TRAINING SET OPTIMIZATION UNDER POPULATION STRUCTURE IN GENOMIC SELECTION

A Dissertation
Presented to the Faculty of the Graduate School of Cornell University in Partial Fulfillment of the Requirements for the Degree of Master of Science

by
Julio Isidro Sánchez
January 2015
The optimization of the training set (TRS) in genomic selection has received much interest in both animal and plant breeding, because it is critical to the accuracy of the prediction models. In this study, five different TRS sampling algorithms, stratified sampling, mean of the coefficient of determination (CDmean), mean of predictor error variance (PEVmean), stratified CDmean (StratCDmean) and random sampling, were evaluated for prediction accuracy in the presence of different levels of population structure. In the presence of population structure, the most phenotypic variation captured by a sampling method in the TRS is desirable. The wheat dataset showed mild population structure, and CDmean and stratified CDmean methods showed the highest accuracies for all the traits except for test weight and heading date. The rice dataset had strong population structure and the approach based on stratified sampling showed the highest accuracies for all traits. In general, CDmean minimized the relationship between genotypes in the TRS, maximizing the relationship between TRS and the test set. This makes it suitable as an optimization criterion for long-term selection. Our results indicated that the best selection criterion used to optimize the TRS seems to depend on the interaction of trait architecture and population structure.
BIOGRAPHICAL SKETCH

Julio Isidro Sánchez was born in Lora del Río (Sevilla, Spain), the second of three children. He earned his Bachelor of Biology from Seville University (Spain) in 2001 and he joined the plant research after reading the Irish famine disaster in 2002. In 2003 he joined the Plant Physiology doctoral program from the University of Granada, Spain. He received his Master of Agrobiology in 2005 from the University of Granada. During his PhD, Julio had two interships in Canada and Argentina to learn key research questions from his thesis. In 2008, he obtained his PhD and moved to Swift Current, SK, Canada to joined the Semiarid Praise Agriculture Research Center as a postdoctoral researcher. In 2012 he enrolled in the master graduate program at Cornell University to study optimization of training population in Genomic Selection.
a mis padres,
a mis hermanos,
a Vanesa Soto del Real,
a todos los bollakidos, breicons, brankatans y bètics del universo.
ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Mark Sorrells for their guidance, patience, encouragement and support. Dr. Jean-Luc Jannink was very understanding and always available for great instruction throughout my Master. I would like to extend my sincere appreciation to all members of Mark and Jean-Luc’s lab for their excellent advice, assistance and support in the last two years. Especially, to my friend Dr. Deniz Akdemir who helped me in all the statistical steps analysis of my project. I would also like to express my appreciation to all the faculty members, students and staffs at the Plant Breeding Department for providing me with such a professional research environment. Finally, I want to particularly thank to Vanesa, my parents and my brother and sister. It is their support and love that make all what I have accomplished so far possible. This work was supported by a Master scholarship from the Regional Ministry of Economy, Innovation and Science of Andalusia, Spain.
# TABLE OF CONTENTS

| Section                                                                 | Page |
|------------------------------------------------------------------------|------|
| Biographical Sketch                                                    | iii  |
| Dedication                                                             | iv   |
| Acknowledgements                                                       | v    |
| Table of Contents                                                      | vi   |
| List of Tables                                                         | vii  |
| List of Figures                                                        | viii |
| **1 Introduction**                                                     | 1    |
| 1.1 Genomic Selection steps                                            | 3    |
| 1.2 Why Genomic Selection?                                             | 6    |
| 1.3 Factors affecting Genomic Selection                               | 7    |
| 1.3.1 Heritability                                                    | 8    |
| 1.3.2 Linkage disequilibrium                                          | 10   |
| 1.3.3 Population Size                                                 | 12   |
| 1.3.4 Marker density and types                                         | 12   |
| 1.3.5 Statistical Models                                              | 15   |
| 1.3.6 Relationship between TRS and TS                                 | 21   |
| 1.3.7 Final summary factors affecting GS                              | 22   |
| 1.4 Drawbacks Genomic Selection                                       | 23   |
| **2 Training set optimization under population structure in genomic** | 25   |
| 2.1 Abstract                                                           | 25   |
| 2.2 Introduction                                                       | 26   |
| 2.3 Material and Methods                                               | 30   |
| 2.3.1 Genetic dataset material                                         | 30   |
| 2.3.2 Training set optimization methods                                | 32   |
| 2.3.3 Method 1-optimization based on stratified sampling by cluster    | 34   |
| 2.3.4 Method 2-optimization criterion based on CDmean and PEVmean     | 35   |
| 2.3.5 Method 3-optimization criterion based on stratified sampling CDmean by cluster | 38   |
| 2.3.6 Heritability calculation and statistical software                | 38   |
| 2.4 Results                                                            | 39   |
| 2.5 Discussion                                                         | 46   |
| 2.6 Acknowledgements                                                   | 53   |
| **Bibliography**                                                       | 54   |
# LIST OF TABLES

1.1 Main differences between MAS versus GS .............................................. 3
1.2 Linkage disequilibrium decay on rice, wheat, maize and sorghum. 10

2.1 Germplasm description summary and heritabilities values for each trait. ......................................................... 30
2.2 Descriptions of wheat and rice clusters identified using hierarchical clustering model analysis. NY New York, Ont Ontario, OH Ohio, MI Michigan, IN Indiana, VA Virginia, IN Indica, CH China, PH Philippines, BR Brasil, US United Stated, AR Argentina, CO Congo, NI Nigeria, EU European Union, JA Japan. 40
LIST OF FIGURES

1.1 Population sets in GS: Calibration Set CS; Training Set, TRS; Validation Set VS; Test Set, TS. ................................. 4
1.2 Genomic Selection scheme. ............................................... 5
1.3 Genomic Selection snapshot. ........................................... 6
1.4 Factors affecting Genomic Selection. ................................. 8
1.5 Number of SNPs needed to reach a certain accuracy depending on TRS size. .......................................................... 13
1.6 Statistical GS methods based on the subject of estimation. .... 16

2.1 Example of optimization of training population set (TRS) scheme in the wheat dataset. The three methods are represented in numerical circles. Number 1 represents the stratified sampling method, number 2 the CDmean and PEVmean approaches and number 3 the stratified CDmean. C1C4 acronyms indicate the number of cluster after analysis. The optimization process was repeated over 50 runs and in a TRS size of 25, 50, 100, 200 and 300 genotypes. ................................................................. 33

2.2 Plots of the first two principal components and the cluster analysis using R with 38,893 GBS and 36,901 SNPs markers on a wheat and b rice germplasm. Each solid circle represents a genotype and the colors indicate clusters membership. Legends summarize the distribution of the subpopulations for both germplasm. Number of genotypes per cluster is indicated in parenthesis. .... 41

2.3 Accuracies of the predictions of the TS genotypes in the wheat germplasm. The calibration sets were defined by maximizing CDmean; minimizing PEVmean; maximizing CDmean within cluster; stratified proportional sampling and random sampling. Four different population sizes (25, 50, 100, 200 and 300) were used for the optimization algorithm in five different traits (a yield, b test weight, c lodging, d heading date, e plant height). Standard error is indicated for each point over the 50 runs. Optimization of CDmean, PEVmean and StratCDmean was made with the heritability measured for each trait in each germplasm. 42
2.4 Accuracies of the predictions of the TS genotypes in the rice germplasm. The calibration sets were defined by maximizing CDmean; minimizing PEVmean; maximizing CDmean within cluster; stratified proportional sampling and random sampling. Four different population sizes (25, 50, 100, 150 and 175) were used for the optimization algorithm in four different traits (a florets per panicle, b flowering time, c plant height, d protein content). Standard error is indicated for each point over the 50 runs. Optimization of CDmean, PEVmean and StratCDmean was made with the heritability measured for each trait in each germplasm.

2.5 Genotypes selected from the optimization algorithm over the 50 run are plotted on the principal components analysis in wheat and rice germplasm. The genotypes were selected based on CDmean (a, d), PEVmean (b, e) and StratCDmean (e, f). Green dots represent the genotypes selected by the algorithm over the 50 runs. Red dots indicate those genotypes that were selected more than 15 and 27 times in wheat and rice germplasm, respectively.

2.6 Relative performance of CDmean versus the random sampling method as a function of the ratio of the phenotypic variance between the top 40 most selected genotypes by CDmean and the overall phenotypic variance in wheat (red-cross) and rice (blue-triangle). Traits per germplasm are indicated in wheat as YLD yield, TWT test weight, LODG lodging, HD heading date, HT plant height and in rice as FLORETS florets per panicle, FLOWERING flowering time, PH plant height, PROTEIN protein content.
Genomic is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes. It refers to the structure and function of an entire genome. Selection is a process in which environmental or genetic influences determine which types of organism thrive better than others, regarded as a factor in evolution. From a plant breeding view, selection consist in choosing the favorable alleles in a population by selecting the genotypes with better fitness for a particular or wide range of traits [24]. Therefore, genomic selection (GS) or genome wide selection (GWS) refers to select genotypes using the genomic information to make selection.

The term GS was first introduced by Haley and Visscher at the 6th World Congress on Genetic Applied to Livestock Production at Armidale, Australia in 1998. Nevertheless, it was not until 2001 when Meuwissen [48] applied the concept of GS to make predictions. Genomic selection refers to the use of a dense panel of genetic markers covering the whole genome to estimate the breeding value (BV) of selection candidates. As the markers are closely spaced, a Quantitative Trait Loci (QTL) located anywhere in the genome will be in linkage disequilibrium (LD) with at least one marker. Hence, GS is not more than marker assisted selection on a genome wide scale. In the traditional selection strategy, pedigree information and phenotypic records are the basic information sources for the prediction of future genotypes. With the advent of high throughput genotyping technologies, GS has become a very intense field of research during the last few years, since it is possible to include information about the variation
of DNA sequences among individuals in the estimation of the BVs (Vazquez et al. 2010). Genomic selection can increase the rates of genetic gain through increased accuracy of estimated BVs, reduction of generation interval, and better utilization of available genetic resources through genome-guided mate selection [51, 75, 84].

Molecular marker technology have been using since the 80’s to help plant breeders to make selection. Quantitative trait loci identification ([61], marker-assisted introgression [10], marker-assisted selection (MAS) [60] and marker-assisted recurrent selection (MARS) [6] have been used as methods to make selections. The main difference of these methods with regard to GS is that, a limited fraction of the genetic variation is explained by the identification of QTL. That is to say, MAS uses only a small number of markers to predict BVs and therefore have poor predictive ability due to that only a limited proportion of genetic variance can be captured by the markers [34]. The reason why a limited fraction of the genetic variation is explained by the identified QTL is that, in order to identify a QTL, very stringent tests for statistical significance have to be performed. These tests are stringent because many positions for the presence of a QTL are tested, and if the tests were not stringent, many false positive would be found. Table 1.1 highlights the main differences between MAS and GS.

The idea of GS is to omit the significance testing, and simply estimates the effects of all genes or chromosome positions simultaneously [48]. Different statistical methods are used to estimate the marker effects. Genomic selection is based on the estimation of Genomic Estimate Breeding Values (GEBVs). These are estimated as the sum of SNP effects of single markers or haplotypes within a chromosomal segment [48, 83]. These single genetic markers or haplotypes are
| Selection based on | MAS   | GS    |
|--------------------|-------|-------|
| Genetic Trait architecture Identification | Markers | GEBVs |
| Threshold Identification | Qualitative | Quantitative |
| Markers needed | QTL | YES |
| Variance Explained | Small | Large |
| Genes involved | Low | High |

Table 1.1: Main differences between MAS versus GS

assumed to explain most of the genetic variance contributed by the QTL [47]. Unavailability of dense marker arrays and the higher cost associated with the genotyping these markers have been the major obstacles for the implementation of genomic selection [48].

### 1.1 Genomic Selection steps

Four types of populations can be differentiating from a base population under selection in GS (Fig.1.1).

1. Calibration set population (CS).
2. Training set population (TRS).
3. Validation set (VS).
4. Test set (TS)

The population is divided into calibration set (CS) and training set (TS) and then CS is separated into training set (TRS) and validation set (VS) population.
In the CS all genotypes are phenotyped and genotyped. Genotypes belonging to the TRS will create the prediction equation. The remaining genotypes in the CS are reserved for evaluating the performance of the model, that is to say the VS. The TS is a set of genotypes from the base population where predictions will be made to make selection (1.1).

![Diagram of population sets in GS: Calibration Set CS; Training Set, TRS; Validation Set VS; Test Set, TS.](image)

The basis steps in GS are:

1. Population under selection is divided in CS and TS. Lines in CS are all phenotyped and genotyped.

2. Calibration set is further divided on TRS and VS. Training set population is where the prediction equation will be created. The prediction equation is essentially a set of estimated marker-effects that can be used to predict GEBVs for non-phenotyped individuals based on their genotype. Validation set will evaluate the performance of the statistical model before make predictions in the TS. The statistical model will simultaneously estimates
allele effects at all loci. A broad range of statistical methods can be used for estimation of marker effects [18, 51].

3. Once the model is validated and the effects of markers in the TRS estimated, GEBVs is calculated multiplying the genotypic information present in the TS by the marker effects previously estimated in the TRS. Summing up all the effects will predict the total BV of the genotype under study (GEBVs). Selection can then be based on GEBVs. Genotypic estimated breeding values is equal to \( GEBV = \sum_{i=1}^{p} X_i \hat{g}_i \) where \( p \) is the number of markers (chromosome segments), \( X_i \) is the doses of marker at each loci and \( g \) is the estimate of marker effect at loci \( i \) (Fig. 1.2 and 1.3).

![Figure 1.2: Genomic Selection scheme.](image-url)
### 1.2 Why Genomic Selection?

The main reasons of why GS have been widely adapted to many breeding programs and the growing interest in the scientific community are:

1. The decrease in the cost of genotyping allowed the implementation of GS in many breeding programs.

2. The increase in the cost of phenotyping allowed finding alternatives to reduce this negative drawback.

3. Most of the traits are governed by QTL with small effects and GS is suitable for quantitative traits rather than qualitative.

4. Simulation and more recently empirical studies, have demonstrated the superiority of GS versus traditional methods.
5. Perhaps, the most important characteristic of GS is that allows increasing the genetic gain, since reducing the generation interval in the breeding program.

1.3 **Factors affecting Genomic Selection**

The main factors affecting the accuracy of GS are:

1. Heritability.
2. Linkage disequilibrium (LD).
3. Population size.
4. Marker density and types.
5. Statistical model.
6. Relationship between TRS and TS.

All these factors do not act independently. There is a complex network relationship among all the factors affecting GS (Fig. 1.4).
1.3.1 Heritability

Heritability ($h^2$) is the degree of correspondence between phenotype and BVs. To estimate BVs we need to estimate its progeny. If the correlation between the BV of the progeny and the parents are high then the heritability is high (normally with qualitative traits). If dominance, epistatic and environment effects are large then heritability is low (quantitative traits). Heritability depends strictly on the population you are working with. Perhaps, the most important characteristic of heritability is its predictive role, expressing the reliability of the phenotypic value as a guide to the BV. Only phenotype can be measured, but it’s the BV that influences the next generation. If a breeder chooses individuals to be parents according to their phenotypic value, his success in changing the pop-
ulation can be predicted only from knowledge of the degree of correspondence between phenotypic values and breeding values. Therefore, heritability has an important impact in the accuracies of the GS models. Daetwyler [14, 15] derived a formula to calculate the correlation between true and estimated additive genetic values (i.e., accuracy).

\[ r_{\hat{g}g} = \sqrt{\frac{Nh^2}{Nh^2 + M_e}} \]

where \( r_{\hat{g}g} \) is the correlation between true and estimated additive genetic values, \( N \) is the number of genotypes in the TRS and is the number of independent chromosome segments where \( M_e = 2N_eL \) and \( N_e \) is the effective population size and \( L \) is the genome length in Morgans [36].

Daetwyler formula indicates that the higher the heritability values the better the accuracies in GS. Therefore, there is a positive relationship between accuracy and heritability in GS. Nevertheless, we still can have high accuracies with low heritability values. When the population size is big and the heritability is low, accuracies can be maintaining constant, since is the product of population size and heritability the driver of Daetwyler’s equation. The above formula is expected to overestimate the accuracy, because ignores that not all of the genetic variance may be explained by the markers, because of insufficient marker density. A critical parameter in the formula is also \( M_e \), because as \( M_e \) increases, accuracy decreases. The more related a population is, the lower the \( M_e \) and the higher the accuracy that can be achieved [36].
1.3.2 Linkage disequilibrium

Linkage disequilibrium is the nonrandom association of alleles at different loci. That is to say, it is the correlation between polymorphisms that is caused by their shared history of mutation and recombination. In a large randomly mated population with loci segregating independently, but in the absence of selection, mutation, or migration, polymorphic loci will be in linkage equilibrium [27]. Linkage refers to physical distribution of alleles in the same chromosome. Dis-equilibrium refers to a statistical concept of correlation, that is, when 2 or more loci are not in Hardy-Weinberg equilibrium. Two loci can have a high correlation and not be in LD, for example, if those loci are in different chromosomes but they always passed to the progeny together. The response to GS relies on LD between specific alleles and QTL. In GS all genes are expected to be in LD with at least some of the markers, therefore LD is a key factor that is driving the genomic prediction process. Linkage disequilibrium decays when recombination occurs, consequently the minimum number of markers to achieve genome-wide coverage depends on LD decays rates which vary widely across species, populations and genomes (Table 1.2).

| Crop       | LD decay     |
|------------|--------------|
| Rice       | 75 – 500 kb  |
| Wheat      | 50 – 100 Mb  |
| Maize      | 0.1 – 1.5 kb |
| Sorghum    | 15 – 20 kb   |

Table 1.2: Linkage disequilibrium decay on rice, wheat, maize and sorghum.

From the coalescence theory, for a population in recombination-drift equilibrium, the LD between marker and QTL is a function of $N_e \times c$, where $c$ is the recombination rate between the loci and $N_e$ is the effective population size. For example, the LD structure will be the same for a population with $N_e = 50$.
and 500 Single nucleotide polymorphism (SNPs) per Morgan (M), compared to a population with $N_e = 500$ and 5,000 SNPs per Morgan. Two populations have same LD if the ratio $\frac{\text{number markers per cM}}{N_e}$ is equal between them.

Effective population size is the number of breeding individuals in a randomly mating population that experiences the same amount of genetic drift, as the base population. It can be also expressed as the minimum number of individuals having the maximum genetic variability from the base population. The lower $N_e$ the more rapidly drift operates, increasing LD among loci. The higher $N_e$ the lower LD. Therefore, marker density needs to scale with $N_e$. If you increase $N_e$ you have to increase the number of markers to be in LD with every QTL. From here, it can be deduced that the number of markers effects to estimate should be proportional to the effective population size [83]. At the same time, LD of the marker-QTL should be conserved between TRS and TS to achieve high accuracies [36]. If LD decay is more rapid in TS population than in TRS, accuracies will decrease because markers could segregate from the QTL they estimated in the TRS. Hence, TRS should have equal or greater LD extend than TS population to reduce the accuracy loss from genetic divergence.

Weedon [89] showed that the ideal of having a marker in perfect LD with each QTL is complicated by the fact that almost all the genetic variation of quantitative traits is due to genes with a small effect. This indicates that i) there are very many QTL so the effect of a single marker may be due to a several QTL in the region ii) estimation of single genes will be difficult due to their small size and LD with other genes and iii) the assumption of constant genetic variance across the genome may be quite realistic [88]. Nevertheless, although LD plays the key in GS, it has been demonstrated previously [42] that BVs can also be pre-
dicted in the absence of linkage between markers and QTL, since the markers can predict family relationships between the individuals.

1.3.3 Population Size

Large population size is needed to accurately estimate marker effects and to maximize the number of QTL in LD with at least one marker. This is clearly seen from Daetwyler’s formula.

\[ r_{gg} = \sqrt{\frac{Nh^2}{Nh^2 + Me}} \]

Therefore, high population size is needed if the trait heritability is low and vice versa. Several studies have found that TRS size impact accuracy more than marker number [62, 83]. Satisfactory marker density and TP size depend on QTL number and trait heritability. The impact of TP size is illustrated in Figure 1.5.

1.3.4 Marker density and types

Developments in marker discovery and high throughput genotyping technology have allowed the use of GS in a routinely manner. With the availability of high-density marker maps and cost-effective genotyping, GS methods provide faster genetic gain than can be achieved by current selection methods based on phenotypes and the pedigree [36, 51, 66]. There are different ways of using marker information for GS.
1. Considering each marker allele at a single locus to be a different haplotype, without any relationship among haplotypes. In this way, BVs are estimated directly for the marker alleles [93].

2. Construct haplotypes from two alleles at adjacent markers and assuming a zero relationship between haplotypes at the same locus [66].

3. Considering two or more surrounding marker alleles and derive identical-by-descent (IBD) probabilities between the different haplotypes at the same locus [66].

The first model is suited for applications where you only consider two haplotypes at a locus. i.e. double-haploid populations with only two segregating genotypes at each locus [93]. The advantage of this model is that determining the linkage phase of the haplotypes is not required and the markers do not need to be mapped. However, no new haplotypes arise as a result of recombination, while such an event actually might change the linkage between the marker and
the QTL alleles. For outbred populations, where the association between markers and QTL might be different in different families, the first model is perhaps less well suited.

In general, accuracy increases with increasing marker density, since higher marker density will help to increase the probability of finding markers that are in consistent LD with the QTL. The lower accuracy normally detected for lower marker densities may be explained by the smaller probability of LD between the markers and the QTL; hence, only a smaller fraction of genetic variation can be explained [15]. Therefore, application of GS regarding the number of markers will be different depending on the type of population because of the extent LD. For example, marker density requirements for biparental populations are much lower than for a set of lines with broad genetic diversity. There might have two main reasons of why an increase of marker density does not translate to an augment of accuracy i) the extent of LD is so high that low density can capture all the genetic variation and ii) training set population size is not big enough to make advantages of more markers.

Meuwissen [85] suggested that an increase in marker density should be coupled with higher TRS size to result in higher accuracies. Nevertheless, recent studies [4] indicated that it is more important in the short term to increase the TRS size rather than to increase marker density to increase accuracy. In general, TRS size should scale with marker numbers to successfully capture the additional information provided by increasing marker density [70].

In GS the most used markers are Genotyping by sequencing (GBS), SNPs, diversity array technology (DarT) and single sequence repeats (SSRs). Recently, GBS have raised a great interest in the scientific community because of the
high flexibility and low cost, which make GBS an ideal approach for genomics-assisted breeding [71]. Solberg [83] found that the SNP markers required a density of two to three times that of the SSR markers.

1.3.5 Statistical Models

In standard genetic models, phenotypes are regressed with a linear model. In matrix notation

\[ y_i = \sum_{j=1}^{p} x_{ij} \beta_j + \epsilon_i \]

where is the regression on the marker covariate

Estimation of \( \beta_j \) via multiple regression is not possible when the number of markers is greater than the number of phenotypes, the so-called large \( p \), small \( n \). In this case, not enough degree of freedoms exists for estimating all predictor effects simultaneously. Thus, the high degree of colinearity among markers can produce an over-fitted model. To confront this problem many statistical models and approaches have been proposed. These include multiple regression on marker genotypes [60], best linear unbiased prediction (BLUP) including effects of a single-marker locus [40], ridge regression [90], Bayesian procedures [19, 31, 48, 93], and semiparametric specifications [29, 32].

These methods differ with respect to assumptions about the marker effects [17], may perform differently for different phenotypes [44, 48, 86] and results may diverge because of differences in genetic architecture among traits [39, 47]. Implementing GS poses several statistical and computational challenges, such as how models can cope with the curse of dimensionality, colinearity between
markers, or the complexity of quantitative traits.

In general, there is not a universal model specification in GS. Different aspects can divide methods in GS:

- If genetic values are approximated by using linear regression procedures parametric [48] or using semi-parametric methods [32].
- Type of shrinkage estimation procedure used.
- Estimation based on liner or marker effects (Figure 1.6)

![Figure 1.6: Statistical GS methods based on the subject of estimation.](image)

Here, we classify the GS models based on estimation procedures performing shrinkage model, variable selection models, kernel methods and dimension reduction methods.
Shrinkage and variable selection methods

1. **Penalized methods**

To avoid over-fitting, a penalty term is introduced in the error function

\[
E = \sum_{i=1}^{N} (y_i - \sum_{j=0}^{M} x_{ij} \beta_j)^2 + \lambda \sum_{j=0}^{M} |\beta_j|^q
\]

where \( \lambda \) is a parameter that controls the effects of the penalty term. Setting \( q = 1, q = 2 \) and \( 0 \leq q \leq 1 \) are known as LASSO (least absolute shrinkage and selection operator), ridge regression (RR) and elastic net, respectively. Whittaker [90] proposed RR [54], also known as random regression best linear unbiased prediction (RRBLUP) for MAS. In this model, rather than categorizing markers whether or not are significant, RR shrinks all marker effects toward a fixed effect point (e.g., 0), and this may increase bias but reduces the variance of the estimator. This method makes the assumption that markers are random effects with common variance. Equal variance does not assume that all markers have the same effect but that marker effects are equally shrunken toward zero. LASSO combines variable selection and shrinkage of estimates. However, LASSO and other subset selection approaches have two main drawbacks. First, subset selection admits at most \( n \) nonzero estimates of regression coefficients, and this limitation is unreasonable in GS. Second, when predictors are correlated, methods performing variable selection are outperformed by RR [46]. Hence, in an effort to combine both methods (RR, LASSO), Zou and Hastie [96] proposed to use a weighted average and term the method the elastic net (EN).

2. **Bayesian shrinkage estimation**

The assumption made by RR-BLUP that genetic effects are evenly spread
across the genome with equal variance is not satisfactory in the genetic framework. Meuwissen et al. [48] proposed the used of Bayesian method analysis to improve genetic predictions. Bayesian methods can also be used for variable selection and shrinkage of estimates and is the prior density of marker effects, which defines whether the model will induce variable selection and shrinkage or shrinkage only. There is an extensive Bayesian alphabet [28] to use in predictions, which its main differences are the assumption of marker effects. In Bayes A, each marker effect is drawn from a normal distribution with its own variance. This allows each marker to be shrunken toward zero to a different degree. The variance parameters are in turn sampled from a scaled inverted $X^2$ distribution. This model is also known as Bayesian shrinkage regression [93]. In Bayes B, a probability is given that a marker has no effect at all. This model would better reflect the underlying genetic architecture if genetic variance was present at few loci and absent at many loci. Bayes B allows a combination of variable selection and shrinkage. Bayes C assumes common markers variances and allows for some markers to have no effect and gives more flexibility to model oligogenic than polygenic traits. The essential difference between many of these models is the prior distribution assigned to the marker coefficients [18, 28].

**Kernel methods**

Reproducing Kernel Hilbert Spaces (RKHS) [3] is the most common kernel used in GS. The RKHS approach first uses a kernel function to convert the marker dataset into a set of distances between pairs of observations that results in a
square matrix to be used in a linear model. Because RKHS regression does not assume linearity it might better capture non-additive effects.

**Dimension reduction methods**

When the number of dimensions becomes high, as in GS, the number of combinations of subsets taken from p variables is 2p. When p exceeds 20, calculation becomes difficult pointing terms of computation time. For these reason, methods for dimension reduction without loss of statistical information are important techniques for data analysis. The most dimensional reduction approach used in GS is principal component Analysis (PCA). Principal component analysis (PCA) is the best, in the mean-square error sense, linear dimension reduction technique. Being based on the covariance it is a second-order method matrix of the variables.

In essence, PCA seeks to reduce the dimension of the data by finding a few orthogonal linear combinations (the PCs) of the original variables with the largest variance. The first PC, $s_1$, is the linear combination with the largest variance. We have $s_1 = x^T w_1$, where the p dimensionale coefficient vector $w_1 = (w_1, 1 \ldots w_1, p)^T$ solves $w_1 = \argmax_{w=1} Var x^T w$. The second PC is the linear combination with the second largest variance and orthogonal to the first PC, and so on. There are as many PCs as the number of the original variables. For many datasets, the first several PCs explain most of the variance, so that the rest can be disregarded with minimal loss of information. Since the variance depends on the scale of the variables, it is customary to first standardize each variable to have mean zero and standard deviation one. After the standardization, the original variables with possibly different units of measurement are all
in comparable units.

**Statistical methods summary**

Empirical analyses have shown only small differences between methods [53, 69], with a slight advantage of models performing selection and shrinkage, such as Bayes B, for traits with large-effect QTL. But in general models such as Bayes A or Bayesian LASSO perform well across traits and GBLUP performs well for most traits. Habier [42] showed that RR-BLUP is more effective at capturing genetic relationships because it fits more markers into the prediction model. RR-BLUP implies that the trait is controlled by many loci with small effects. Nevertheless, Bayes B is more effective capturing LD between markers and QTL. Because markers are in tight linkage with the QTL, recombination does not cause them to decay rapidly and accuracies from Bayes B persist longer those from RR-BLUP. The unexpected generally good performance of GBLUP in real data is due to:

1. The real genetic architecture of traits appears less extreme than expected based on QTL mapping results.

2. Most of the gain in accuracy due to using markers in current applications arises from explaining the Mendelian sampling term, rather than from tracing signals generated at individual QTL.

Therefore, empirical evidence across populations has shown that in many cases these methods obtain very similar accuracies of the GEBVs (Moser et al. 2010). This suggests that many genes with a small effect controlled the additive genetic variation, somewhat like Fisher’s [26] infinitesimal model. Most of the
models assume additivity and might not perform well if non-additive genetic effects were important. In these situations, RKHS seems to perform well [53].

1.3.6 Relationship between TRS and TS

The theory of GS is based on the prediction of the genetic markers effects in LD with QTL. This denotes that a single SNP marker or a group of markers can be associated with QTL effects [38]. The overall implication is that predictions of BV may persist for several generations, allowing for:

1. Reduce number of phenotypic measurements in each generation [70].
2. Make accurate predictions across different populations provided sufficient marker density [35].

In case of linkage equilibrium, the accuracy of GEBVs is not necessarily zero but will approach the accuracy of pedigree-based BLUP-EBVs as the number of SNPs fitted in the model increases. The reason is that SNPs capture additive-genetic relationships irrespective of the amount of LD in the population as demonstrated by Habier [42] and Gianola [30]. Habier [42] proposed that genomic predictions also rely on the genetic relationships between individuals with phenotypic records to accurately predict genetic values, because genomic predictions are more accurate when predicted individuals are more closely related to a TRS. Therefore, a better understanding of what GS is predicting is important, especially because if the relationship hypothesis is true, then the predictive ability based on genomic data would persist only for one or two generations ahead, and continuous measurements of phenotypes of individuals that
are related to selection candidates would be needed [13].

Several studies have highlighted the importance of relatedness measures on genomic prediction accuracy [12, 13, 42, 44, 76, 81]. These studies indicating that the overall prediction of BV rely on the degree of relationship between the predicted individuals and those in the TRS, because the less related the predicted individuals were to those in the TRS, the lower the accuracy of prediction [12, 47]. This has important implications for breeding programs. If there are QTL with large effects, then accurate predictions may persist over generations, but long term predictions may not be as accurate when variation is controlled by a larger number of genes. Therefore, the larger the number of small genes controlling variation the more important it is that genotypes included in the reference population are genetically more related to selection candidates. In turn, Habier [44] and Pszczola [76] found that the individual reliability strongly depended on the average square relationship to the reference population. That is to say, the average relationship within the genotypes included in the TRS should be low and the relationship between the TRS and the TS should be maximized. These findings also imply that the optimal design of the TRS may differ from one breeding application to another and depend on the desired breeding strategy. In fact, Isidro [55] have showed that under population structure there is not a best criterion to optimize the TRS.

1.3.7 Final summary factors affecting GS

Population size, the relationships between TRS and TS, the marker density, and the statistical model are the factors affecting prediction accuracy that can be
controlled by the researcher. Nevertheless, trait heritability, genetic architecture, and to a large extent LD, cannot be controlled; Among the factors that are under control of the researcher, the size of the TRS and the strength of genetic relationships between training and validation samples are by far the most important factors affecting prediction accuracy. The model of choice is also important; however, the differences between models reported by simulation studies have not always been confirmed by real data analysis. Empirical analyses have shown only small differences between methods, with a slight advantage of models performing selection and shrinkage such as BayesB for traits with large-effect QTL [53] But in general thick-tailed models such as BayesA or Bayesian LASSO perform well across traits and G-BLUP performs well for most traits. An important reason is that, due to the fact that $p \approx n$, there are a multitude of different prediction equations that yield about the same likelihood and minimum prediction error rate [9].

1.4 Drawbacks Genomic Selection

- Implementation of GS is still an issue due to the fact that genotyping is still expensive.

- After selection phenotyping gives better accuracies than GS because phenotyping obtain new information on the entire genome every generation. However, GS selects every generation accurately for the same part of the genome.

- Understanding the biological and statistical mechanisms that drive the short-medium and long-term impact of genomic selection.
• There are no uniform benchmarks regarding the statistical methods used, the design of TRS schemes and the reporting of genomic prediction results.

• Genomic selection has yet to be demonstrated for most crop species. Where to apply GS in the breeding cycle (which generations) and how many lines to select for genotyping requires further research is essential in crop breeding.

• Genomic selection does not use functional annotation of markers. If so, GS could detect specific function genes affecting a trait, providing both biological insight and information about functional attributes of genes that may be influential across populations. That is to say, switch from the black box to white box.
CHAPTER 2
TRAINING SET OPTIMIZATION UNDER POPULATION STRUCTURE IN GENOMIC SELECTION

2.1 Abstract

Key message Population structure must be evaluated before optimization of the training set population. Maximizing the phenotypic variance captured by the training set is important for optimal performance. Abstract The optimization of the training set (TRS) in genomic selection has received much interest in both animal and plant breeding, because it is critical to the accuracy of the prediction models. In this study, five different TRS sampling algorithms, stratified sampling, mean of the coefficient of determination (CDmean), mean of predictor error variance (PEVmean), stratified CDmean (StratCDmean) and random sampling, were evaluated for prediction accuracy in the presence of different levels of population structure. In the presence of population structure, the most phenotypic variation captured by a sampling method in the TRS is desirable. The wheat dataset showed mild population structure, and CDmean and stratified CDmean methods showed the highest accuracies for all the traits except for test weight and heading date. The rice dataset had strong population structure and the approach based on stratified sampling showed the highest accuracies for all traits. In general, CDmean minimized the relationship between genotypes in the TRS, maximizing the relationship between TRS and the test set. This makes it suitable as an optimization criterion for long-term selection. Our results indicated that the best selection criterion used to optimize the TRS seems to depend on the interaction of trait architecture and population structure.
2.2 Introduction

Genomic selection (GS) emerged from the need to improve prediction of complex traits based on marker information [48]. The objective of GS is to improve the precision of selection by generating a genomic-estimated breeding value (GEBV) for selection candidates by simultaneously using genome-wide molecular marker information. Genomic selection uses a training population set (TRS) of individuals that have been both genotyped and phenotyped to train a model that takes genotypic information from a candidate population of untested individuals and produces GEBVs for selection [48]. Genomic selection modeling takes advantage of the increasing abundance of molecular markers through modeling of many genetic loci with small effects [43, 72, 83, 90, 93, 95].

Over the last decade, simulation and empirical cross-validation studies in plants have shown GS to be more effective than strategies that use only a subset of markers with significant effects [7, 16, 37, 49, 51, 56, 62, 65]. Genomic Selection is superior to phenotype based estimates for increasing gains per unit time even if both models show the same efficiency, because in principle, there is no need to record phenotypes of the candidates for the selection, hence shortening the length of the breeding cycle [50].

The most commonly used methods to estimate GEBVs are (1) best linear unbiased prediction from mixed model analysis using a genomic-estimated relationship matrix (GBLUP) [42] [95] and (2) random regression-best linear unbiased predictions (RR-BLUP) [48, 90]. Genomic best linear unbiased prediction is a method that utilizes a genomic relationship matrix and potentially pedigree information to estimate the genetic merit of an individual. Elements of
the genomic relationship matrix are estimated based on the proportion of the genome that two individuals share and predictions may be more accurate than those based on pedigree alone. For RR-BLUP, marker effects in the calibration set (CS) are estimated and then the GEBVs of the selection candidates are calculated by multiplying their marker scores by these estimates. Nevertheless, Habier [42] showed that both methods are equivalent.

The prediction accuracy of the GEBVs is normally evaluated using the correlation between the GEBVs and the true breeding values (TBV), $r(GEBV, TBV)$. This correlation provides an estimate of selection accuracy and is directly related to selection response [24], where $R = irσ_A$, $i = $ selection intensity, $r = $ accuracy, and $σ_A = $ the square root of the additive genetic variance [24]. Response to selection is important for determining gain per unit time and cost and for comparing breeding strategies. While new studies demonstrate that GS has great potential to increase rates of genetic gain, parameters determine its effectiveness for any specific breeding population. Factors that affect prediction accuracy include the number of markers used for estimating the GEBVs [82], trait heritability [51] calibration population size ([56], statistical models [53], number and type of molecular markers [11, 72], linkage disequilibrium [42], effective population size [14], relationship between calibration and test set (TS) [1, 12, 13, 76] and population structure [20, 41, 80, 81, 91]

In this study, we focus on the impact of population structure on GS accuracy. As a consequence of having different population genetic histories, distinct subpopulations could have differences in allele frequencies for many polymorphisms throughout the genome. If the populations have different overall values for the phenotype, any polymorphisms that differ in frequency between the two
populations will be associated with the phenotype even though they are not casual or in strong linkage disequilibrium with casual polymorphisms [64, 73, 74].

Population structure is a key factor affecting predictions of breeding values with genomic models and could result in biased accuracies of genomic predictions [77, 81, 92]. Accordingly, population structure needs to be taken into account because it could lead to unrealistic assessments of accuracy [77, 91] and preferential selection of individuals within a single subpopulation, which would result in a loss of diversity in the breeding program.

Recently, the design of the TRS has attracted much interest in both animal and plant breeding, since it is critical to the accuracy of the prediction models. Knowing the predictability of a model is one of the key elements for a better allocation of resources in plant breeding, especially due to the high costs of phenotyping. Several studies have noted that the accuracy of genomic predictions is highly influenced by the population used to calibrate the model [1, 12, 13, 42, 45, 76]. Larger TRSs tend to increase accuracy but simulations suggest that, in some cases, small TRSs can be just as accurate [43]. Generally, larger TRSs are required for traits controlled by more genes with smaller effects [35]. From the mixed model framework, given the trait heritability, marker data, and a TRS, it is possible to derive a measure of the quality of prediction for a set of genotypes. Two of those measures are the prediction error variance (PEV) and the coefficient of determination (CD). Rincent [78] used those criteria in an optimization procedure to choose a TRS of a given size in a maize diversity panel.

In quantitative genetics the PEV is central to the calculation of accuracies of estimated breeding values [52], to the restricted maximum likelihood (REML)
algorithms for the estimation of variance components (Patterson and Thompson 1971), and to methods that restrict the variance of response to selection [67]. The trends in genetic variance over time can be explored using breeding values and PEV of Mendelian sampling deviations. Choosing a TRS by seeking to minimize the PEV, however, may (1) result in the sampling of close relatives since the PEV does not take into account the genetic variance within the TRS (2) lead to TRSs that diverge between traits of differing heritability. To mitigate the first problem, Rincent [78] used the CD [59] that maximizes the expected reliabilities of contrasts between each selection candidate and the population mean. The CD can be defined as the squared correlation between the true and the predicted contrast of genetic values. It is a function of the PEV and of the genetic variance. Rincent [78] proposed CDmean as a criterion to maximize the consistency of prediction for several CS sizes. This criterion gave higher predictions than random samples and the PEVmean, because CDmean took into account covariance among the TRS genotypes and avoided the selection of closely related individuals. When all the genotypes are independent, PEVmean and CDmean are equivalent [59].

The purpose of this study was to compare the performance of different optimization criteria, including one proposed by Rincent [78], in the presence of population structure and to evaluate how population structure interacts with these criteria in the choice of the TRS. During the different optimization methods, the genotypes for all the individuals in the CS are used, but the phenotypes were only required for individuals selected in the TRS at the model building stage. Finally, accuracies of the models were evaluated by calculating Pearson correlations between the predicted values and the observed phenotype values in the TS.
2.3 Material and Methods

2.3.1 Genetic dataset material

Wheat Dataset

A population of 1,127 soft winter wheat varieties and $F_5$-derived advanced breeding genotypes resulting from many different crosses in the Cornell University Wheat Breeding Program (Ithaca, NY) were analyzed in this study. Lines were genotyped with 38,893 genotyping-by-sequencing (GBS) markers (Table 2.1).

| Worksheet | Rice | Wheat |
|-----------|------|-------|
| Population size | 1,127 | 405 |
| Markers | 38,893 GBS | 36,901 SNPs |
| Subpopulation | 4 | 3 |
| Environments | 3 | 2 |
| Years | 6 | 2 |
| Trait | $h^2$ | Trait | $h^2$ |
| YLD | 0.79 | FP | 0.78 |
| TWT | 0.92 | FT | 0.85 |
| LODG | 0.78 | PH | 0.89 |
| HD | 0.94 | PC | 0.70 |
| HT | 0.95 | |

Table 2.1: Germplasm description summary and heritabilities values for each trait.

Information about the construction and elaboration of the GBS libraries can be found in [72] and the latest updates on the GBS approach for wheat can be found on the website http://www.wheatgenetics.org/research). In summary, the GBS libraries were constructed in 95-plex using the P384A adaptor set. Genomic DNA was co-digested with the restriction enzymes PstI (CTGCAG) and MspI (CCGG) and barcoded adapters were ligated to genotype samples. Samples
were pooled by plate into a single library and polymerase chain reaction amplified. Each library was sequenced on a single lane of Illumina HiSeq 2000 (Cornell Life Science Core Laboratory Center). Missing marker values were imputed using a multivariate normal (MVN)-expectation maximization (EM) algorithm [72]. The EM algorithm represents a general approach to calculating maximum likelihood estimates of unknown parameters when data are missing [21].

The EM imputation was designed for use with genotyping-by-sequencing (GBS) markers, which tend to be high density but have lots of missing data. Phenotypic data for five traits in the wheat dataset were analyzed: grain yield, test weight, lodging, heading date and plant height (Table 1). The experiments were carried out over 6 years from 2007 to 2012, with one location in 2007 and three locations per year from 2008 to 2012 near Ithaca, NY. Each location was arranged in an unreplicated augmented, row-column design [25] with six check varieties replicated ten times each. First, in a mixed effect model an analysis was used to calculate best linear unbiased estimates (BLUEs) of locations and year effects [68] and BLUPs for the genotypes (i.e., varieties or accessions) as random effects in ASRmel- R [33]. Subsequently, these BLUPS were used for model building and the calculation of the accuracies of the models.

**Rice Dataset**

The rice diversity panel consisted of 413 diverse accessions of inbred lines of rice (O. sativa) from 82 countries, including many landraces, representing all the major ricegrowing regions of the world. This panel was genotyped with a 44-K chip (44,100 SNPs) and after filtering a total of 36,901 SNP markers were retained for genetic analysis [2] (Table 1). Across the 12 chromosomes of rice, SNPs cover roughly 380 Mb of the genome at a density of about 1 SNP per 10 Kb.
Each line was evaluated for important agronomic traits over 2 years with two replicates from 2006 to 2007. From this dataset, four different traits were selected (florets per panicle, flowering time in Arkansas, plant height and protein content) and phenotypic means of each inbred line across years and replicates were used for analysis (Table 1). All of the data from this study are publicly available at http://www.ricediversity.org and more details can be found in Zhao et al. (2011) and their supplementary data.

2.3.2 Training set optimization methods

In this study, three different methods were developed to study the optimization of the TRS. Method 1 optimizes the TRS by stratified sampling, method 2 by CDmean, PEVmean and random sampling and method 3 combined previous methods to build the TRS. More details about the methods can be found in supplementary information S1, S2 and S3. Initially, the overall population was randomly divided into a calibration set (CS) and a test set (TS). Next, the CS was further divided into a training set population (TRS) and a remaining set (RS). Genotypes belonging to the TRS were used to create the prediction equation by a mixed model. The remaining genotypes in the RS were used to build the TRS in method 2 and method 3. The TS is the set of genotypes from the base population where predictions will be made, that is to say, where GEBVs are calculated to make selection. In our study, for all methods, the CS and the TS were randomly obtained from the overall population (Fig. 2.1, number 1). To ensure an accurate comparison among methods, the same CS and TS genotypes were used for each one of the TRS methodologies. In this study, we used datasets with information for all the phenotypes and genotypes. This allowed
us to evaluate the accuracy of different TRS optimization methods. Nevertheless, in a real scenario the phenotypes are only available when the TRS is selected after the optimization process. Consequently, when selecting the TRS, only marker information was used. From the CS a subset of genotypes will be selected for phenotyping, which will build the TRS. The model built based on the phenotypes and genotypes in the TRS will be used to estimate the GEBVs for the genotypes in the TS. Here, we imposed the same population structure between CS and TS to avoid a potential prediction accuracy deflation that could arise when the TS population is not similarly stratified [91].
2.3.3 Method 1-optimization based on stratified sampling by cluster

In this method, two random samples from the base population were taken to generate the Calibration set (CS) and the test set (TS). Then, a cluster analysis was run on the CS as follows: Genotypic markers were used to calculate the Euclidean distances between genotypes. Hierarchical clustering analysis using the Ward criterion (i.e., at each step the pair of clusters with minimum between cluster distance are merged, generating clusters that were more equal in size) was applied to the Euclidean distance matrix. Principal components analysis (PCA) on genotypic data was used to visualize the structure of our populations. For the Cornell wheat program population, we selected four distinct subpopulations, based on genetic relationship and breeder’s knowledge. For the rice dataset we selected three distinct subpopulations. When the cluster analysis is obtained, the TRS is created by selecting a number of genotypes from each cluster proportional to the size of the cluster. Consequently, clusters with more genotypes will have a larger representation in the TRS than smaller clusters. With this method, we selected 25, 50, 100, 200 and 300 wheat genotypes and 25, 50, 100, 150 and 175 rice genotypes for the TRS. This methodology was repeated 50 times, and each time CS and TS were saved to assure a legitimate comparison among methods. The same CS and TS generated here were used to build the CS and TS for methods 2 and 3. Stratified sampling in clusters assured a high degree of genetic variability in the TRS, since each subpopulation was represented proportionally to its size. The optimization framework is shown in Fig. 2.1 number 1.
2.3.4 Method 2-optimization criterion based on CDmean and PEVmean

The same CS and TS obtained in method 1 were used here to initiate the optimization. Firstly, a random sample of the target TRS size was obtained and the CDmean was calculated. Then, the optimization algorithm code provided by Rincent et al. was applied [78] to our datasets. At each iteration, the algorithm randomly exchanged one genotype between the TRS and the set of RS genotypes. CDmean and PEVmean were then calculated. If the criterion was improved, the genotype exchange was accepted and otherwise rejected. The TRS optimization sizes sampled were the same as method 1. For each panel, 50 repetitions of the algorithm were performed and 2,000 iterations were needed to reach a plateau in the CDmean or PEVmean. The optimization framework is shown in Fig. 2.1 number 2.

PEV and CD optimization

A detailed description of the prediction model and optimization criteria was provided by [59] and [78]. We highlight here the model details and the calculation of PEVmean and CDmean. The criteria are based on the use of GBLUP [87] [42] to calculate the GEBVs. GBLUP mixed model can be formulated as

$$ y = X\beta + Z\mathbf{u} + \epsilon $$

where \(y\) is a vector of phenotypes, \(\beta\) is a vector of fixed effects (population mean in our case), \(\mathbf{u}\) is a vector of random genetic values, \(\epsilon\) is the vector of random residual, \(X\) and \(Z\) are design matrices.

The variance of the random effects \(\mathbf{u}\) is \(\text{var}(\mathbf{u}) = \text{G}\sigma_g^2\), where \(G\) is the genomic
relationship matrix and $\sigma^2_g$ is the additive genetic variance in the panel. The variance of the residuals is $\text{var}(\epsilon) = I \sigma^2_e$, where $I$ is the identity matrix.

Criteria of optimization

The prediction error variance of $u$ can be derived from the Henderson equation:

$$
\begin{pmatrix}
XX' & X'Z \\
Z'X & Z'Z + \lambda G^{-1}
\end{pmatrix} \begin{pmatrix}
\hat{\beta} \\
\hat{u}
\end{pmatrix} = \begin{pmatrix}
X'y \\
Z'y
\end{pmatrix}
$$

where $\lambda = \sigma^2_e / \sigma^2_g$ is the ratio between the residual and the additive variances and $G$ is the genomic relationship matrix. Using the notation

$$
\begin{bmatrix}
XX' & X'Z \\
Z'X & Z'Z + \lambda G^{-1}
\end{bmatrix} = \begin{bmatrix}
C_{11} & C_{12} \\
C_{21} & C_{22}
\end{bmatrix}
$$

$$
\text{var}(u | \hat{u}) = \text{var}(\hat{u} | u) = (Z'MZ + \lambda G^{-1})^{-1} \times \sigma^2_e
$$

where $M$ is a projector, orthogonal to the vector subspace spanned by $X$ columns ($MX=0$), $M = I - X(X'X)^{-1}X'$ where $(X'X)^{-1}$ is a generalized inverse of $X'X$ [59] and therefore

$$
\text{PEV}(\hat{u}) = \text{var}(u | \hat{u}) = \text{diag}C_{22} \times \sigma^2_e
$$

Contrasts allow us to compare the precision of comparisons between genotypes. The contrast will perform the comparison between genotype $i$ and $j$, therefore for any contrast $c$ of the predicted performances PEV can be calculated as:

$$
\text{PEV} = \text{diag} \left[ \frac{c'(Z'MZ + \lambda G^{-1})^{-1}c}{c'c} \right] \times \sigma^2_e
$$
where \( c \) is a vector of a particular linear combination whose elements sum to 0.

The aim in statistics is to minimize the error. Therefore, minimizing the mean of the PEVs of the contrast between each RS genotype and the mean of the CS panel is the goal of the optimization with PEV. Lalo [59] defined CD as the squared correlation between the true and the predicted contrast of genetic values. The CD can be expressed as

\[
CD = R^2 = \frac{\text{TSS-RSS}}{\text{TSS}} = \frac{\text{var}(u) - \text{var}(u | \hat{u})}{\text{var}(u)}
\]

where TSS is the total sum of squares, RSS is the residual of sum of squares, \( c \) is the contrast between genotypes, \( \text{var}(u) \) is the total genetic variance and \( \text{var}(u | \hat{u}) \) is the residual error variance or PEV. Making the corresponding substitution and calling

\[
(ZM'Z + \lambda G)^{-1} = \theta
\]

\[
CD(c) = \frac{\sigma^2_c c'Gc - \sigma^2_e c'\theta c}{\sigma^2_c c'Gc} = 1 - \frac{\sigma^2_e c'\theta c}{\sigma^2_c c'Gc} = 1 - \frac{\lambda c'\theta c}{c'Gc} = \frac{c'(G - \lambda \theta)c}{c'Gc}
\]

and taking the diagonal elements of this matrix the CD can be expressed as

\[
CD = \text{diag} \left[ \frac{c'(G - \lambda(Z'MZ + \lambda G^{-1})c}{c'Gc} \right]
\]

The CD corresponds to the expected reliability of the contrast between the predicted value of a given individual of the RS population and the population mean. The CD always lies within the unit interval. In this case, the optimization criteria will maximize the mean of the CD of the contrast between each non-phenotyped genotype (of the RS set) and the mean of the population [78].
The relationship matrix used for the calculation of PEVmean and CDmean was the genomic relationship matrix (G). The relationship matrix is estimated as $G = \frac{WW'}{f}$ where $W_{ik} = X_{ik} - 2p_k$ is the the mean centered marker $k$ for individual $i$, $p_k$ is the frequency of the 1 allele at marker $k$ for the entire population, and $X_{ik}$ denotes the number of minor alleles for the $i$th individual at marker $k$. Using a normalization constant of $f = 2\sum p_k(1 - p_k)$, the mean of the diagonal elements is $1 + f$ [23].

### 2.3.5 Method 3-optimization criterion based on stratified sampling CDmean by cluster

The goal in this approach is to combine the strengths of methods 1 and 2. In this method, after the cluster analysis, the algorithm will create the TRS based on CDmean applied within each cluster. That is, rather than random stratified sampling, TRS members are selected within each cluster by the CDmean method. The same conditions on TRS size, number of iterations and repetitions were applied in this method as described in previous methods. The optimization framework is shown in Fig. 2.1 number 3 and supplementary information S3.

### 2.3.6 Heritability calculation and statistical software

Trait heritability was estimated across $e$ environments and $r$ replicates using a mixed model where environment was treated as a fixed effect and genotypes and genotype x environment interaction as random effects.
\[ h^2 = \frac{\sigma^2_g}{\sigma^2_g + \frac{\sigma^2_{ge}}{e} + \frac{\sigma^2_e}{er}} \]

where \( \sigma^2_g \), \( \sigma^2_{ge} \), \( \sigma^2_e \) are the additive, genotype by environment and residual variance components, \( e \) is number of environments and \( r \) is the number of replicates per environment. All analyses were performed using R version 3.0 (2013). The package rrBLUP version 4.2 [22] (http://cran.r-project.org/web/packages/rrBLUP/) was used to calculate GEBVs. We assessed the predictive ability of the models by the Pearson correlation coefficients between the GEBVs and the observed phenotypes in the TS (referred to here as accuracy). The training population set was also obtained by random sampling from the CS.

\section*{2.4 Results}

\subsection*{Population structure}

We performed PCA to summarize the genetic variation in both datasets. The analyses revealed structure in both populations (Fig. 2.2).

\subsection*{Wheat}

Cluster analysis revealed that all of the clusters can be separated in the first two PC axes that accounted for 12.7 and 8.3\% of the genetic variance, respectively (Fig. 2.2a). The number of lines per cluster ranged from 107 to 516 (Table 2). The largest subpopulation size (516) corresponds to ancestral varieties from New York, Ontario, Ohio and Michigan, followed by genotypes derived from
Hurus/Houser/SuMei crosses. The third was formed by Elite Eastern soft winter from the eastern United States. Finally, genotypes from Geneva/Cayuga crosses (New York) formed the forth cluster. The structure explained in yield, test weight and height was 6.1, 5.7 and 8.2% respectively. Lodging and test weight showed the highest proportion of variance explained by the clusters with 15.4 and 13.0%, respectively.

| Cluster | Number of lines | Origins    | Representative line          |
|---------|-----------------|------------|-------------------------------|
| Wheat   |                 |            |                               |
| C1      | 516             | NY,Ont,OH,MI | Ancestral varieties         |
| C2      | 350             | NY,Ont,China | Harus/Houser/SuMei          |
| C3      | 154             | NY,MI,OH,IN,VA | Eliste Eastern           |
| C4      | 107             | NY          | Geneva/Caledonia           |
| Rice    |                 |            |                               |
| C1      | 145             | IN,CH,PH,BR | Indica/Aus                  |
| C2      | 127             | US,BR,AR,CO,NI | Tropial Japonica       |
| C3      | 99              | UE,JA,CH    | Temperate Japonica          |

Table 2.2: Descriptions of wheat and rice clusters identified using hierarchical clustering model analysis. NY New York, Ont Ontario, OH Ohio, MI Michigan, IN Indiana, VA Virginia, IN Indica, CH China, PH Philippines, BR Brasil, US United Stated, AR Argentina, CO Congo, NI Nigeria, EU European Union, JA Japan.

Rice

The rice dataset is a very diverse panel from 82 countries and the analysis of population structure revealed three clear subpopulations. Clusters were separated in the first two PCs axes and accounted for 39.2 and 8.3% of the total variance (Fig. 2b). Population sizes within clusters varied from 99 to 145 genotypes. A more detailed description of the accessions and geographical distribution of the rice germplasm can be found in Zhao [94].
Training set prediction accuracies

Figures 2.3 and 2.4 show the accuracies of the predictions for the wheat and rice datasets. In general, accuracy values were lower in wheat. Accuracies ranged from 0.12 to 0.59 and from 0.20 to 0.72 in wheat and rice, respectively.

In both populations, accuracies increased as the TRS size increased. Different heritability values and $\lambda$ did not change the patterns of accuracy for either dataset. Nevertheless, there were noteworthy differences in GS accuracies among TRS selection methods of optimization studied here. In the wheat dataset, predictions using the CDmean and StratCDmean methods showed the highest accuracies for all the traits except for test weight and heading date. In general, CDmean and StratCDmean were not significantly different within traits, although some exceptions were found for the smaller TRS sizes (Fig. 2.3 a, d). At the lowest TRS size, CDmean and StratCDmean showed the high-
The calibration sets were defined by maximizing CDmean; minimizing PEVmean; maximizing CDmean within cluster; stratified proportional sampling and random sampling. Four different population sizes (25, 50, 100, 200 and 300) were used for the optimization algorithm in five different traits (a yield, b test weight, c lodging, d heading date, e plant height). Standard error is indicated for each point over the 50 runs. Optimization of CDmean, PEVmean and StratCDmean was made with the heritability measured for each trait in each germplasm.

Predictions using the rice dataset showed higher accuracies than the wheat dataset overall even though the CS size (250) was smaller than for wheat (627) (Fig. 2.4). The stratified sampling method showed the highest accuracies for all traits. In this dataset, the calibration set of random sampling was always lower or equivalent to those obtained by stratified sampling for all traits.
the smallest population size, CDmean and StratCDmean showed the highest reliabilities but were not significantly different from stratified and random sampling. As the population size increased, their accuracies dropped below the stratified sampling approach, especially for plant height and protein content (Fig. 2.4 c, d). Similar to the wheat population, PEVmean accuracies followed a pattern similar to CDmean and the differences between accuracies of CDmean and PEVmean were significant only for florets per panicle and flowering time at intermediate population size. For these traits, PEVmean showed the lowest accuracies (Fig. 2.4 a, b).

**Selection optimization of the training sets**

For both populations, Fig. 2.5 shows for the TRS size of 25, the PCA axes for the genotypes selected by the algorithms based on CDmean, PEVmean and stratCDmean methods. This figure illustrates the functional role of the algorithm in selecting the best genotypes to generate the optimized TRS as well as the variability of the panel captured by the TRS. In both populations, CDmean frequently selected most of the genotypes from the center of the PCs, and only rarely selected genotypes from the extremes of the clusters. This feature was observed more clearly for wheat than for rice (Figs. 2.3, 2.4, 2.5a, d). These patterns were stable across runs and traits. For the wheat dataset, most of the TRS genotypes selected using the PEVmean method were from the Elite Eastern cluster, with few genotypes from the center of the PCs (Fig. 2.5b). This pattern was also observed in the rice population, where PEVmean did not select genotypes from the Temperate Japonica cluster and more frequently selected genotypes from the Indica/Aus cluster (Fig. 2.5e). Although StratCDmean selected genotypes
more disperse within clusters than other sampling algorithm, this was not reflected in an increase of the accuracies (Fig. 2.5c, f). Although, the algorithm forced CDmean to pick genotypes within clusters, most of the genotypes that were repeatedly selected tended to be from the center of the PCs in both wheat and rice populations.
Figure 2.5: Genotypes selected from the optimization algorithm over the 50 run are plotted on the principal components analysis in wheat and rice germplasm. The genotypes were selected based on CDmean (a, d), PEVmean (b, e) and StratCDmean (e, f). Green dots represent the genotypes selected by the algorithm over the 50 runs. Red dots indicate those genotypes that were selected more than 15 and 27 times in wheat and rice germplasm, respectively.

Relative phenotypic variance and accuracy.

Because of the different behavior of the test weight and heading date traits in wheat, we conducted additional analyses to determine the relationship between the phenotypic variance and accuracy. The ratio of the phenotypic variance of the genotypes most selected by CDmean and the total variance was plotted against the relative accuracy between CDmean and random sampling methods in a TRS size of 40 genotypes. If the value of the ratio between CDmean and the total variance is greater than 1.0, it means that extreme phenotypes are overrepresented in the TRS, while close-to average phenotypes are underrepresented. There was a positive overall relationship between the phenotypic variance cap-
tured by the TRS, and the relative accuracy of CDmean versus the random sampling method. Within the dataset the same relationship was observed more clearly in the wheat population than in rice. Figure 2.6 shows that CDmean only performed well when the ratio of the phenotypic variance of the TRS and the total phenotypic variance was greater than or equal to two. Yield, lodging and plant height were the only traits in the wheat germplasm, where CDmean performed better than random sampling. In the wheat dataset, CDmean did not perform well for test weight and heading date (Fig. 2.3 c,d). For these traits, the accuracies were the lowest and CDmean did not capture a larger phenotypic variance. In the rice dataset, CDmean did not perform better than random sampling (Fig. 2.4). Here, as observed for wheat, rice traits were grouped together and showed the same positive relationship between the relative phenotypic variance and accuracy.

2.5 Discussion

In a scenario where we have a diverse panel of genotypes that have been genotyped but not phenotyped, the first question that arises is how to select the best genotypes to create a TRS to build our statistical model for making predictions in the TS. The goal is to select the minimum number of genotypes that assure an optimal accuracy on the TS population. Several studies [12, 13, 63, 76, 81] and more recently by Guo [41] have highlighted the criteria to build an optimal TRS.

Rincent [78] developed algorithms to select an improved TRS that strategically sampled the genotypic space when developing training sets for genomic prediction. In this paper, our aim was to compare the performance of five algo-
Figure 2.6: Relative performance of CDmean versus the random sampling method as a function of the ratio of the phenotypic variance between the top 40 most selected genotypes by CDmean and the overall phenotypic variance in wheat (red-cross) and rice (blue-triangle). Traits per germplasm are indicated in wheat as YLD yield, TWT test weight, LODG lodging, HD heading date, HT plant height and in rice as FLORETS florets per panicle, FLOWERING flowering time, PH plant height, PROTEIN protein content.

Algorithms, including the procedures from Rincent [78], in the presence of population structure using three different TRS optimization selection methods. These methods were tested on two different germplasm panels with different origins, different population structure effects, and in nine different traits with heritabilities ranging from 0.70 to 0.95 (Table 2.2). Our results indicated that the best selection criterion used to optimize the TRS was not consistent among populations. This seems to indicate that the interaction of trait architecture and population structure plays an important role in the optimization of the TRS. In six out of the nine traits studied in this analysis, the stratified sampling method showed higher accuracies than CDmean, PEVmean and StratCDmean, indicating that the degree of population structure is important in the design of the TRS.
In populations with strong structure, as observed for the rice population, stratified sampling performed better than other methods (Fig. 2.4). In contrast, CD-mean and StratCDmean showed better accuracies where population structure effects were mild, as observed for the wheat germplasm (Fig. 2.3). The similarity of accuracies for CDmean and StratCDmean can be explained by the fact that the contrasts used for both approaches to calculate the CDmean statistics were the same.

The divergence in selection method performance for test weight and heading date traits in wheat was unexpected and additional analyses were performed to explain the result. We found that these different results for test weight and heading date could be explained by the total phenotypic variance sampled for the trait (Fig. 2.6). For those traits, stratified sampling showed higher accuracies than CDmean, PEVmean and StratCDmean methods but was not different from random sampling.

Our results in Fig. 2.6 indicated that, when CDmean captures most of the phenotypic variance the accuracies increased, as indicated for the traits yield, lodging and plant height. In contrast to this observation, the large genotypic variance obtained by CDmean does not always translate into a higher phenotypic variance ratio in the TRS. This might explain why CDmean performed poorly for test weight and heading date, because on average it produced TRS with reduced phenotypic variance compared to a random sampling. In addition, the lower phenotypic variance for test weight and heading date could be due to fewer genes affecting these traits in comparison to the other traits. These results seem to indicate that the best strategy may be to maximize the phenotypic variance captured by the TRS. In fact, recent studies [8, 57] have shown
that strategies that maximize the phenotypic variance, through picking individuals from the two-tail distribution, are preferable to using genotypes with the largest or lowest phenotypic deviation. Empirical studies are needed to endorse the simulation results. Capturing most of the phenotypic variance in the training set seems to be key for optimal performance. CDmean showed higher accuracies than PEVmean among traits and populations, with the only exception being for intermediate TRS set sizes for test weight for wheat (Fig. 2.3d).

The optimal design for a TRS population for use in genomic prediction should minimize the relationship among genotypes in the TRS and maximize the relationship of the TS genotypes to the TRS. Consequently, the genotypes belonging to the TRS should not be closely related to each other but should be representative of the entire population. This is the main benefit to using the CDmean, because it takes into account the covariance among the candidate genotypes preventing the selection of closely related genotypes [59, 78]. The CDmean algorithm most frequently selected genotypes situated near the center of the PCA under the effect of population structure, indicating that CDmean minimized the genetic distance to each cluster resulting in optimal performance when there was mild population structure. In contrast to the results found by Rincent [78], the CDmean method did not include all the extreme genotypes from each cluster. For example, in the wheat population the most frequent genotypes selected by CDmean belonged to the Ancestral Varieties (Fig. 2.5a). The StratCDmean method was chosen to force the CDmean algorithm to select more extreme genotypes from different clusters. Although, StratCDmean improved the sampling of the extremes of the genotypes in different clusters, our results indicated that this strategy did not improve the accuracies of the predictions in either the rice or the wheat datasets. This could be due to the fact that
the contrasts used in the CDmean and StratCDmean were the same. CDmean and StratCDmean gave the highest accuracies for the smallest TRS size in both populations (Figs. 2.4, 2.5). This indicated that under the effect of population structure, CDmean and StratCDmean will perform better, on average, than the other methods, and therefore would be favored among these methods when the size of the TRS is small.

As observed by Rincent [78], the performance of the PEVmean in both populations revealed patterns similar to CDmean. One pitfall to using PEVmean to optimize the TRS is that, in contrast to CDmean, PEVmean selected a high number of related genotypes to create the TRS, which was not optimal (Fig. 2.5b, e). While accuracies between PEVmean and CDmean were not very different among traits and germplasm, the fact that it included more closely related genotypes would limit long-term gains from selection needed in plant breeding schemes [56]. Nevertheless, PEV is still an appropriate selection criterion for a measure of connectedness [58]. As shown in Fig.5b, genotypes selected by PEVmean did not cover a wide genotypic space from the relationship matrix, but it selected a larger sample of Elite wheat genotypes.

The efficiency of the methods in terms of computational time also plays an important role in choosing a method for optimization. From the three methods used here, stratified sampling was the most efficient (less than a day), followed by StratCDmean (2 days), and CDmean (4 days). The fact that StratCDmean did not show large differences in accuracies in comparison with CDmean, and also improved the speed of the algorithm, made it more suitable than CDmean in the presence of population structure. It is also important to note that the size of the CS can limit the use of CDmean. The algorithm requires the inversion
of large matrices at each iteration to optimize the TRS, making it computationally intensive for large population CS sizes. For example in our study, the time to find the optimum took 50% more time using the wheat dataset compared to rice. In addition, for stratified sampling and StratCDmean methods to be effective, a sufficient number of genotypes per cluster is required for the sampling algorithm. When the number of genotypes per cluster is too small, the stratified sampling is less useful.

Training set design for GS has attracted much attention in both animal and plant breeding in recent years because it is critical to the accuracy of the prediction models. However, less consideration has been given to the test population in the optimization process. We believe that the use of information from the test set could be valuable to improve accuracies of prediction models for TRS design. In this sense, an alternative to the maximization of the CDmean in the TRS could be the minimizing the PEVmean in the test set. Thus, the information about the test dataset could be used, while building the prediction model, by selecting the genotypes for the TRS that minimize the PEV of the test set. In our optimization criteria, as well as in Rincent [78], the information from performance of relatives was incorporated through the use of a relationship matrix to calculate GEBVs. This is appropriate if major genes are not involved in the trait of interest. If the genetic distance based on genome-wide markers does not reflect the variability of the trait because major genes are involved, markers are not expected to be efficient for guiding the sampling of the TRS. If the optimal calibration set depends on the trait considered, this might be a problem for the implementation of GS in breeding programs because selection objectives usually involve multiple traits. Instead of using genomic prediction models for traits with major genes, it might be better to use models that include large effect loci as fixed effects in
GS models. Studies have shown that including large effect loci in GS models can improve significantly the prediction accuracies [5, 28, 53, 79]. The information about the trait architecture learned from these models could be used in the future for developing new criteria for optimization. In addition, it should be mentioned that our results come from an additive genetic model and it might be worthwhile to explore the use of other models that can capture genetic effects such as epistasis and genotype-by-environment interaction.

In this study, we only measure the effect of population structure on the optimization of the TRS, however some of the variation observed in our results could be due to other unmeasured features, because accuracies from prediction models depend on a complex network of different, interrelated factors. We showed that population structure played an important role in the optimization of the TRS. When population structure effects are minor, CDmean performed better than other selection methods and captured most of the genetic variability for most traits in the TRS. This makes it suitable as an optimization criterion for long-term selection. However, under strong population structure stratified sampling performed better than CDmean, indicating that population structure must be evaluated before optimization to be sure the algorithm used does not reduce the phenotypic variation. Our results indicate that the overall optimization method works best when the trait under study is polygenic, because the genome-wide relationship measured by the G matrix captures the phenotypic relationship adequately. If the underlying genetic control of the trait is not polygenic, then the success of the training optimization techniques will similarly depend on whether or not the alleles of the trait are aligned with the overall structure. Stratified sampling is expected to perform best if the alleles controlling the traits are distributed according to the structure.
2.6 Acknowledgements

The Regional Ministry of Economy, Innovation and Science of Andalusia, Spain provided financial support for Julio Isidro Sánchez. The authors would like to thank researchers and institutions that contributed to the development of the rice diversity panel. In addition, the authors would like to express gratitude to Dr. R. Rincent who gently provided us the script model to optimize the training set based on CDmean. We also thank to all collaborators involved in conducting the trials. This research was supported in part by the National Research Initiative Competitive Grant 2011-68002-30029 (Triticeae-CAP) from the USDA National Institute of Food and Agriculture and by Hatch project 149-449.
[1] Theresa Albrecht, Valentin Wimmer, Hans-Jürgen Auinger, Malena Erbe, Carsten Knaak, Milena Ouzunova, Henner Simianer, and Chris-Carolin Schön. Genome-based prediction of testcross values in maize. *Theoretical and Applied Genetics*, 123(2):339–350, 2011.

[2] Jetty SS Ammiraju, Meizhong Luo, José L Goicoechea, Wenming Wang, Dave Kudrna, Christopher Mueller, Jayson Talag, HyeRan Kim, Nicholas B Sisneros, Barbara Blackmon, et al. The oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert bac libraries that represent the 10 genome types of the genus oryza. *Genome Research*, 16(1):140–147, 2006.

[3] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, pages 337–404, 1950.

[4] Franco G Asoro, Mark A Newell, William D Beavis, M Paul Scott, and Jean-Luc Jannink. Accuracy and training population design for genomic selection on quantitative traits in elite north american oats. *The Plant Genome*, 4(2):132–144, 2011.

[5] Rex Bernardo. Genomewide selection when major genes are known. *Crop Science*, 54(1):68–75, 2014.

[6] Rex Bernardo and Alain Charcosset. Usefulness of gene information in marker-assisted recurrent selection: a simulation appraisal. *Crop Science*, 46(2):614–621, 2006.

[7] Rex Bernardo and Jianming Yu. Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47(3):1082–1090, 2007.

[8] AA Boligon, N Long, LG Albuquerque, KA Weigel, D Gianola, and GJM Rosa. Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection. *Journal of animal science*, 90(13):4716–4722, 2012.

[9] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[10] Alain Charcosset et al. Marker-assisted introgression of quantitative trait loci. *Genetics*, 147(3):1469–1485, 1997.
[11] X Chen and PF Sullivan. Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *The pharmacogenomics journal*, 3(2):77–96, 2003.

[12] Samuel A Clark, John M Hickey, Hans D Daetwyler, and Julius HJ van der Werf. 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), 2012.

[13] Samuel A Clark, John M Hickey, and Julius HJ Van der Werf. Different models of genetic variation and their effect on genomic evaluation. *Genet Sel Evol*, 43:18, 2011.

[14] Hans D Daetwyler, Beatriz Villanueva, and John A Woolliams. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, 3(10):e3395, 2008.

[15] HD Daetwyler, JM Hickey, JM Henshall, S Dominik, B Gredler, JHJ Van Der Werf, and BJ Hayes. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Animal Production Science*, 50(12):1004–1010, 2010.

[16] Gustavo de Los Campos, Daniel Gianola, and David B Allison. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, 11(12):880–886, 2010.

[17] Gustavo de los Campos, John M Hickey, Ricardo Pong-Wong, Hans D Daetwyler, and Mario PL Calus. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327–345, 2013.

[18] Gustavo de los Campos, Yann C Klimentidis, Ana I Vazquez, and David B Allison. Prediction of expected years of life using whole-genome markers. *PloS one*, 7(7):e40964, 2012.

[19] Gustavo De Los Campos, Hugo Naya, Daniel Gianola, José Crossa, Andrés Legarra, Eduardo Manfredi, Kent Weigel, and José Miguel Cotes. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375–385, 2009.

[20] APW De Roos, BJ Hayes, and ME Goddard. Reliability of genomic predictions across multiple populations. *Genetics*, 183(4):1545–1553, 2009.
[21] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[22] Jeffrey B Endelman. Ridge regression and other kernels for genomic selection with r package rrblup. *The Plant Genome*, 4(3):250–255, 2011.

[23] Jeffrey B Endelman and Jean-Luc Jannink. Shrinkage estimation of the realized relationship matrix. *G3: Genes— Genomes— Genetics*, 2(11):1405–1413, 2012.

[24] DS Falconer and TFC Mackay. Introduction to quantitative genetics. *Longman*, 19(8):1, 1996.

[25] Walter T Federer and D Raghavarao. On augmented designs. *Biometrics*, pages 29–35, 1975.

[26] Ronald A Fisher. Xv.the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52(02):399–433, 1919.

[27] Sherry A Flint-Garcia, Jeffry M Thornsberry, and Buckler IV. Structure of linkage disequilibrium in plants*. *Annual review of plant biology*, 54(1):357–374, 2003.

[28] Daniel Gianola. Priors in whole-genome regression: the bayesian alphabet returns. *Genetics*, 194(3):573–596, 2013.

[29] Daniel Gianola and Gustavo de Los Campos. Inferring genetic values for quantitative traits non-parametrically. *Genetics Research*, 90(06):525–540, 2008.

[30] Daniel Gianola, Gustavo de Los Campos, William G Hill, Eduardo Manfredi, and Rohan Fernando. Additive genetic variability and the bayesian alphabet. *Genetics*, 183(1):347–363, 2009.

[31] Daniel Gianola, Miguel Perez-Enciso, and Miguel A Toro. On marker-assisted prediction of genetic value: beyond the ridge. *Genetics*, 163(1):347–365, 2003.

[32] Daniel Gianola and Johannes BCHM van Kaam. Reproducing kernel
hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178(4):2289–2303, 2008.

[33] Arthur R Gilmour, Robin Thompson, and Brian R Cullis. Average information reml: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, pages 1440–1450, 1995.

[34] ME Goddard and BJ Hayes. Genomic selection. *Journal of Animal Breeding and Genetics*, 124(6):323–330, 2007.

[35] Michael E Goddard and Ben J Hayes. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10(6):381–391, 2009.

[36] Mike Goddard. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136(2):245–257, 2009.

[37] JM González-Camacho, G De Los Campos, P Pérez, D Gianola, JE Cairns, G Mahuku, R Babu, and J Crossa. Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, 125(4):759–771, 2012.

[38] Laura Grapes, Mehmet Z Firat, Jack CM Dekkers, Max F Rothschild, and Rohan Luigi Fernando. Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. *Genetics*, 172(3):1955–1965, 2006.

[39] Dario Grattapaglia and Marcos DV Resende. Genomic selection in forest tree breeding. *Tree Genetics & Genomes*, 7(2):241–255, 2011.

[40] RL Fernando M Grossman. unbiased prediction. 1989.

[41] Zhigang Guo, Dominic M Tucker, Christopher J Basten, Harish Gandhi, Elhan Ersoz, Baohong Guo, Zhanyou Xu, Daolong Wang, and Gilles Gay. The impact of population structure on genomic prediction in stratified populations. *Theoretical and Applied Genetics*, 127(3):749–762, 2014.

[42] D Habier, RL Fernando, and JCM Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007.
[43] David Habier, Rohan L Fernando, and Jack CM Dekkers. Genomic selection using low-density marker panels. *Genetics*, 182(1):343–353, 2009.

[44] David Habier, Rohan L Fernando, Kadir Kizilkaya, and Dorian J Garrick. Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics*, 12(1):186, 2011.

[45] David Habier, Jens Tetens, Franz-Reinhold Seefried, Peter Lichtner, and Georg Thaller. The impact of genetic relationship information on genomic breeding values in german holstein cattle. *Genetics Selection Evolution*, 42(1):5, 2010.

[46] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[47] Ben John Hayes, Peter M Visscher, and Michael E Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(01):47–60, 2009.

[48] BJ Hayes, ME Goddard, et al. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

[49] Elliot L Heffner, Jean-Luc Jannink, and Mark E Sorrells. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome*, 4(1):65–75, 2011.

[50] Elliot L Heffner, Aaron J Lorenz, Jean-Luc Jannink, and Mark E Sorrells. Plant breeding with genomic selection: gain per unit time and cost. *Crop science*, 50(5):1681–1690, 2010.

[51] Elliot L Heffner, Mark E Sorrells, and Jean-Luc Jannink. Genomic selection for crop improvement. *Crop Science*, 49(1):1–12, 2009.

[52] Charles R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975.

[53] Nicolas Heslot, Hsiao-Pei Yang, Mark E Sorrells, and Jean-Luc Jannink. Genomic selection in plant breeding: a comparison of models. *Crop Science*, 52(1):146–160, 2012.
[54] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[55] Julio Isidro, Jean-Luc Jannink, Deniz Akdemir, Jesse Poland, Nicolas Heslot, and Mark E. Sorrells. Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics*, 40:1–14, 2014.

[56] Jean-Luc Jannink, Aaron J Lorenz, and Hiroyoshi Iwata. Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics*, 9(2):166–177, 2010.

[57] JA Jiménez-Montero, O Gonzalez-Recio, and R Alenda. Genotyping strategies for genomic selection in small dairy cattle populations. *animal*, 6(08):1216–1224, 2012.

[58] BW Kennedy and D Trus. Considerations on genetic connectedness between management units under an animal model. *Journal of animal science*, 71(9):2341–2352, 1993.

[59] Denis Laloë. Precision and information in linear models of genetic evaluation. *Genetics Selection Evolution*, 25(6):557–576, 1993.

[60] Russell Lande and Robin Thompson. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743–756, 1990.

[61] Eric S Lander and David Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199, 1989.

[62] Robenzon E Lorenzana and Rex Bernardo. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, 120(1):151–161, 2009.

[63] Steven Maenhout, Bernard De Baets, and Geert Haesaert. Graph-based data selection for the construction of genomic prediction models. *Genetics*, 185(4):1463–1475, 2010.

[64] Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5):512–517, 2004.

[65] Jon M Massman, Andres Gordillo, Robenzon E Lorenzana, and Rex
Bernardo. Genomewide predictions from maize single-cross data. *Theoretical and Applied Genetics*, 126(1):13–22, 2013.

[66] The Meuwissen, Mike E Goddard, et al. Prediction of identity by descent probabilities from marker-haplotypes. *Genetics Selection Evolution*, 33(6):605–634, 2001.

[67] The Meuwissen and JA Woolliams. Maximizing genetic response in breeding schemes of dairy cattle with constraints on variance of response. *Journal of dairy science*, 77(7):1905–1916, 1994.

[68] J Möhring and H-P Piepho. Comparison of weighting in two-stage analysis of plant breeding trials. *Crop science*, 49(6):1977–1988, 2009.

[69] Gerhard Moser, Bruce Tier, Ron E Crump, Mehar S Khatkar, Herman W Raadsma, et al. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide snp markers. *Genet Sel Evol*, 41(1):56, 2009.

[70] WM Muir. Comparison of genomic and traditional blup-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics*, 124(6):342–355, 2007.

[71] Jesse A Poland, Patrick J Brown, Mark E Sorrells, and Jean-Luc Jan-nink. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS one*, 7(2):e32253, 2012.

[72] Jesse A Poland and Trevor W Rife. Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, 5(3):92–102, 2012.

[73] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[74] Jonathan K Pritchard and Peter Donnelly. Case–control studies of association in structured or admixed populations. *Theoretical population biology*, 60(3):227–237, 2001.

[75] JE Pryce, BJ Hayes, and ME Goddard. Novel strategies to minimize
progeny inbreeding while maximizing genetic gain using genomic information. *Journal of dairy science*, 95(1):377–388, 2012.

[76] M Pszczola, T Strabel, HA Mulder, and MPL Calus. Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of dairy science*, 95(1):389–400, 2012.

[77] Christian Riedelsheimer, Jeffrey B Endelman, Michael Stange, Mark E Sorrells, Jean-Luc Jannink, and Albrecht E Melchinger. Genomic predictability of interconnected biparental maize populations. *Genetics*, 194(2):493–503, 2013.

[78] Renaud Rincent, Denis Laloë, Stéphane Nicolas, Thomas Altmann, Dominique Brunel, Pedro Revilla, Victor M Rodriguez, J Moreno-Gonzalez, A Melchinger, Eva Bauer, et al. 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(2):715–728, 2012.

[79] Jessica E Rutkoski, Jesse A Poland, Ravi P Singh, Julio Huerta-Espino, Sridhar Bhavani, Hugues Barbier, Matthew N Rouse, Jean-Luc Jannink, Mark E Sorrells, J Huerta Espino, et al. Genomic selection for quantitative adult plant stem rust resistance in wheat.

[80] M Saatchi, SR Miraei-Ashtiani, A Nejati Javaremi, and H Mehrabani-Yeghaneh. The impact of information quantity and strength of relationship between training set and validation set on accuracy of genomic estimated breeding values. *African Journal of Biotechnology*, 9(4), 2010.

[81] Mahdi Saatchi, Mathew C McClure, Stephanie D McKay, Megan M Rolf, JaeWoo Kim, Jared E Decker, Tasia M Taxis, Richard H Chapple, Holly R Ramey, Sally L Northcutt, et al. Accuracies of genomic breeding values in american angus beef cattle using k-means clustering for cross-validation. *Genet Sel Evol*, 43:40, 2011.

[82] LR Schaeffer. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123(4):218–223, 2006.

[83] TR Solberg, AK Sonesson, JA Woolliams, et al. Genomic selection using different marker types and densities. *Journal of Animal Science*, 86(10):2447–2454, 2008.
[84] AK Sonesson et al. Maximising genetic gain whilst controlling rates of genomic inbreeding using genomic optimum contribution selection. In *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, 1*, number 892, 2010.

[85] Meuwissen The. Accuracy of breeding values of unrelated individuals predicted by dense snp genotyping. *Genet Sel Evol*, 41:35, 2009.

[86] M Graziano Usai, Mike E Goddard, and Ben J Hayes. Lasso with cross-validation for genomic selection. *Genetics research*, 91(06):427–436, 2009.

[87] PM VanRaden. Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423, 2008.

[88] Peter M Visscher, Stuart Macgregor, Beben Benyamin, Gu Zhu, Scott Gordon, Sarah Medland, William G Hill, Jouke-Jan Hottenga, Gonneke Willemsen, Dorret I Boomsma, et al. Genome partitioning of genetic variation for height from 11,214 sibling pairs. *The American Journal of Human Genetics*, 81(5):1104–1110, 2007.

[89] Michael N Weedon, Hana Lango, Cecilia M Lindgren, Chris Wallace, David M Evans, Massimo Mangino, Rachel M Freathy, John RB Perry, Suzanne Stevens, Alistair S Hall, et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics*, 40(5):575–583, 2008.

[90] John C Whittaker, Robin Thompson, and Mike C Denham. Marker-assisted selection using ridge regression. *Genetical research*, 75(02):249–252, 2000.

[91] Vanessa S Windhausen, Gary N Atlin, John M Hickey, Jose Crossa, Jean-Luc Jannink, Mark E Sorrells, Babu Raman, Jill E Cairns, Amsal Tarekegne, Kassa Semagn, et al. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3: Genes—Genomes—Genetics*, 2(11):1427–1436, 2012.

[92] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher. Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, 14(7):507–515, 2013.

[93] Shizhong Xu. Estimating polygenic effects using markers of the entire genome. *Genetics*, 163(2):789–801, 2003.
[94] Keyan Zhao, Chih-Wei Tung, Georgia C Eizenga, Mark H Wright, M Liakat Ali, Adam H Price, Gareth J Norton, M Rafiul Islam, Andy Reynolds, Jason Mezey, et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa. Nature communications, 2:467, 2011.

[95] Shengqiang Zhong, Jack CM Dekkers, Rohan L Fernando, and Jean-Luc Jannink. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics, 182(1):355–364, 2009.

[96] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.