develop a model to predict the remaining individuals in a family, unless family size is very large (Lorenz 2013; Endelman et al. 2014; Riedelsheimer and Melchinger 2013).

2.2.2 Training and Target Populations Include Related and Unrelated Genotypes

Realistically, models built from single biparental populations are limited in their applications outside of breeding systems with easy access to large population sizes and efficient doubled haploid technologies. The time required to develop and phenotype biparental populations diminishes the potential time savings of implementing genomic selection in place of phenotypic selection. Therefore, methods that combine data across multiple related and/or unrelated families would be valuable for breeders. This can be a challenge because many additional factors come into play when combining data across populations, and adding more individuals to the training population does not necessarily result in greater prediction accuracy as we will discuss below.

The inclusion of related and unrelated genotypes in training and target populations can be further broken down into two scenarios for our purposes here. One scenario includes the development and testing of large families, often consisting of DH lines, as is used in hybrid maize breeding. Families often consist of 150 progenies or more. Under this scenario, it would be possible and appropriate to pool together a few well-chosen families into a single training population. A second common scenario is the development of many, small families. This scenario is common in crops such as soybean and small grains, where crossing is followed by multiple generations of inbreeding followed by visual selection on simply inherited traits and on molecular markers tagging large-effect QTL. The number of progenies per family reaching the yield trial phase is typically small (~20–40) which excludes the possibility of within-family training populations as well as the pooling together of only a few families to form a training population. Rather, training populations would need to be formed by pooling together progenies that are derived from various pedigrees and genetic backgrounds, spanning levels of relatedness. If the populations have been genotyped, ancestral relationships among individuals in the training and the target populations can be used to optimize the selection of training set.

Numerous studies in both plant and animal breeding systems have shown that prediction accuracy suffers when training populations are not related to the target population (Pszczola et al. 2012; Windhausen et al. 2012; Ly et al. 2013; Technow et al. 2013; Albrecht et al. 2014; Lorenz and Smith 2015). Analysis of genomic selection in sheep showed that the strongest predictor of prediction accuracy of each individual was the strength of relationship between the individual being predicted and the top ten relatives in the training population (Clark et al. 2012). In contrast, the mean relationship of the training population to the individual being predicted was a weak predictor of prediction accuracy. Therefore, for an individual to be predicted well using genomic prediction, the training population must include several close relatives to that individual.

Along these same lines, results looking at pooling together large families (first scenario described above) to predict a specific target family have generally

indicated that the best results are obtained when the families being pooled share one parent with the target family. The addition of families sharing one parent with the family-specific training population could increase model accuracy above the family-specific training population, especially if the target family is small in size (Schulz-Streeck et al. 2012; Jacobson et al. 2014). Lehermeier et al. (2014) found that the predictive ability of pooled half-sib training populations could achieve similar accuracy to family-specific training populations, but models built using 375 half-sib individuals were needed to reach the accuracy of models built using only 50 full-sib individuals. Riedelsheimer et al. (2013) found that half-sib training populations that shared one parent in common with the target population only reached 50% of the predictive ability of family-specific training populations. This study, however, only included a limited number of families (six), and in reality, breeding programs would likely include many more families from which to pool data.

The use of data from families unrelated to the target population (family) is more problematic. Training populations consisting of only individuals unrelated to the target population generally result in zero or near-zero prediction accuracy (Riedelsheimer et al. 2013; Jacobson et al. 2014; Lehermeier et al. 2014). Moreover, the addition of unrelated families to a family-specific training population can reduce prediction accuracy compared to the family-specific training population alone (Riedelsheimer et al. 2013; Jacobson et al. 2014) or have no effect despite increasing the training population size by up to sixfold (Zhao et al. 2012). Lorenz and Smith (2015) showed a decline in prediction accuracy when individuals less and less related to the target population were added to the training population. Model accuracy was maximized by using smaller training populations that were more closely related to the target population, and the addition of less related individuals (mostly from a different breeding program) reduced accuracy of predictions for all traits. High marker densities may enhance the sharing of information between families and improve prediction accuracy by pooling unrelated families (Hickey et al. 2014). Hickey et al. (2014) found that training populations consisting of families unrelated to the target family could produce models with accuracies reaching 0.70, but only with population sizes approaching 20,000 individuals and marker numbers greater than 10,000. It is possible that such training populations could be constructed within the seed industry, but to our knowledge, nothing in the public sector has yet come close to this scale.

2.2.3 Training and Target Populations Include Lines from a Diverse Germplasm Collection

Besides predicting the genetic value of progenies comprising an active breeding program, another role of genomic prediction includes the prediction of diverse accessions comprising a germplasm collection. Germplasm collections can be very

large, containing up to hundreds of thousands of plant accessions. Advancements in genotyping have made it possible to genotype entire germplasm collections (Hearne et al. 2015; Song et al. 2015), opening up the possibility of predicting the performance of all accessions (Jarquin et al. 2016). Phenotyping entire collections, on the other hand, is often not feasible.

In this scenario, the training and target populations are essentially two subsets of the same population, and thus the training population should be selected to represent the entire population. Several studies have examined the performance of chosen statistical criteria and accompanying optimization algorithms in choosing informative training populations.

Two criteria for assessing population design derived from mixed linear model theory have been proposed: prediction error variance (PEV) and the generalized coefficient of determination (CD). The PEV quantifies the error of prediction of each random effect in the model. It is a function of the ratio of the model error to genetic variance, the number of times an individual is measured, the number of relatives of the individual included in the dataset, and the strength of their relationship. The CD is defined as the amount of variation in true contrasts of genetic values by predicted contrasts of genetic values, where the contrast is between each individual being predicted in the target population and target population mean (Laloë et al. 1996). Optimizing the reliability of these contrasts rather than of the predictions per se takes the covariances among the individuals comprising the target population into account and thus prevents the selection of closely related individuals for training population formation (Rincent et al. 2012). Because genetic variance is not included in the calculation of PEV, using this method may result in selecting a relatively narrow training population that contains many close relatives. These statistics are calculated for each individual in the target population, and the average value across the target population (i.e., PEVmean and CDmean) is the final optimization criteria.

Criteria related to minimizing PEV have been previously used to optimize data collection in animal breeding programs (Laloë and Phocas 2003; Kuehn et al. 2007). Rincent et al. (2012) expanded the use of these criteria for training population design and genomic selection in plant populations by implementing them in combination with a simple exchange algorithm. An exchange algorithm involves removal and replacement of one individual in the training population, followed by calculation of the optimization criteria (e.g., PEVmean, CDmean) for the newly formed training population. If the removal and replacement results in an improvement measured by the chosen criteria, then the newly added individual remains; else it is removed in place of another randomly sampled individual from the pool of candidates. Rincent et al. (2012) found that an optimization scheme based on a CDmean-optimized training population resulted in models of higher accuracy compared to random sampling. An optimized population of 100 individuals achieved the same prediction accuracy as a randomly selected population of 200, indicating large reduction in costs associated with phenotyping if this method is applied. The CDmean criteria typically outperformed PEVmean and other diversity criteria such as mean genetic relatedness of the selected training population

measured by the genomic relationship matrix. Isidro et al. (2015) applied these same criteria to rice and wheat panels. These authors found that a simple, stratified sampling method that ensured representation of each subpopulation in the training set was superior for the highly structured rice population, whereas the CDmean method was superior for the minimally structured wheat population. This indicates that training population optimization does depend on the population, as well as the trait.

Akdemir et al. (2015) also showed a consistent benefit to optimizing training populations using relationship-based selection procedures. These authors focused on a principal component-based approach that increased computational efficiency and selected training populations with regard to a specified target population, rather than relationships within the training population itself. Their results suggest that such methods hold great potential to help choose maximally informative training populations. Software that implement these methods have been made available to the general user (Rincen et al. 2012; Akdemir et al. 2015).

2.2.4 Sources of Information and Population Genomic Architecture Influence Training Population Design

The overall theme of the literature reviewed above is that relationships between training and target populations are highly important for genomic prediction. It is clear that small training populations can be used, and are likely superior, if they are closely related to the target population. Very large training populations are needed if little to no relationship exists. Some researchers (Campos de los et al. 2013; Habier et al. 2013) have contributed a theoretical basis to the importance of relationships and their interaction with marker density and prediction model. By far the most common methods for performing genomic prediction are ridge regression best linear unbiased prediction (RR-BLUP) and genomic best linear unbiased prediction (G-BLUP). Although these two models are mathematically equivalent under the properties of the multivariate normal distribution (Habier et al. 2013), practitioners of breeding and genomic selection view the information sharing of these models from two different perspectives. From the RR-BLUP perspective, information is shared between training populations and target populations through the LD that exists between markers and QTL. Because of this, as marker-QTL LD increases, prediction accuracy is expected to increase. From the G-BLUP perspective, information is shared via the realized genomic relationships of the training and target individuals, which reflect the higher degree of resemblance of more closely related individuals. Prediction of selection candidates is a function of the weighted sum of phenotypes of individuals in the training population, with weights being proportional to the genomic relationships (Campos de los et al. 2013). Depending on the family structure and distribution of relationships, only a few close relatives could be heavily weighted in the calculation of the genomic predictions, or weights

could be more uniformly distributed among individuals in training populations that are distantly related to the target population.

Ultimately, it is the genetic relationships at causal loci that influence the effectiveness of training populations to predict trait values in prediction sets and not genetic relationships calculated according to markers (Habier et al. 2013; Campos de los et al. 2013). The genomic relationship matrix, calculated using genome-wide markers, is an estimate of the genomic relationship matrix at the causal polymorphisms. Therefore, the accuracy of this estimation is what determines the effectiveness of G-BLUP (Campos de los et al. 2013). The resemblance between the genomic relationship matrix at causal polymorphisms and the estimated genomic relationship matrix based on markers is determined by marker-QTL LD, which in turn is determined by pedigree relationships of the population, population history and diversity, and marker density. Formula for calculating PEV and reliability of predictions using expected genomic relationships based on pedigree data was derived by Henderson (1975). Under these expectations, the reliability of predictions approach 1.0 as the population size goes to infinity. This is even the case if the training population is distantly related to the target population, although the number of individuals required to increase accuracy is much higher compared to the addition of more closely related individuals (Campos de los et al. 2013). Campos et al. (Campos de los et al. 2013) showed that the marker-QTL LD sets an upper limit to prediction accuracy. This limit is lowered when there is a lack of relationship between the training and target populations due to a decrease in marker-QTL LD. This is especially true for distantly related individuals where genomic relationships can be variable with respect to which markers are in high LD with QTL, leading to a major source of error in the G-BLUP model (Hill and Weir 2011). The expected value of realized or pedigree relationship decreases, while the variance of the realized relationship increases (Hill and Weir 2011).

Another way to look at this problem is by partitioning the information contained in the genomic relationship matrix into three components: (1) marker-QTL LD, which is an association between alleles among the population founders; (2) linkage or co-segregation of alleles created by pedigree relationships at QTL; and (3) additive genetic relationships captured by markers (Habier et al. 2013). Habier et al. (2013) used simulations and models to partition these three sources of information. First, they showed that large population sizes and high marker densities are needed to exploit the LD source of information. Secondly, the proportion of accuracy from shared additive genetic relationships is reduced if training populations are expanded by adding unrelated individuals. Accuracy due to LD, however, might be able to compensate for low relatedness if very large training population sizes and/or high marker densities are available. Still, Habier et al. (2013) present an example from cattle data where the increase in the accuracy from LD could not compensate for the loss of information from additive genetic relationships, and an overall decrease in accuracy was observed after the addition of unrelated individuals. However, in their maize example, additive genetic relationship accuracy was not changed by increasing training population size, possibly due to a stronger family structure with many more close relatives in the maize training population.

2.3 Resource Allocation for Phenotyping for Genomic Prediction Model Calibration: To Rep or Not to Rep

A key design aspect of breeding programs is the allocation of resources among breeding trials in terms of population size, number of replications, and locations. Allocation decisions are multifaceted, involving consideration of trait logistics, selection intensity, breeding stage of the materials being tested, and any associated genotyping costs. These decisions affect the genetic gain that is possible, as well as the power to detect QTL or accurately estimate marker effects. Considering selection in general, the fundamental trade-off is between achieving accurate estimates of genotypic value by increasing replication and sampling a greater number of individuals to increase the chance of identifying superior genotypes (Gauch and Zobel 1996). Bos (1983) explored the optimum replication scheme for breeding programs with respect to heritability. Because replication decreases phenotypic variance, it also increases heritability. However, this increase only occurs to a point, after which, fundamental changes to the experimental design would be needed to improve heritability (Gauch and Zobel 1996). Therefore, more replication generally results in better selection outcomes, with the exception of situations where heritability is high and selection intensity is relaxed (Bos 1983). Gauch and Zobel (1996) extended the scope of the Bos (1983) findings to consider the precision of data collected and the relative efficiency of data collected. They found that in experiments with high precision, adding replication beyond two is much less efficient than in lower precision experiments that retain efficiency at greater replication numbers.

When considering markers, the focus changes from identifying the genotypic value of individuals to estimating the additive genetic values of alleles. Knapp and Bridges (1990) identified sources of variation in a QTL mapping experiment and found that increasing population size instead of replication resulted in higher power to detect QTL, particularly when residual genetic variation existed in the population. Other studies reported similar findings, where larger population sizes generally result in higher power of QTL detection, and only moderately sized populations of 150–300 individuals benefit from replication (Schön et al. 2004). Because of the similarities between QTL mapping and MAS, resource allocation recommendations for QTL mapping seem to transfer well to the context of MAS. Moreau et al. (Moreau 2000) showed that larger population sizes resulted in maximum gain from selection when traits were controlled by 5–10 QTL and when genotyping costs were equal to phenotyping costs. The shift toward genomic selection has required a reevaluation of these resource allocation recommendations in the context of a cultivar development program.

In contrast to MAS, genomic selection aims to improve traits that are influenced by many more QTL. In addition, because MAS considers marker effects as fixed and statistical thresholds are used to determine which markers are used to calculate marker scores, the success of MAS is closely related to QTL detection power. Here, we will explore recent published literature that aims to address the resource allocation questions relevant to genomic selection breeding programs and are not

sufficiently addressed by previous MAS studies. Specifically, we review: (1) the value of replication for calibrating genomic selection models, (2) allocation of plots to stages within the breeding cycle, and (3) allocation of plots to within versus across environment replication.

2.3.1 Replication and Plot Allocation for Calibrating Genomic Selection Models

To determine whether resource allocation recommendations for MAS can be extrapolated to genomic selection models, Lorenz (2013) compared the accuracies of genomic selection models (RR-BLUP) and MAS models (ordinary least squares, OLS) under varying resource allocation schemes. The factors studied included total plot budget, relative cost of genotyping in comparison to phenotyping, population size, number of replications, heritability, and percentage of phenotyped individuals. A very clear distinction in resource optimization between GS and MAS models was found. Prediction accuracy was always substantially lower with MAS, and the effect of replication was more apparent for MAS. Within a set budget, the addition of replications and a consequent reduction of total individuals screened lead to a decrease in accuracy with MAS. In contrast, the RR-BLUP model remained fairly constant across different resource allocation scenarios, with low heritability, high marker cost scenarios slightly favoring fewer individuals, and more replication. When the total number of individuals was varied, the accuracy of genomic selection models began to level off around 50–75 individuals, whereas MAS models took many more individuals to achieve moderate prediction accuracies and continued to improve as the numbers increased. These results suggest that the underlying considerations for MAS are different from genomic selection.

2.3.2 Allocation of Resources Across Preliminary and Advanced Breeding Tests

Breeding programs are generally structured with less replications in early generation screening, followed by greater replication, larger scale, and higher-cost trials in later generations (Bernardo 2010). Breeders must take this tiered structure of the breeding program into account when planning for genomic selection implementation. The stage at which genomic selection is implemented can affect genetic gain as well as costs. Bassi et al. (2016) compared a series of wheat breeding schemes that implemented genomic selection starting in generations F_2 , F_3 , F_4 , or F_7 . They found that without including phenotypic selection at some stage in the program, early generation F_2 implementation had the highest potential for gain per year, but also the highest genotyping costs. Longin et al. (2015) found that genomic selection

without a stage of phenotypic selection would only be useful with very high prediction accuracies, possibly unrealistically high accuracies.

When accuracies are low, genomic selection can fill the role of a pretest, whereby a low selection intensity is applied to remove the lowest performing individuals (Longin et al. 2015). Most studies have focused on overall accuracy of genomic selection, without considering the effectiveness of these selection schemes to accurately remove the worst individuals or include the best. Endelman et al. (2014) proposed using a response to selection metric R_{\max} based on the maximum genotypic value of selections instead of R_{mean} based on mean values for selection to analyze genetic gain in preliminary yield trials. Because the mean of the selected population decreases as more individuals are selected, the R_{mean} measure of genetic gain may encourage overly stringent selection in early generations that have less precise phenotypic estimates. Additional studies are needed to expand on the use of genomic selection for early generation screening.

In contrast to early generation genomic selection, Bassi et al. (2016) compared intermediate and later generation schemes. They found that implementation in the F_3 and F_4 was a good compromise between no stage of phenotypic selection and the minimal benefits of F_7 implementation. While F_7 implementation might be attractive to breeders because of its ease of implementation and lower genotyping costs, this scheme resulted in minimal benefit over phenotypic selection alone. Longin et al. (2015) concluded that for traits such as yield in wheat, with prediction accuracies of approximately 0.3, one stage of genomic selection followed by one stage of phenotypic selection provides the best compromise between genomic and phenotypic selection.

2.3.3 Across Environment Versus Within Environment Replication

For simplicity, much of the literature surrounding the topic of resource allocation focuses on trade-offs within single environments, but the distribution of plot resources across environments is a major consideration for breeders. Riedelsheimer and Melchinger (2013) attempted to tackle this issue by developing a resource allocation planning tool for distributing plot resources across and within environments. Their tool is limited to a single cycle of selection in biparental populations, and it requires some degree of estimation based on previous experimental data. Their calculations extend those developed by Daetwyler et al. (2008) to include considerations of multi-environment testing. They found that larger budgets favored more environments, with a lower proportion of plots being allocated to the training set. As the budget decreased, the training set became a larger proportion of the plots, and the number of environments tested decreased. Furthermore, they emphasize that under low-budget scenarios, the optimization has much less

flexibility than under large-budget scenarios. Overall, their findings suggest relatively few environments are needed for high prediction accuracy.

Endelman et al. (2014) looked at the effect of spreading replicates across locations in preliminary yield trials under fixed budgets. They found that accuracy increased as individuals were replicated across locations, but under a fixed budget, the optimum accuracies were obtained without replication across locations, unless the budget forced a relatively small training population size. That is, each individual should be phenotyped in only one environment, and population size should be maximized to the extent the total number of plots across environments allows. Markers provide the connectivity between environments. In contrast, across environment phenotypic estimates based on phenotyping alone were poor when individuals were phenotyped in single environments. This reinforces the idea that shared marker information does provide the connectivity between individuals, providing potential cost savings for breeders implementing genomic selection.

2.3.4 Conclusions

The role of phenotyping in plant breeding is rapidly shifting from its previous sole purpose of providing information for making breeding line advancement, parent selection, and variety release decisions to providing the necessary data to train genomic prediction models to enable genomic selection. As this new role of phenotyping increases in relative importance, plant breeders need to rethink how they design field trials, allocate plot resources to genotypes, and which individuals are included in field trials. This review provides a short and simple introduction to this literature. We have two basic conclusions at this time: (1) Training population selection and design should take genetic relationships with the target population into consideration, and optimization criteria such as PEVmean and CDmean combined with exchange algorithms are useful methods for selecting training populations. (2) The number of individuals phenotyped should be maximized by allocating only one field plot to each genotype in most situations. Further research is needed to develop a comprehensive theoretical framework for phenotyping for genomic selection.

References

- Akdemir D, Sanchez JI, Jannink J-L (2015) Optimization of genomic selection training populations with a genetic algorithm. *Genet Sel Evol* 47:38. doi:[10.1186/s12711-015-0116-6](https://doi.org/10.1186/s12711-015-0116-6)
- Albrecht T, Auinger H-J, Wimmer V et al (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theor Appl Genet* 127:1375–1386. doi:[10.1007/s00122-014-2305-z](https://doi.org/10.1007/s00122-014-2305-z)
- Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. *Trends Plant Sci* 19:52–61. doi:[10.1016/j.tplants.2013.09.008](https://doi.org/10.1016/j.tplants.2013.09.008)

- Bassi FM, Bentley AR, Charmet G et al (2016) Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.) *Plant Sci* 242:23–36. doi:[10.1016/j.plantsci.2015.08.021](https://doi.org/10.1016/j.plantsci.2015.08.021)
- Bernardo R (2010) Breeding for quantitative traits in plants, 2nd edn. Stemma Press, Woodbury, MN
- Bernardo R, Yu J (2007) Prospects for Genomewide selection for quantitative traits in maize. *Crop Sci* 47:1082–1090. doi:[10.2135/cropsci2006.11.0690](https://doi.org/10.2135/cropsci2006.11.0690)
- Bos I (1983) The optimum number of replications when testing lines or families on a fixed number of plots. *Euphytica* 32:311–318. doi:[10.1007/BF00021439](https://doi.org/10.1007/BF00021439)
- Cabrera-Bosquet L, Crossa J, von Zitzewitz J et al (2012) High-throughput Phenotyping and genomic selection: the Frontiers of crop breeding ConvergeF. *J Integr Plant Biol* 54:312–320. doi:[10.1111/j.1744-7909.2012.01116.x](https://doi.org/10.1111/j.1744-7909.2012.01116.x)
- Campos de los G, Vazquez AI, Fernando R et al (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet* 9:e1003608. doi:[10.1371/journal.pgen.1003608](https://doi.org/10.1371/journal.pgen.1003608)
- 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. doi:[10.1186/1297-9686-44-4](https://doi.org/10.1186/1297-9686-44-4)
- Combs E, Bernardo R (2013) Accuracy of Genomewide selection for different traits with constant population size, heritability, and number of markers. *Plant Genome* 6:0. doi: [10.3835/plantgenome2012.11.0030](https://doi.org/10.3835/plantgenome2012.11.0030)
- Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3:e3395. doi:[10.1371/journal.pone.0003395](https://doi.org/10.1371/journal.pone.0003395)
- Endelman JB, Atlin GN, Beyene Y et al (2014) Optimal Design of Preliminary Yield Trials with genome-wide markers. *Crop Sci* 54:48–59. doi:[10.2135/cropsci2013.03.0154](https://doi.org/10.2135/cropsci2013.03.0154)
- Gauch HG, Zobel RW (1996) Optimal replication in selection experiments. *Crop Sci* 36:838–843. doi:[10.2135/cropsci1996.0011183X003600040002x](https://doi.org/10.2135/cropsci1996.0011183X003600040002x)
- Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics* 194:597–607. doi:[10.1534/genetics.113.152207](https://doi.org/10.1534/genetics.113.152207)
- Hearne S, Franco J, Chen J et al (2015) Genome wide assessment of maize Genebank diversity: synthesis of next generation technologies and GIS based approaches. San Diego, USA
- 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
- Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31:423–447. doi:[10.2307/2529430](https://doi.org/10.2307/2529430)
- Hickey JM, Dreisigacker S, Crossa J et al (2014) Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Sci* 54:1476–1488. doi:[10.2135/cropsci2013.03.0195](https://doi.org/10.2135/cropsci2013.03.0195)
- Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet Res* 93:47–64. doi:[10.1017/S0016672310000480](https://doi.org/10.1017/S0016672310000480)
- Isidro J, Jannink J-L, Akdemir D et al (2015) Training set optimization under population structure in genomic selection. *Theor Appl Genet* 128:145–158. doi:[10.1007/s00122-014-2418-4](https://doi.org/10.1007/s00122-014-2418-4)
- Jacobson A, Lian L, Zhong S, Bernardo R (2014) General combining ability model for Genomewide selection in a Biparental cross. *Crop Sci* 54:895–905. doi:[10.2135/cropsci2013.11.0774](https://doi.org/10.2135/cropsci2013.11.0774)
- Jannink J-L (2005) Selective Phenotyping to accurately map quantitative trait loci. *Crop Sci* 45:901–908. doi:[10.2135/cropsci2004.0278](https://doi.org/10.2135/cropsci2004.0278)
- Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* 9:166–177. doi:[10.1093/bfpg/elq001](https://doi.org/10.1093/bfpg/elq001)
- Jarquín D, Specht J, Lorenz A (2016) Prospects of genomic prediction in the USDA soybean Germplasm collection: historical data creates robust models for enhancing selection of accessions. *G3 GenesGenomesGenetics*:g3.116.031443. doi:[10.1534/g3.116.031443](https://doi.org/10.1534/g3.116.031443)

- Johnson R (2004) Marker-assisted selection. *Plant Breeding Reviews*. John Wiley & Sons, In
- Knapp SJ, Bridges WC (1990) Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for Unreplicated and replicated progeny. *Genetics* 126:769–777
- Kuehn LA, Notter DR, Nieuwhof GJ, Lewis RM (2007) Changes in connectedness over time in alternative sheep sire referencing schemes. *J Anim Sci* 86:536–544. doi:[10.2527/jas.2007-0256](https://doi.org/10.2527/jas.2007-0256)
- Laloë D, Phocas F (2003) A proposal of criteria of robustness analysis in genetic evaluation. *Livest Prod Sci* 80:241–256. doi:[10.1016/S0301-6226\(02\)00092-1](https://doi.org/10.1016/S0301-6226(02)00092-1)
- Laloë D, Phocas F, Méniissier F (1996) Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genet Sel Evol* 28:1–20. doi:[10.1186/1297-9686-28-4-359](https://doi.org/10.1186/1297-9686-28-4-359)
- Lehermeier C, Krämer N, Bauer E et al (2014) Usefulness of Multiparental populations of maize (*Zea mays* L.) for genome-based prediction. *Genetics* 198:3–16. doi:[10.1534/genetics.114.161943](https://doi.org/10.1534/genetics.114.161943)
- 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. doi:[10.1007/s00122-015-2505-1](https://doi.org/10.1007/s00122-015-2505-1)
- Lorenz AJ (2013) Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. *G3 GenesGenomesGenetics* 3:481–491. doi:[10.1534/g3.112.004911](https://doi.org/10.1534/g3.112.004911)
- Lorenz AJ, Smith KP (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Sci* 55:2657–2667. doi:[10.2135/cropsci2014.12.0827](https://doi.org/10.2135/cropsci2014.12.0827)
- Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor Appl Genet* 120:151–161. doi:[10.1007/s00122-009-1166-3](https://doi.org/10.1007/s00122-009-1166-3)
- Ly D, Hamblin M, Rabbi I et al (2013) Relatedness and genotype \times environment interaction affect prediction accuracies in genomic selection: a study in cassava. *Crop Sci* 53:1312–1325. doi:[10.2135/cropsci2012.11.0653](https://doi.org/10.2135/cropsci2012.11.0653)
- 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. doi:[10.1111/pbr.12317](https://doi.org/10.1111/pbr.12317)
- Massman JM, Jung H-JG, Bernardo R (2013) Genomewide selection versus marker-assisted recurrent selection to improve grain yield and Stover-quality traits for cellulosic ethanol in maize. *Crop Sci* 53:58–66. doi:[10.2135/cropsci2012.02.0112](https://doi.org/10.2135/cropsci2012.02.0112)
- Moreau L, Lemarie S, Charcosset A, Gallais A (2000) Economic efficiency of one cycle of marker-assisted selection. *Crop Sci* 40:329–337. doi:[10.2135/cropsci2000.402329x](https://doi.org/10.2135/cropsci2000.402329x)
- 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. doi:[10.3168/jds.2011-4338](https://doi.org/10.3168/jds.2011-4338)
- Riedelsheimer C, Endelman JB, Stange M et al (2013) Genomic predictability of interconnected Biparental maize populations. *Genetics* 194:493–503. doi:[10.1534/genetics.113.150227](https://doi.org/10.1534/genetics.113.150227)
- Riedelsheimer C, Melchinger AE (2013) Optimizing the allocation of resources for genomic selection in one breeding cycle. *Theor Appl Genet* 126:2835–2848. doi:[10.1007/s00122-013-2175-9](https://doi.org/10.1007/s00122-013-2175-9)
- Rincent R, Laloë D, Nicolas S et al (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize Inbreds (*Zea mays* L.) *Genetics* 192:715–728. doi:[10.1534/genetics.112.141473](https://doi.org/10.1534/genetics.112.141473)
- Schön CC, Utz HF, Groh S et al (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics* 167:485–498

- Schopp P, Muller D, Technow F, Melchinger A (2017) Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium. *Genetics* 205:441–454
- Schulz-Streeck T, Ogutu JO, Karaman Z et al (2012) Genomic selection using multiple populations. *Crop Sci* 52:2453–2461. doi:[10.2135/cropsci2012.03.0160](https://doi.org/10.2135/cropsci2012.03.0160)
- Sen S, Johannes F, Broman KW (2009) Selective genotyping and Phenotyping strategies in a complex trait context. *Genetics* 181:1613–1626. doi:[10.1534/genetics.108.094607](https://doi.org/10.1534/genetics.108.094607)
- Song Q, Hyten DL, Jia G et al (2015) Fingerprinting soybean Germplasm and its utility in genomic research. *G3 GenesGenomesGenetics* 5:1999–2006. doi:[10.1534/g3.115.019000](https://doi.org/10.1534/g3.115.019000)
- Technow F, Bürger A, Melchinger AE (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for Heterotic groups. *G3 GenesGenomesGenetics* 3:197–203. doi:[10.1534/g3.112.004630](https://doi.org/10.1534/g3.112.004630)
- Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. *Genet Res* 75:249–252. doi: null
- Windhausen VS, Atlin GN, Hickey JM et al (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3 GenesGenomesGenetics* 2:1427–1436. doi:[10.1534/g3.112.003699](https://doi.org/10.1534/g3.112.003699)
- Zhao Y, Gowda M, Liu W et al (2012) Accuracy of genomic selection in European maize elite breeding populations. *Theor Appl Genet* 124:769–776. doi:[10.1007/s00122-011-1745-y](https://doi.org/10.1007/s00122-011-1745-y)

Genomic Selection for Crop Improvement
New Molecular Breeding Strategies for Crop
Improvement

Varshney, R.K.; Roorkiwal, M.; Sorrells, M.E. (Eds.)
2017, XII, 258 p. 14 illus., 8 illus. in color., Hardcover
ISBN: 978-3-319-63168-4