INTRODUCING SPARSITY INTO SELECTION INDEX METHODOLOGY WITH APPLICATIONS TO HIGH-THROUGHPUT PHENOTYPING AND GENOMIC PREDICTION

By

Marco Antonio Lopez Cruz

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Breeding, Genetics and Biotechnology – Crop and Soil Sciences – Doctor of Philosophy

2020

ABSTRACT

INTRODUCING SPARSITY INTO SELECTION INDEX METHODOLOGY WITH APPLICATIONS TO HIGH-THROUGHPUT PHENOTYPING AND GENOMIC PREDICTION

By

Marco Antonio Lopez Cruz

Research in plant and animal breeding has been largely focused on the development of methods for a more efficient selection by altering the factors that affect genetic progress: selection intensity, selection accuracy, genetic variance, and length of the breeding cycle. Most of the breeding efforts have been primarily towards increasing selection accuracy and reducing the breeding cycle.

Genomic selection has been successfully adopted by many public and private breeding organizations. Over years, these institutions have developed and accumulated large volumes of genomic data linked to phenotypes from multiple populations and multiple generations. This data abundance offers the opportunity to revolutionize genetic research. However, these data sets are also increasingly heterogeneous, with many subpopulations and multiple generations represented in the data. This translates into potentially heterogeneous allele frequencies and different LD patterns, thus leading to SNP-effect heterogeneity.

Genomic selection methods were developed with reference to homogeneous populations in which SNP-effects are assumed constant across the whole population. These methods are not necessarily optimal for the contemporary available data sets for model training. Therefore, a first focus of this dissertation is on developing novel methods that can leverage the large-scale of modern data sets while coping with the heterogeneity and complexity of this type of data.

In recent years, there have also been important advances in high-throughput phenotyping (HTP) technologies that can generate large volumes of data at multiple time-points of a crop. Examples of this include hyper-spectral imaging technologies that can capture the reflectance of electromagnetic power by crops at potentially thousands of wavelengths. The integration of HTP in genetic evaluations represents a great opportunity to further advance plant breeding; however,

the high-dimensional nature of HTP data poses important challenges. Therefore, a second focus of this dissertation is on the development of a novel approach to efficiently incorporate HTP data for breeding values prediction.

Thus, this dissertation aims to contribute novel methods that can improve the accuracy of genomic prediction by optimizing the use of large, potentially heterogeneous, genomic data sets and by enabling the integration of HTP data. We present a novel statistical approach that combines the standard selection index methodology with variable-selection methods commonly used in machine learning and statistics, and developed software to implement the method. Our approach offers solutions to both genomic selection with potentially highly heterogeneous genomic data sets, and the integration of HTP in genetic evaluations.

Dedicated to my mother and to the memory of my father.

ACKNOWLEDGEMENTS

I want to express my special gratitude to Dr. Gustavo de los Campos for his mentoring during the curse of my doctorate, for all his valuable academic and personal advice, and for the constructive suggestions during the planning and development of this research.

I want to thank my graduate committee, Dr. Eric Olson, Dr. Gustavo de los Campos, Dr. David Douches, and Dr. Dechun Wang for accepting being part of my Ph.D. committee and for their advice and assistance.

I want to express my gratitude to Dr. Jose Crossa and Dr. Susanne Dreisigacker from the International Maize and Wheat Improvement Center (CIMMYT) to trust me and encourage me to pursue a Ph.D., and for their support in getting funding for my studies.

I offer my special acknowledgment to Dr. Paulino Perez for being my first contact with the field of Statistical Genetics and his support in the development of the software generated from this research. I wish to thank lab members of the QuantGen group and other friends I met at MSU for their support and for making an impact on my life during my journey at MSU.

I want to express my acknowledgments to the Monsanto's Beachell-Borlaug International Scholarship Program (MBBISP) for sponsoring me during the first four years of my doctorate. Likewise, I extend my gratitude to the MSU Graduate School for providing me the dissertation completion fellowship in the last semester of my Ph.D. I would also like to thank Dr. Jason and Dr. Dana Lily, and Mr. Chris and Mrs. Judith Rossman for granting me the graduate student funds.

Lastly, I would like to thank my family, my parents Amparo Cruz and Antonio Lopez, and my siblings Juan, Vlady, and Luis, for all their unconditional support to make this dream possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Plant breeding began thousands of years ago when humans started domesticating wild species turning them into crops (Tang et al., 2010). Domestication occurred when the most preferable plants were selected to be propagated for a purpose (e.g., food requirement). The transformation from shattering to non-shattering cereals (e.g., rice and wheat) to facilitate harvest, and the developing of the cultivated maize from its wild relative (teosinte) are two important examples of plants' domestication. In the nineteenth century, hybridization experiments and the discovery of the rules of inheritance by Gregor Mendel and Darwin's theory of natural selection shed light on the foundations of adaptation by means of natural or artificial selection.

Mass phenotypic selection has been extremely successful in producing modern cultivars and hybrids of great agronomic potential (Jiang, 2013). These advances required not only selection for yield potential but also for more robust plants; the development of the high-yielding semi-dwarf wheat and rice varieties during the "Green Revolution" is a clear example of how plant architecture and plan physiology needs to be adapted to improve agronomic performance.

The expected rate of genetic progress from selection is determined by the interplaying of four factors: selection intensity, selection accuracy, genetic variance, and length of the breeding cycle. Despite of the great progress achieved by means of direct phenotypic selection, this technology has several limitations: (*i*) selection accuracy is bounded by trait heritability, (*ii*) extensive (and expensive) phenotypic testing is required to achieve high selection intensity, and (*iii*) the opportunities to shorten the breeding cycle are limited. In the last century, research in plant and animal breeding has been largely focused on the development of technologies that can improve selection by altering the four factors that affect genetic progress.

More accurate predictions of breeding values (BV) can be obtained using statistical methods (e.g., selection indices, Best Linear Unbiased Prediction, BLUP) that smooth-out environmental components of inter-individual differences in phenotypes and enable borrowing of information

between related individuals.

Selection Index (SI) methodology began by predicting BVs obtained by combining individual performance data and progeny evaluations (Lush, 1935). Smith (1936) and Hazel (1943) generalized ideas used in progeny testing to a more general framework (SI) which uses all informative data. A phenotypic measurement is informative about the BV of a selection candidate if it is genetically correlated with the BV. Informative phenotypic measurements include the records of the selection goal on the selection candidate or in relatives of the selection candidate, and measurements of traits genetically correlated with the selection objective. In a selection index, the total genetic value of an individual is predicted using a weighted sum of the available phenotypic measurements. The weights on each of the measured phenotypes depend on genetic and phenotypic (co)variances.

Henderson's BLUP: By the middle of the twenty century, C.R. Henderson (1963) introduced the concept of the Best Linear Unbiased Predictor (BLUP) of the breeding values. BLUPs of BVs are obtained using mixed models that incorporate genetic relationships among all evaluated individuals. Henderson's BLUP can be shown to be equivalent to a SI; however, the BLUP methodology provides an adequate framework for simultaneous modeling of genetic and non-genetic effects (e.g., location, year, year-location, block). BLUP also provides a framework for the estimation of genetic and environmental (co)variance parameters.

Selection index and pedigree BLUP become the method of choice for BV prediction in the second half of the twenty century and the beginning of the current century. However, pedigree information presents some limitations; for example, pedigree and ancestral data cannot predict inter-individual differences in genetic values between members of a bi-parental family.

The development of molecular markers (e.g., DNA markers) has benefited plant science in many aspects, including germplasm characterization, gene introgression, and the development of DNA-assisted prediction/selection methods (Xu & Crouch, 2008).

Marker-assisted selection (MAS) relies on the identification and validation of DNA markers predictive (i.e., tightly linked) of quantitative trait loci (QTL) genotypes. MAS can increase selection accuracy and may enable early screening (e.g., at the seedling stage). MAS has been an

efficient tool in the detection and validation in the improvement of qualitative traits (e.g., disease resistance and grain quality) in various crops (e.g., William et al., 2007; Xu et al., 2009; Miah et al., 2013). However, most of the traits of agronomic performance (e.g., grain yield, oil concentration) are affected by a large number of small-effect loci. Detecting marker-QTL associations for complex traits requires an exceedingly larger sample size (Melchinger et al., 1998) and specific designs aiming to maximize power (e.g., Yu et al., 2008; Zhu et al., 2008). The difficulty of mapping small-effect QTL limited the adoption of MAS for the improvement of traits affected by a large number of QTLs.

Genomic Selection (GS): The continued development of genotyping and sequencing technologies has led to a steady decrease in genotyping costs. In the first decade of the 21st century, arrays, including tens of thousands of SNPs, become available for most agricultural species. These genotyping arrays offered the opportunity to track, via linkage disequilibrium with causal variants, genetics signals distributed over the entire genomes. In a landmark publication, Meuwissen et al. (2001) proposed using a large number of SNPs for breeding value prediction. Genomic selection exploits multi-locus linkage disequilibrium (LD) with the causal variants; with enough marker density, GS can potentially capture all genetic signals (Heffner et al., 2009).

Genomic selection has been successfully adopted by many public and private breeding organizations. Sample size has been recognized as one of the main factors limiting the accuracy of GS (Daetwyler et al., 2008; Goddard, 2009; de los Campos et al., 2013a). Early implementations of GS in plant breeding were based on relatively small training data sets. However, over years, private and public organizations have accumulated large volumes of genomic data linked to phenotypes from multiple populations and multiple generations. The very large sample size of these data sets implies that highly complex genomic prediction models can now be accurately calibrated. However, these data sets are also increasingly heterogeneous, with many subpopulations and multiple generations represented in the data. This translates into potentially heterogeneous allele frequencies and different LD patterns, thus leading to SNP-effect heterogeneity.

GS methods were developed with reference to homogeneous populations in which SNP-effects

can be assumed constant across subsets of the data sets; these methods are not necessarily optimal for the type of data sets available for model training. Therefore, an important focus of this dissertation is on developing novel methods that can leverage the large sample size of modern genomic data sets while coping with the challenges posed by the inherent heterogeneity and complexity of these data sets.

The advent of high-throughput phenotyping (HTP): In recent years, there has been an important improvement in HTP technologies. Modern phenotyping platforms can generate large volumes of data at multiple time-points of a crop (Montes et al., 2007). Examples of this include the use of hyper-spectral imaging technologies which can capture the absorbance of electromagnetic power by crops at potentially thousands of wavelengths. The integration of HTP in genetic evaluations represents a great opportunity to further advance plant breeding; however, the high-dimensional nature of HTP data poses important challenges. Therefore, a second focus of this dissertation is on the development of a novel approach for BV prediction using HTP data.

Thus, the overall aim of this dissertation is to contribute novel methods that can improve the accuracy of genomic prediction by optimizing the use of large, potentially heterogeneous, genomic data sets and by enabling the integration of HTP data. To achieve these goals, we developed a novel statistical approach that combines the standard SI methodology with sparsity-inducing methods commonly used in the field of statistics and machine learning. The procedure that we develop, which we named as sparse selection index (SSI), offers solutions to both GS with potentially highly heterogeneous genomic data sets, and the integration of HTP in genetic evaluations.

1.1 Chapter 2: Incorporating hyper-spectral image data into selection indices for breeding value prediction

The use of HTP technologies can enable screening a larger number of genotypes over many environments and locations, thus offering opportunities to increase selection intensity. Furthermore, indirect selection based on phenotypes collected early in the growing cycle can lead to a reduction in the length of the breeding cycle of perennials. Therefore, the integration of HTP in breeding

evaluations can lead to an increase of genetic progress by either enabling a more intensive and/or faster selection.

Most of the research efforts on crop imaging have focused on developing methods to predict phenotypes from HTP data. For instance, vegetation indices (VI) derived from spectra data (e.g., NDVI; Tucker, 1979), have been used to predict green biomass, leaf area, chlorophyll content, and yield in wheat and maize under field conditions (e.g., Babar et al., 2006; Garriga et al., 2017; Haboudane et al., 2002; Rutkoski et al., 2016); and to detect abiotic (e.g., Roemer et al., 2012) and abiotic stress (e.g., Mahlein et al., 2012) in crops. VIs using information from a reduced number of wavelengths in the spectrum. More recently, researchers have considered developing methods that use whole-spectra data. One approach consists of using dimension-reduction techniques (e.g., principal components, PC, and partial least squares, PLS regression) to predict agronomic traits, e.g., biomass (e.g., Hansen & Schjoerring, 2003) or grain yield in wheat (e.g., Ferrio et al., 2005; Hernandez et al., 2015) and in maize (e.g., Weber et al., 2012; Aguate et al., 2017). Another approach consists of introducing whole-spectra data into Bayesian or penalized regression; this approach has been used to predict milk components from milk-spectra data (Ferragina et al., 2015) and grain yield in wheat (e.g., Montesinos-López et al., 2017) and in maize (e.g., Aguate et al., 2017) from hyper-spectral crop imaging.

The approaches described in the preceding paragraph are well-suited for phenotypic prediction but can be sub-optimal for selection because the best predictor of a phenotype is not necessarily the best predictor of the genetic value.

Therefore, to address the limitations of existing methods, in ***Chapter 2*** we present a novel methodology to develop penalized and reduced-rank SIs using high-dimensional phenotypes. Our approach integrates into a unified framework standard SI methodology with methods used in high-dimensional regression that can prevent over-fitting. We evaluate the proposed methodology using a multi-environment wheat data set ($n \sim 3,200$) containing hyper-spectral ($p = 2,250$ wavebands) and grain yield information. Our results show that penalized and reduced-rank SIs offer improved selection accuracy ($\sim 10 - 40\%$ gain) relative to the standard (i.e., non-regularized) SI.

1.2 Chapters 3 and 4: Improving the accuracy of genomic prediction using sparse selection indices

These two chapters further develop the penalized selection index methodology introduced in Chapter 2 to develop methods to improve the prediction accuracy of GS. The context that motivates the development of these methods is that of large and potentially heterogeneous genomic data sets for which the assumption of effect-homogeneity may not hold.

The challenges posed by the availability of large, potentially heterogeneous, data sets have been recognized in recent years, and there have been several attempts to develop methods to confront these challenges. A first line of research involves using models that include interactions between SNPs and genetic groups. In these models, SNP-effects effects are represented as the sum of a main effect plus a deviation that is group-specific (e.g., Schulz-Streeck et al., 2012; de los Campos et al., 2015; Veturi et al., 2019). A similar (in some cases statistically equivalent) approach is to use multivariate models in which the SNP effects are assumed to be different but correlated between different genetic groups (e.g., Olson et al., 2012; Lehermeier et al., 2015). These methods are well-suited for data sets in which individuals cluster into defined groups which may be known in advance (e.g., breeds, families, pools) or may be inferred using statistical methods (e.g., STRUCTURE; Pritchard et al., 2000). The methods just described have shown promise; however, they are not well suited for data sets in which genetic heterogeneity exhibits more complex/cryptic patterns and thus cannot be reduced to the clustering of individuals into well-defined distinct groups.

Another line of research consists of identifying, for a given prediction set a subset of the training data that may be optimal to predict breeding values of the selection candidates. This approach is motivated by the fact that genetic similarities (e.g., family relationships) have a great impact on prediction accuracy (Habier et al., 2010). Indeed, several studies have shown that the accuracy of GS can be very low when subjects in the training set are genetically distant to those in the prediction set (Clark et al., 2012; Lorenz & Smith, 2015). Furthermore, some studies (de los Campos et al., 2013b; Wolc et al., 2016) suggest that prediction accuracy can be reduced when individuals distantly related to those of the prediction set are also used for model training. Based on

these observations, several studies seek to develop methods for training set optimization/designing (e.g., Clark et al., 2012; Jacobson et al., 2014; Lehermeier et al., 2014; Lorenz & Smith, 2015). Research in this area has focused on comparing various optimization criteria (e.g., Rincent et al., 2012; Akdemir & Isidro-Sanchez, 2019; Roth et al., 2020). However, all these approaches assume that a single training set is optimal for all the subjects in the prediction set; this assumption may not be true because each candidate of selection may draw useful information from different training data points.

Therefore, to overcome the limitations of the existing methods, in ***Chapter 3*** we present a genomic prediction approach that identifies, for each individual in the prediction set, an optimal training set (i.e., a set of support points). Our approach is based on a sparse selection index (SSI) which integrates SI methodology with sparsity-inducing methods (i.e., an L1-penalization). The SSI can be seen as a sparse version of the populate genomic-BLUP (G-BLUP, VanRaden, 2008). We developed software that implements SSI and evaluated the methodology using two multi-environmental wheat data sets, a relatively small ($n = 599$) highly structured one and a very large ($n \sim 29,000$) data set that has a relatively more cryptic structure. In both cases, we found that, compared with the G-BLUP, the SSI can yield an improvement in prediction accuracy of about 5-10%.

Finally, ***Chapter 4*** presents an application of the SSI methodology to a very large ($n = 3,039$) multi-generation double-haploid (DH) maize data set comprising genotype, grain yield, and plant height records. Here, we applied the SSI methodology using additive genomic relationships and also using a (non-linear) Gaussian kernel (K-BLUP). Like the SSI, the Gaussian kernel can be tuned to maximize the borrowing of information between closely related individuals. The sparse and non-sparse K-BLUP models performed similarly with up to 28% of gain in prediction accuracy compared with G-BLUP based on additive relationships. In some cases, the sparse K-BLUP outperformed the non-sparse K-BLUP.

CHAPTER 2

REGULARIZED SELECTION INDICES FOR BREEDING VALUE PREDICTION USING HYPER-SPECTRAL IMAGE DATA

[Material published in: Lopez-Cruz *et al.* 2020. *Scientific Reports* 10(8195)]

Marco Lopez-Cruz[1], Eric Olson[1], Gabriel Rovere[2,3,4], Jose Crossa[6], Susanne Dreisigacker[6], Suchismita Mondal[6], Ravi Singh[6], and Gustavo de los Campos[3,4,5,*]

[1]Department of Plant, Soil and Microbial Sciences, Michigan State University, USA

[2]Department of Animal Science, Michigan State University, USA

[3]Department of Epidemiology and Biostatistics, Michigan State University, USA

[4]Institute for Quantitative Health Science and Engineering, Michigan State University, USA

[5]Department of Statistics and Probability, Michigan State University, USA

[6]International Maize and Wheat Improvement Center (CIMMYT), Mexico

*Correspondence and request should be addressed to G.D.L.C. (e-mail: gustavoc@msu.edu).

2.1 Abstract

High-throughput phenotyping (HTP) technologies can produce data on thousands of phenotypes per unit being monitored. These data can be used to breed for economically and environmentally relevant traits (e.g., drought tolerance); however, incorporating high-dimensional phenotypes in genetic analyses and in breeding schemes poses important statistical and computational challenges. To address this problem, we developed regularized selection indices; the methodology integrates techniques commonly used in high-dimensional phenotypic regressions (including penalization and rank-reduction approaches) into the selection index (SI) framework. Using extensive data from CIMMYT's (International Maize and Wheat Improvement Center) wheat breeding program we show that regularized SIs derived from hyper-spectral data offer consistently higher accuracy for grain yield than those achieved by standard SIs, and by vegetation indices commonly used to predict agronomic traits. Regularized SIs offer an effective approach to leverage HTP data that is routinely generated in agriculture; the methodology can also be used to conduct genetic studies using high-dimensional phenotypes that are often collected in humans and model organisms including body images and whole-genome gene expression profiles.

2.2 Introduction

High-throughput phenotyping (HTP) technologies have been adopted at a fast pace in agriculture; applications range from the use of HTP in highly controlled environments (e.g., growth chambers (Nagel et al., 2012)) to extensive HTP using sensing devices mounted on aerial (e.g., hyper-spectral cameras mounted on aerial vehicles (Araus & Cairns, 2014)) and terrestrial equipment such as tractors and combine harvesters (Montes et al., 2006). Modern agricultural production systems use HTP data to optimize management practices (White et al., 2012), forecast agricultural outputs (Ferrio et al., 2005) and to assess the quality (e.g., protein content) of agricultural commodities (Spielbauer et al., 2009). HTP data can also be a valuable input for breeding programs. For instance, extensive HTP may enable an expansion of genetic testing that can lead to higher

intensity of selection and faster genetic progress. Moreover, HTP data may offer opportunities to improve traits such as drought tolerance that are otherwise difficult to measure and breed for.

Sensors can generate data on hundreds or thousands of phenotypes per unit being monitored. For example, hyper-spectral cameras can generate reflectance of electromagnetic power at hundreds of wavelengths in the visible and infrared spectrum. These measurements can be considered as indicator phenotypes that can be used to predict other traits. An extensive body of research deals with the use HTP data to predict phenotypes such as grain yield (Ferrio et al., 2005; Garriga et al., 2017; Weber et al., 2012; Aguate et al., 2017), dry matter (Montes et al., 2006), oil and protein content (Garnsworthy et al., 2000; Oblath et al., 2016). However, there has been much less research on how to integrate HTP data in genetic studies and in breeding schemes. In genetics, the problem of predicting the genetic merit of a target trait given a set of correlated phenotypes was first addressed by Smith (1936) and Hazel (1943) who introduced the concept of selection index (SI) in plant and animal breeding, respectively.

A SI seeks to improve a target trait y_i (e.g., grain yield) using information from another set of measured traits (e.g., hyper-spectral image data). A linear SI is a weighted sum of the measured phenotypes with weights derived to maximize the correlation between the genetic merit for the selection target and the SI. Thus, the SI methodology offers a natural framework for integrating HTP data into breeding decisions. However, when the measured phenotype is high-dimensional, the naïve application of the SI can lead to overfitting and sub-optimal accuracy of indirect selection.

To address this problem, we developed regularized selection indices (including penalized and reduced-rank methods) that are tailored to achieve accurate prediction of genetic values using high-dimensional phenotypes. The proposed methodology integrates into the SI framework methods often used to prevent overfitting in high-dimensional phenotypic regressions (Hastie et al., 2009). Using extensive multi-environment crop imaging data from CIMMYT's wheat breeding program we show that regularized SIs offer improved accuracy of indirect selection in both optimal and stress environments.

2.3 Results

A selection index is a linear combination of p measured phenotypes, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, of the form $\mathcal{I}_i = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1 ..., \beta_p)'$ is a vector of regression coefficients whose entries define the weights of each of the measured phenotypes in the SI. In a standard SI the weights are derived by minimizing the expected squared deviation between the genetic merit for the selection goal (g_{y_i}, e.g., the genetic merit for grain yield of the i^{th} genotype) and the index, that is:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \frac{1}{2}\mathbb{E}\left(g_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \tag{2.1}$$

The solution to this optimization problem is (see *Methods* section):

$$\hat{\boldsymbol{\beta}} = \mathbf{P}_x^{-1}\mathbf{G}_{x,y}, \tag{2.2}$$

where $\mathbf{G}_{x,y} = \mathbb{E}(\boldsymbol{x}_i)$ is a vector containing the genetic covariances between the selection objective (y_i) and each of the measured traits ($\boldsymbol{x}_i$), and $\mathbf{P}_x$ is the (population) phenotypic variance-covariance matrix of the measured phenotypes, that is, $\mathbf{P}_x = \mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$. Thus, a standard SI takes the form $\mathcal{I}_i = \boldsymbol{x}_i'\mathbf{P}_x^{-1}\mathbf{G}_{x,y}$. The theory underlying the derivation of SIs and response to indirect selection is well established (Bulmer, 1985; Falconer & Mackay, 1996).

The SI is by construction the best linear predictor (BLP) of the genetic merit for the selection goal; this property holds when $\mathbf{G}_{x,y}$ and $\mathbf{P}_x$ are known. However, when the number of measured phenotypes is large, errors in the estimation of $\mathbf{P}_x$ and $\mathbf{G}_{x,y}$ may lead to overfitting and sub-optimal accuracy of indirect selection.

2.3.1 Regularized selection indices

Reduced-rank (e.g., principal components methods) and penalized regression (Hastie et al., 2009) are two approaches commonly used to confront overfitting in high-dimensional regression problems. These methodologies were developed for regression problems involving an observable phenotype (y_i). In the SI, the response (g_{y_i}) is unobservable; however, the same principles that are applied in phenotypic reduced-rank and penalized regressions can be integrated into the SI framework.

2.3.1.1 Reduced-rank selection indices

In principal components (PC) regression, the response is regressed on a reduced number ($q < p$) of PCs extracted from a set of predictors ($\boldsymbol{x}_i$); the same concept can be used to derive a reduced-rank SI. For instance, one can extract a reduced number of PCs from a crop image and the resulting PCs can be used as 'measured traits' in Equation (2.1). The solution of Equation (2.1) will render estimates of the regression coefficients for the PCs, which can be transformed back to coefficients applicable to the measured traits (see *Methods*). Thus, a reduced-rank SI (referred to as PC-SI) can be derived following these steps: (i) extract, using the singular value decomposition, q PCs from the matrix containing the measured phenotypes, (ii) estimate the genetic covariances between the first q PCs and the selection objective, (iii) use these estimated (co)variances to derive coefficients associated with the top q PCs; finally, (iv) transform these coefficients into coefficients for the measured phenotypes. This process can be done using $q = 1, 2, ..., p$ PCs ($q = p$ renders the standard SI). For the sequence of estimates ($\hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, ..., \hat{\boldsymbol{\beta}}^{(p)}$), one can evaluate the accuracy of indirect selection of the resulting SI and an optimal rank for the PC-SI can be chosen to maximize the accuracy of indirect selection.

2.3.1.2 Penalized selection indices

In a penalized regression, regularization is achieved by including in the objective function a penalty on model complexity. In the context of a SI, we have

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left[\frac{1}{2} \mathbb{E} \left(g_{y_i} - \boldsymbol{x}_i' \boldsymbol{\beta} \right)^2 + \lambda J(\boldsymbol{\beta}) \right], \tag{2.3}$$

where λ is a penalty parameter ($\lambda = 0$ yields the coefficients for the standard SI) and $J(\boldsymbol{\beta})$ is a penalty function. Commonly used penalties include the L2 ($||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^{p} \beta_j^2$) and L1 ($||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$) norms (Fu, 1998), or a weighted sum of the two (Zou & Hastie, 2005). Using $J(\boldsymbol{\beta}) = 1/2 \sum_{j=1}^{p} \beta_j^2$ in Equation (2.3) renders a Ridge-regression-type PSI (RR-PSI, see *Methods*):

$$\hat{\boldsymbol{\beta}}^{L2} = (\mathbf{P}_x + \lambda \mathbf{I})^{-1} \mathbf{G}_{x,y},$$

where $\mathbf{I}$ is a $p \times p$ identity matrix. The RR-PSI (referred to as the L2-PSI) yields shrunken estimates of the regression coefficients.

In many applications, variable selection (i.e., a SI that is a function of a subset of the measured phenotypes) may be desirable. This property can be obtained using penalties involving the L1-norm, either alone, $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$ (LASSO (Tibshirani, 1996)), or in combination with the L2-norm, $J(\boldsymbol{\beta}) = 1/2(1-\alpha)\sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|$ (elastic-net (Zou & Hastie, 2005)). Unlike the L2-PSI, the LASSO and elastic-net SIs (hereinafter referred to as L1-PSI and EN-PSI, respectively) do not have a closed-form solution (Hastie et al., 2009). However, solutions for PSIs involving an L1-penalty can be obtained using iterative procedures such as the coordinate descent (Friedman et al., 2007) and the least angle regression (Efron et al., 2004) (LARS) algorithms (see *Methods*). As with the PC-SI, an optimal PSI can be obtained by choosing the values of the regularizing parameters (λ, α) that maximize the accuracy of indirect selection.

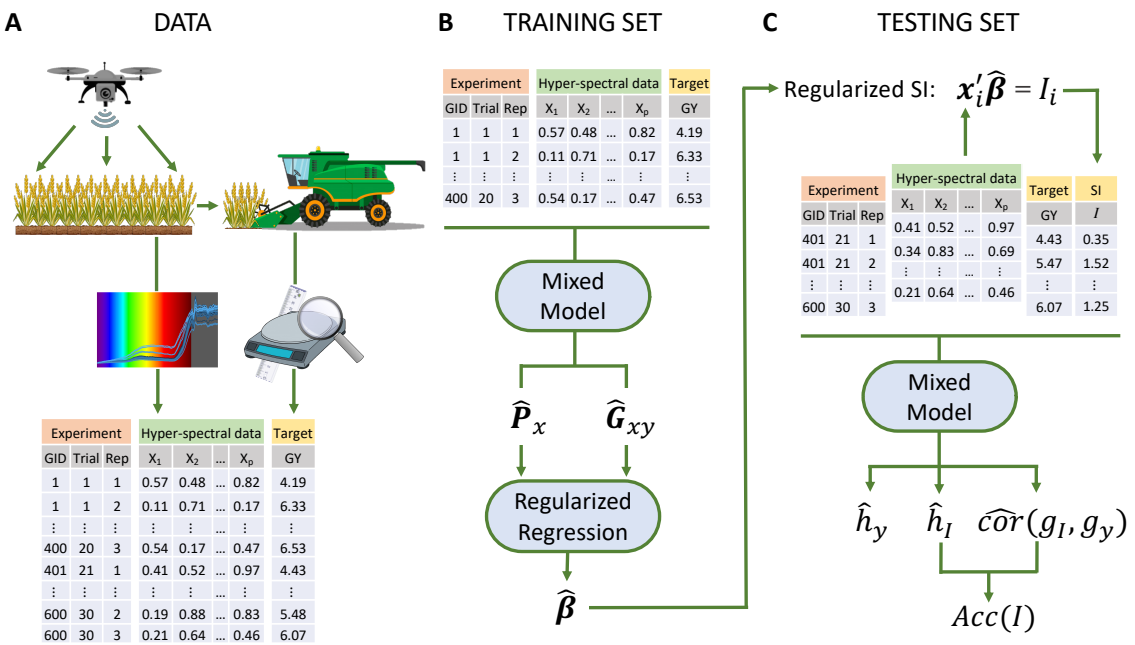

Figure 2.1: Prediction of the genetic merit for grain yield using hyper-spectral crop image data. (A) Data consists of hyper-spectral reflectance data ($\boldsymbol{x}_i$) and phenotypic measurements of the target trait (y_i, e.g., grain yield). (B) A subset of the data (the training set) is used to derive the coefficients of a selection index. (C) These coefficients are then applied to image data of individuals in the testing set to derive the index ($\mathcal{I}_i$) for each individual. The predictive ability of the index is assessed by calculating the accuracy of indirect selection ($Acc(\mathcal{I})$) in the testing set.

2.3.2 Accuracy of indirect selection

Indirect selection accuracy is defined as the correlation between the index used to rank genotypes and the genetic merit of the selection objective, that is, $Acc(\mathcal{I}) = cor(\mathcal{I}_i, g_{y_i})$. This parameter is equal to the product of the square root of the heritability of the SI ($h_{\mathcal{I}}$) times the genetic correlation between the SI and the selection target, $cor(g_{\mathcal{I}_i}, g_{y_i})$ (Falconer & Mackay, 1996). To avoid estimation bias $Acc(\mathcal{I})$ must be estimated using data that was not used to derive the coefficients of the index (Figure 2.1); therefore, in the application presented below we: (i) partitioned the data into training and testing sets, (ii) derived the coefficients of the SI in the training set, (iii) applied these coefficients to image data of the testing set ($\mathcal{I}_i = \boldsymbol{x}_i'\boldsymbol{\beta}$), and ($iv$) estimated $h_{\mathcal{I}}$, $cor(g_{\mathcal{I}_i}, g_{y_i})$, and $Acc(\mathcal{I})$ in the testing set. Furthermore, we quantified the efficiency of indirect selection relative to mass phenotypic selection (RE) using $RE = \frac{h_{\mathcal{I}}}{h_y} cor(g_{\mathcal{I}_i}, g_{y_i})$ (Falconer & Mackay, 1996).

2.3.3 Regularized selection indices for wheat grain yield using hyper-spectral image data

We applied the methodology described in the previous section to data ($n = 3,276$) from the CIMMYT's Global Wheat Program consisting of grain yield (ton ha^{-1}) and hyper-spectral image data. The data were collected at CIMMYT's experimental station in Ciudad Obregon, Sonora, Mexico (27°20' N, 109°54' W, 38 masl) from 39 yield trials in which a total of 1,092 genotypes were tested. Rainfall in Obregon is very limited; therefore, four different environments were generated representing a combination of planting methods (*Flat* or *Bed*), controlled irrigation (minimal, 2 or 5 irrigations), and planting dates (optimum or early-heat). As expected, average yield decreased as drought stress intensity increased (see Table 2.1 and Supplementary Figure A.1 for boxplots of yield by environment).

Image data was collected using an infrared and an hyper-spectral camera and consisted of reflectance of electromagnetic power at 250 wavebands within the visible and near-infrared spectrums (392-850 nm). Separate images were collected at 9 time-points covering vegetative (VEG), grain filling (GF), and maturity (MAT) stages of the crop (see Supplementary Figure A.2). Grain yield

and image data were pre-adjusted using mixed-effects model that accounted for genotype, trial, replicate, and sub-block (see *Methods* section).

Table 2.1: Average grain yield and heritability by environmental condition.

Planting conditions		Number of irrigations	Abbreviation	Average (SD) Yield	Heritability (SD)
Date	System				
	Flat	Minimal	Flat-Drought	2.06 (0.58)	0.83 (0.016)
Optimum		2	Bed-2IR	3.67 (0.43)	0.66 (0.032)
	Bed	5	Bed-5IR	6.11 (0.61)	0.43 (0.025)
Early		5	Bed-EHeat	6.43 (0.73)	0.61 (0.018)

SD: standard deviation.

2.3.4 Regularization improves the heritability and the accuracy of the index

To assess the effect of regularization on the accuracy of indirect selection we fitted an L1-PSI over a grid of values of the regularization parameter ($\lambda^{(1)} > \lambda^{(2)} > ... > 0$ in Equation (2.3), using $\lambda = 0$ renders a standard SI). For each of the solutions ($\hat{\boldsymbol{\beta}}(\lambda^{(1)}), \hat{\boldsymbol{\beta}}(\lambda^{(2)}), ...$) we estimated the heritability of the resulting index and the genetic correlation between the index and the selection target, and from those estimates we derived the accuracy of indirect selection. The same approach was used to evaluate the accuracy of indirect selection of PC-SIs with a varying number ($1, 2, ...$) of PCs.

We first fitted PSIs and PC-SIs using data from a single time-point; the results from the latest time-point (corresponding to MAT or late GF stages depending on the environment) are presented in Figure 2.2 (see Supplementary Figures A.3 to A.5 for other time-points). The heritability of the L1-PSI (Figure 2.2A) decreased as more bands became active in the index. Likewise, the heritability of PC-SI (Figure 2.2B) decreased with the number of PCs used. However, the genetic correlation increased as either more bands become active in the L1-PSI or more PCs were used in the PC-SI. Consequently, the maximum accuracy of indirect selection was achieved with a SI of intermediate complexity (with anywhere between 20 and 60 of the 250 bands being active in the L1-PSI, and between 20-60 PCs in the PC-SI). Results for other time-points and environments (Supplementary Figures A.3 to A.5) exhibited similar patterns with some differences between environments. The

accuracy of indirect selection of the optimal L1-PSI was always close to that of the optimal PC-SI and that of the optimal L2-PSI (Supplementary Table A.1). Importantly, in all cases the accuracy of indirect selection of the optimal regularized SIs was considerably higher than that of the standard SI, which is the one corresponding to 250 active bands or 250 PCs (i.e., the right-most results in the plots in Figure 2.2).

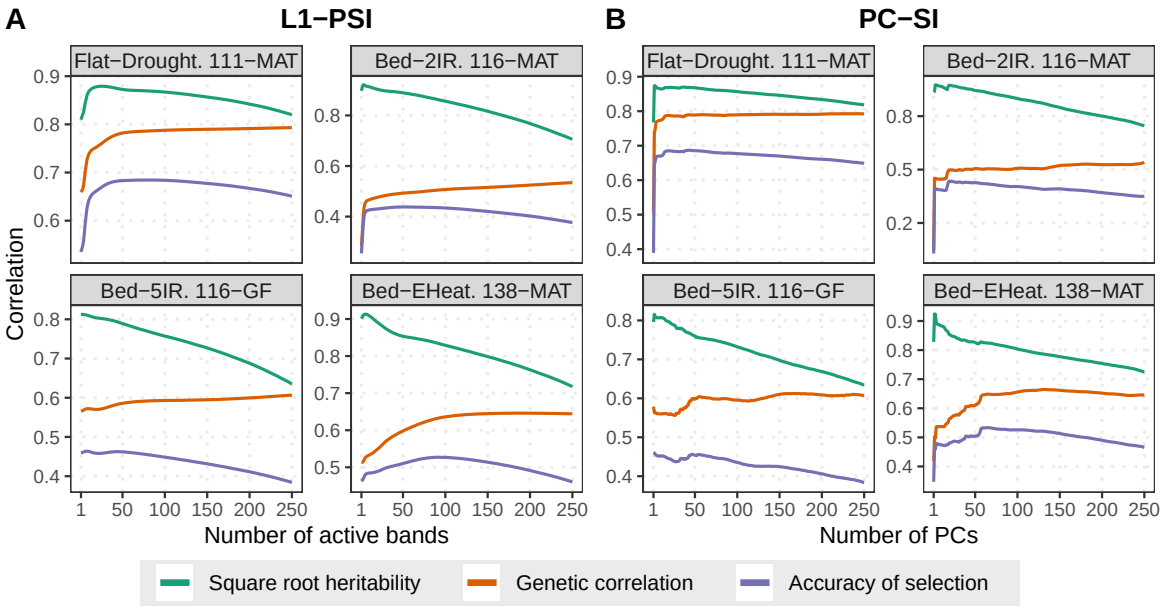

Figure 2.2: Accuracy of indirect selection of regularized SIs and its components. Square root heritability (green), genetic correlation (orange), and accuracy of indirect selection (purple, all averaged over 100 training-testing partitions), versus the number of predictors used to build the index: (A) number of active bands in the case of the L1-PSI, or (B) number of PCs in the PC-SI. Each panel represents one environment (latest time-point).

Figure 2.3 displays the accuracy of indirect selection across time-points for the optimal (i.e., the one with the highest accuracy of indirect selection) L1-PSI and PC-SI. For comparison we also display in the plot the accuracy of indirect selection achieved by a standard SI (in green). The estimated 95% confidence intervals of the accuracy of the regularized SIs (either PC-SI or L1-PSI) are all above (and do not overlap) with the confidence intervals for the accuracy of the standard SI, except for a single time-point (57 DAS in environment *Bed-2IR*). Results from Tukey's Honest Significance Difference confirmed that the accuracy of the regularized SIs is statistically different (higher) than the standard SI at a 5% of significance (Supplementary Table A.1) for all

but one (57 DAS in environment *Bed-2IR*) time-point/environment. Regularization increased the selection accuracy across time-points and environments. Regularized SIs (either PC-SI or L1-PSI) had an accuracy of indirect selection that was in average 10-40% higher than the accuracy achieved by a standard SI. These gains in accuracy were stronger in the optimal environment (*Bed-5IR* with a median of 36%) and smaller in the stressed environments (*Flat-Drought* and *Bed-EHeat* with a median of 16%). Interestingly, there were no sizable differences between the accuracy of indirect selection achieved with the optimal L1-PSI and that of the optimal PC-SI. Compared with a standard SI, regularized SIs had higher heritability (Supplementary Figure A.6); this was achieved without compromising the genetic correlation (Supplementary Figure A.7), thus leading to a higher accuracy of indirect selection achieved by either penalization or rank-reduction strategies.

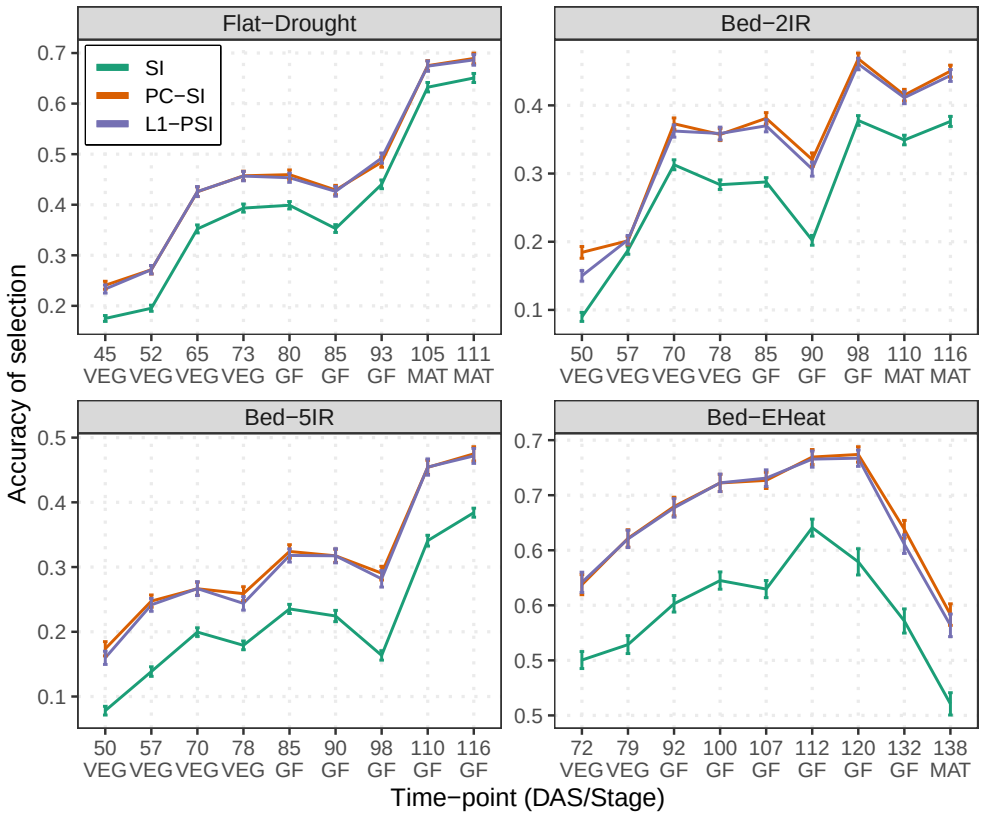

Figure 2.3: Accuracy of indirect selection achieved by a standard (SI) and by regularized (PC-SI and L1-PSI) selection indices. The lines provide the average accuracy over 100 training-testing partitions. Vertical lines represent a 95% confidence interval for the average. The horizontal axis gives the time-point at which images were collected and are expressed in both days after sowing (DAS) and stages (VEG=vegetative, GF=grain filling, MAT=maturity).

2.3.5 Using data from multiple time-points further improves selection accuracy

The results presented above were based on data from a single time-point. We also generated selection indices using data from multiple time-points (in this case, $\boldsymbol{x}_i$ was a vector containing 2,250 traits, corresponding to 250 wavebands measured at each of 9 time-points). Integrating data from multiple time-points further increased the accuracy of L1-PSI by a margin that ranged from 1 to 8 points on the correlation scale (Table 2.2). The gains in selection accuracy obtained using data from multiple time-points were more evident in environments with lower accuracy; similar results were obtained for the PC-SI and L2-PSI (Supplementary Table A.1).

Table 2.2: Accuracy and relative efficiency of indirect selection of an L1-penalized SI using data from one and nine time-points.

	Accuracy (SD)		**Relative Efficiency (SD)**	
Environment	**Best single time-point***	**Nine time-points combined**	**Best single time-point***	**Nine time-points combined**
Flat-Drought	0.69 (0.05)	0.70 (0.05)	0.74 (0.05)	0.75 (0.05)
Bed-2IR	0.46 (0.04)	0.54 (0.03)	0.57 (0.05)	0.67 (0.04)
Bed-5IR	0.47 (0.06)	0.55 (0.05)	0.72 (0.08)	0.83 (0.08)
Bed-EHeat	0.68 (0.04)	0.71 (0.04)	0.88 (0.05)	0.91 (0.04)

Values are presented as an average across 100 training-testing partitions. SD: standard deviation. *For each environment we include the time-point that gave the highest accuracy of selection (see Figure 2.3 for other time-points).

2.3.6 L1-penalization leads to sparse selection indices

Figure 2.4 shows a heatmap for the solutions of the optimal L1-PSI that integrated data from the 9 time-points. Each panel represents an environment, horizontal bands represent time-points. Within each time-point wavebands not entering in the solution are in grey and non-zero coefficients are represented in a yellow-red scale (red indicates large absolute-value coefficients). The well-irrigated environments (*Bed-5IR* and *Bed-EHeat*) had considerably sparser indices with only a reduced number of wavebands in the solutions; these were mostly located in the violet, blue and red regions of the spectrum. In stressed environments (*Flat-Drought* and *Bed-2IR*) there were also

a few wavebands in the green and infrared regions that were active. In all the indices, there were wavebands from several time-points that were active in the optimal solution, suggesting that data from both early and late phenological stages are informative about the genetic merit for grain yield.

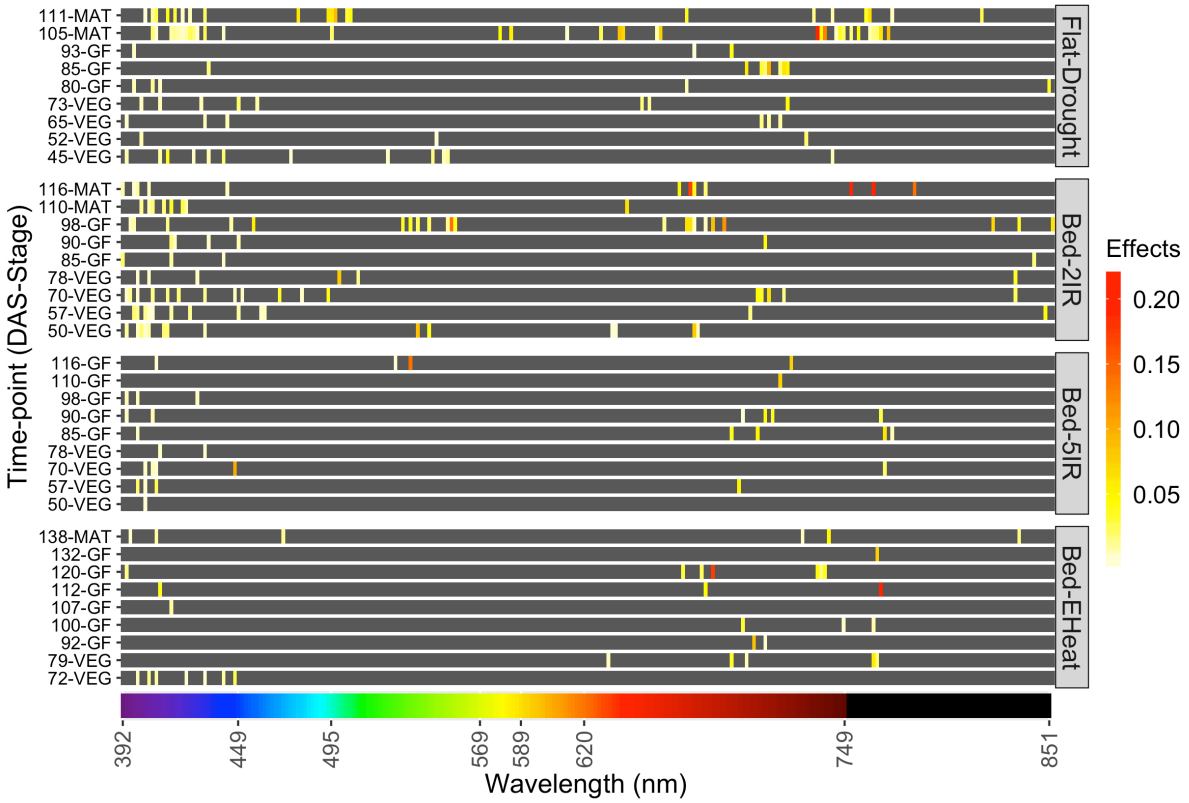

Figure 2.4: Heatmap of regression coefficients for L1-penalized selection indices. Separate indices were derived for each environment using multi time-point data. DAS=days after sowing, VEG, GF, MAT represent vegetative, grain-filling and maturity stages, respectively. The bottom color-bar shows the light color associated with each waveband in the visible spectrum ($\leq$ 750 nm); black was used to represent the near-infrared spectrum (wavelength > 750 nm).

2.3.7 Comparison with phenotypic prediction

We compared the accuracy of indirect selection of the PSI and PC-SI with vegetation indices and penalized phenotypic prediction. Vegetation indices are often used to predict yield (Tattaris et al., 2016), biomass, and chlorophyll content (Babar et al., 2006; Haboudane et al., 2002). We considered two vegetation indices: the Red and Green Normalized Difference Vegetation Indices

(RNDVI (Tucker, 1979) and GNDVI (Gitelson et al., 1996) respectively). For each of these indices we estimated the genetic correlation with grain yield, as well as their heritability and accuracy of indirect selection (Supplementary Table A.1). Overall, the accuracy of indirect selection of the GNDVI and RNDVI was lower than the one achieved with a PSI (the average difference in accuracy between RNDVI and the L1-PSI varied by environment from 0.02 to 0.14 points in correlation, Supplementary Table A.1, in favor of the L1-PSI). The heritability of the GNDVI and RNDVI was similar and superior in some cases to that of the L1-PSI (Supplementary Figure A.6); however, the genetic correlation between the vegetation indices and grain yield was (in most time-points and environments) lower than the genetic correlation between the L1-PSI and grain yield (Supplementary Figure A.7). Thus, the main driver of the difference in accuracy between the L1-PSI and the vegetation indices was the difference in genetic correlation.

We also fitted L1-penalized phenotypic prediction (L1-Phen) and compared the accuracy of indirect selection of these phenotypic prediction methods with that of penalized SIs. Overall, the L1-Phen achieved an accuracy of indirect selection very close to that of the L1-PSI (Supplementary Table A.1); however, in a few environments at some time-points, the L1-PSI achieved a higher accuracy of indirect selection than the phenotypic prediction.

2.4 Discussion

High-throughput phenotyping has been extensively adopted in agricultural research and commercial production. Extracting interpretable information from HTP data poses important statistical challenges. The clear majority of research in this area has focused on calibrating equations to predict phenotypes (e.g., total biomass, grain yield) using HTP data as inputs. This approach is well-suited for phenotypic prediction; however, the same approach can be sub-optimal for selection because the best predictor of a phenotype is not always the best predictor of the genetic merit of the same trait.

The best (linear) phenotypic predictor is the sum of the best linear predictor of the genetic merit (g_y) plus the best linear predictor of the environmental term (ε_y), that is, $\mathbb{E}(y|\boldsymbol{x}) = \mathbb{E}(g_y|\boldsymbol{x})+\mathbb{E}(\varepsilon_y|\boldsymbol{x})$.

The first term, $\mathbb{E}(g_y|\boldsymbol{x})$, is the SI and it is, by construction, maximally correlated with the genetic merit. The second term, $\mathbb{E}(\varepsilon_y|\boldsymbol{x})$, is relevant for phenotypic prediction but represents noise when the problem is that of selecting the best genotypes.

Selection indices exploit genetic covariances, while phenotypic prediction relies on phenotypic covariances between the selection target and the measured phenotype (e.g., crop imaging). Thus, the two methods yield different results whenever the patterns of phenotypic correlations are sufficiently different from the patterns of genetic correlations. In our data set, environmental conditions were highly controlled, with relatively low un-controlled within-trial variability in environmental conditions. Consequently, the patterns of phenotypic and genetic correlations were very similar (see Supplementary Figure A.8). This was true for many time-points and environments but not in others (e.g., 80, 85 and 93 DAS in *Flat-Drought*, and 90 and 98 DAS in *Bed-2IR*); it was exactly in those time-points and environments that the L1-PSI achieved higher accuracy of indirect selection than the L1-Phen method (Supplementary Table A.1).

A standard SI (Equation (2.1)) is, by construction, maximally correlated with the genetic merit of the selection objective. This optimality property holds when the genetic and phenotypic (co)variance matrices that are needed to derive the coefficients of the SI (see Equation (2.2)) are known without error. However, when the measured phenotype is high-dimensional, estimation errors in the phenotypic (co)variance matrix ($\mathbf{P}_x$), as well as in the genetic covariances ($\mathbf{G}_{x,y}$), can make the standard SI sub-optimal. Our empirical results confirm this: standard SIs over-fitted the data; this leads to a SI with low heritability and low accuracy of indirect selection.

To prevent overfitting, we considered integrating ideas commonly used in high-dimensional regression into the SI methodology. Our empirical results show that regularization consistently improves the accuracy of indirect selection relative to standard SIs. We verified this for various environmental conditions and for crop imaging data collected at 9 different time-points. The optimal PSI and the optimal PC-SI achieved almost the same accuracy of indirect selection for all the environments and time-points, suggesting that either type of regularization can be effective.

Reduced-rank selection indices are appealing because after dimension reduction the problem

of deriving a SI is trivial and can be dealt with methods commonly used to derive standard SIs. Moreover, after HTP has been reduced to a few derived-traits (say the top 10 PCs), these traits can be integrated into genetic evaluations (either pedigree-based (Henderson & Quaas, 1976) or genomic-enabled (Meuwissen et al., 2001)) using standard multi-trait models.

Principal components-based methods have been considered before in the analysis of Fourier-transformed infrared (FTIR) spectra derived from milk samples. For instance, Soyeurt et al. (2010) used a reduced number of FTIR-derived PCs to estimate variance components for selection objectives (e.g., fat or protein content in milk). Building upon this idea, Dagnachew et al. (2013) suggested predicting the genetic merit for milk fatty acids using FTIR-derived PCs as 'traits' in a genetic evaluation. However, when mapping from genetic predictions of PC-lodgings onto genetic predictions for the selection objective the authors used coefficients derived from a phenotypic (partial least squares) regression. This does not guarantee that the resulting index is maximally correlated with the genetic merit of the selection target. The penalized and PC-SI presented in this study address that problem by using coefficients that are derived using genetic (and not phenotypic) covariances.

A disadvantage of the PC-SI is that the methodology does not naturally provide variable selection, a feature that may be desirable when the measured phenotype is high-dimensional.

Penalized selection indices can perform variable selection based on genetic covariances. While the derivation of a PSI is a bit more challenging than that of the PC-SI, the computational burden involved in the derivation of a PSI is not extremely high.

2.4.1 Integration of PSI and PC-SI into genetic evaluations

The SIs considered here predict genetic merit for a selection target from a set of traits measured on an individual ($\mathcal{I}_i = \boldsymbol{x}_i'\boldsymbol{\beta}$); such indices exploit borrowing of information between traits within an individual. Borrowing of information between individuals increases selection accuracy; we envision two ways in which regularized SIs can be integrated into pedigree or genomic-based genetic evaluations.

One possibility is to use a two-steps approach whereas in the first step a PSI or a PC-SI is used to predict the genetic merit using within-individual information. This step can be considered as a task where patterns attributable to genetic covariances are extracted and those attributable to environmental covariances are smoothed-out. Then, in a second step, the resulting index-data $\{\mathcal{I}_1, ..., \mathcal{I}_n\}$ could be used as a trait in a genetic evaluation.

Our study shows that the use of a regularized SI leads to a derived-phenotype that has higher genetic accuracy than standard SIs, and that of best phenotypic prediction. In principle, using a more accurate phenotype should lead to a higher accuracy of the predicted breeding values in the second step. However, further studies are needed to determine whether the gains in accuracy at the level of the HTP-derived phenotype will fully translate into a higher accuracy of the predicted breeding values in a two-steps procedure.

A one-step approach is conceptually feasible and statistically more efficient as it offers the possibility of considering correlations between traits, relationships between genotypes, and the effects of non-genetic factors jointly; however, the implementation of the one-step approach using high-dimensional phenotypes can be computationally challenging. To implement a one-step approach, the optimization problem of Equation (2.3) can be modified by replacing $\boldsymbol{x}_i$, the vector with the measured phenotypes on the i^{th} individual, with a vector $\boldsymbol{x} = (\boldsymbol{x}'_1, \boldsymbol{x}'_2, ..., \boldsymbol{x}'_n)'$ that contains all the available HTP data (measured on all n individuals); after expanding the squared error loss and taking expectations we get

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\beta_i} \left[\frac{1}{2}\mathbb{E}\left(g^2_{y_i}\right) - \boldsymbol{\beta}'_i \mathbf{G}_{gx} + \frac{1}{2}\boldsymbol{\beta}'_i \mathbf{P}_x \boldsymbol{\beta}_i + \lambda J(\boldsymbol{\beta}_i) \right]$$

where $\mathbf{G}_{gx}$ is a $pn \times 1$ vector of genetic covariances including between-traits-within-individual (co)variances and between-subjects covariances. In standard genetic models, $\mathbf{G}_{gx}$ takes a Kronecker form $\mathbf{G}_{gx} = \mathbf{A}_i \circ \mathbf{G}_{x,y}$, where $\mathbf{A}_i$ are genetic (either DNA- or pedigree-derived) relationships between the candidate for selection and each of the individuals in the training set, and $\mathbf{G}_{x,y}$ is, as before, a vector of genetic covariances between the selection objective and the measured traits ($\boldsymbol{x}$). Likewise, $\mathbf{P}_x$ is a $pn \times pn$ phenotypic (co)variance matrix. Estimating $\mathbf{P}_x$ would require estimating all the genetic and environmental covariances among the measured traits.

2.4.2 Impact of the use of high-throughput phenotypes in breeding programs

According to breeders' equation (Falconer & Mackay, 1996; Lush, 1937), the rate of genetic gain from selection is directly proportional to selection accuracy and selection intensity. Thus, relative to the use of standard SIs, the use of regularized SIs is expected to increase selection gains by a factor equal to the gains observed in accuracy, that is between 10% and 40%. Relative to mass phenotypic selection, the PSIs had efficiencies, RE, ranging from 60% to 90%; therefore, relative to direct phenotypic selection, selection decisions based on PSI derived from images are expected to yield lower genetic gains than the ones that could be achieved via direct mass selection. However, the use of HTP technologies (e.g., crop monitoring using hyper-spectral cameras mounted in drones) may enable the expansion of the number of genotypes tested/measured as well as the number of locations where those genotypes are tested. This could lead to an increase in selection intensity which will in turn increase selection gains. For instance, if the use of HTP enables doubling the number of genotypes tested, the increase in selection gains that could be achieved with HTP may range from 20% (in the case where the PSI has RE of 60%) to 80% (for the traits/environments with RE of 90%).

The discussion in the preceding paragraph is entirely based on breeders' equation, which does not contemplate the long-term impacts of selection in genetic diversity. A more accurate and more intensive selection may affect diversity and long-term response to selection. To address this problem, attention to diversity will be needed with regularized SIs as with any other selection criteria.

2.4.3 Regularized selection indices can also be a valuable tool in genetic research

High-dimensional phenotypes are also becoming increasingly available in genetic studies involving human subjects and model organisms. Performing genetic studies (e.g., genome-wide association analyses) on high-dimensional phenotypes is challenging and the burden of multiple testing across hundreds or thousands of phenotypes (e.g., RNA-abundance across thousands of genes) may critically compromise power. The PSI and PC-SI presented in this study could be used to extract

genetic patterns from high-dimensional phenotype data such as brain imaging or whole-genome gene expression profiles and these patterns can then be used as traits in genetic studies.

Conclusion. We proposed two novel methods for predicting the genetic merit for selection objectives from high-dimensional phenotypes. These phenotypes are becoming increasingly available as the adoption of HTP in crop and animal production increases. The proposed methods integrate regularization procedures commonly used in high-dimensional regressions into the SI methodology. Regularization prevents overfitting and increases the accuracy of the index. The methods proposed here can be used to extract genetic patterns from almost any kind of high-dimensional phenotype, including not only HTP data emerging in agriculture but also high-dimensional phenotypes that emerge in genetic studies involving human subjects and model organisms.

2.5 Methods

2.5.1 Standard selection index

The weights on a SI are derived as the solution to the optimization problem of Equation (2.1):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \frac{1}{2}\mathbb{E}\left(g_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

The right-hand side can be expressed as $\mathbb{E}(g_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = \mathbb{E}(g_{y_i}^2) - 2\mathbb{E}(g_{y_i}\boldsymbol{x}_i)'\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')\boldsymbol{\beta}$. The first term, $\mathbb{E}(g_{y_i}^2)$, does not involve $\boldsymbol{\beta}$; therefore, it can be dropped from the objective function. Furthermore, if $\boldsymbol{x}_i$ has null mean, and assuming that the environmental effects on $\boldsymbol{x}_i$ are orthogonal to g_{y_i}, then $\mathbb{E}(g_{y_i}\boldsymbol{x}_i) = \mathbf{G}_{x,y}$ is a vector containing the genetic covariances between the selection target and each of the measured phenotypes. Likewise, $\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i') = \mathbf{P}_x$ is the phenotypic (co)variance matrix of $\boldsymbol{x}_i$. Therefore, the problem in Equation (2.1) can be written as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}\left[-\mathbf{G}_{x,y}'\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}'\mathbf{P}_x\boldsymbol{\beta}\right].$$

Differentiating the right-hand side with respect to vector $\boldsymbol{\beta}$ and setting the derivatives equal to zero leads to the first order conditions: $\mathbf{P}_x\hat{\boldsymbol{\beta}} = \mathbf{G}_{x,y}$; therefore,

$$\hat{\boldsymbol{\beta}} = \mathbf{P}_x^{-1}\mathbf{G}_{x,y}.$$

2.5.2 Reduced-rank selection index

Recall that the singular value decomposition of a real-valued matrix, $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]'$ (individuals in rows, phenotypes in columns) takes the form $\mathbf{X} = \mathbf{UDV}'$, where $\mathbf{U} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_p]$ is the matrix containing the left-singular vectors that span the row-space of $\mathbf{X}$, $\mathbf{V} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_p]$ is the matrix with the right-singular vectors, and $\mathbf{D} = diag(d_1, ..., d_p)$ is a diagonal matrix with positive or zero elements. The PCs $\mathbf{W} = \mathbf{XV} = \mathbf{UD}$ are linear combinations of the measured phenotypes. A reduced-rank regression uses the first q PCs ($\tilde{\mathbf{W}} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_q]$, $q \leq p$) as "measured phenotypes" in the SI:

$$\hat{\boldsymbol{\gamma}}^{(q)} = \arg\min_{\gamma} \frac{1}{2}\mathbb{E}\left(g_{y_i} - \tilde{\boldsymbol{w}}_i'\boldsymbol{\gamma}^{(q)}\right)^2,$$

where $\tilde{\boldsymbol{w}}_i$ is a vector containing the scores for the i^{th} observation on the first q PCs. The solution to the optimization problem takes the form $\hat{\boldsymbol{\gamma}}^{(q)} = \mathbf{P}_{\tilde{w}}^{-1}\mathbf{G}_{\tilde{w},y}$, where $\mathbf{P}_{\tilde{w}}$ is the phenotypic (co)variance matrix of the first q PCs and $\mathbf{G}_{\tilde{w},y}$ is a vector containing the genetic covariances between each of the top q PCs and the selection objective. Since the left-singular vectors are orthonormal (i.e., $\boldsymbol{u}_j'\boldsymbol{u}_j = 1$ and $\boldsymbol{u}_j'\boldsymbol{u}_k = 0$, for $j \neq k$), then $\mathbf{W}'\mathbf{W} = \mathbf{D}^2 = diag(d_1^2, \ldots, d_p^2)$. Hence, a method-of-moments estimate of the phenotypic (co)variance matrix of $\tilde{\mathbf{W}}$ contains only the first q elements $\tilde{\mathbf{D}}^2 = diag(d_1^2, \ldots, d_q^2)$; this is

$$\hat{\mathbf{P}}_{\tilde{w}} = \frac{1}{n-1}\tilde{\mathbf{D}}^2$$

Using $\hat{\mathbf{P}}_{\tilde{w}}$ makes the coefficients of the PCs to be proportional to the genetic covariance between each of the PCs and the selection objective: $\hat{\boldsymbol{\gamma}}^{(q)} = (n-1)(\tilde{\mathbf{D}}^2)^{-1}\mathbf{G}_{\tilde{w},y}$. This solution can be mapped to coefficients for the measured traits using $\hat{\boldsymbol{\beta}}^{(q)} = (n-1)\tilde{\mathbf{V}}(\tilde{\mathbf{D}}^2)^{-1}\mathbf{G}_{\tilde{w},y}$, where $\tilde{\mathbf{V}}$ is the matrix containing only the first q right-singular vectors.

2.5.3 Penalized selection indices

The objective function of the penalized SI is given by Equation (2.3). Here we considered PSIs using either L1 or L2-norms or a combination of the two.

L2-PSI. Using an L2-norm as penalty, $J(\boldsymbol{\beta}) = 1/2 \sum_{j=1}^{p} \beta_j^2 = 1/2\boldsymbol{\beta}'\boldsymbol{\beta}$, in Equation (2.3) leads to the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{L2} = \arg\min_{\beta} \left[\frac{1}{2}\mathbb{E}\left(g_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 + \lambda\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta} \right]$$

Therefore:

$$\hat{\boldsymbol{\beta}}^{L2} = \arg\min_{\beta} \left[-\mathbf{G}_{x,y}'\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}'\mathbf{P}_x\boldsymbol{\beta} + \lambda\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta} \right]$$

The second and third right-hand side terms can be combined to obtain:

$$\hat{\boldsymbol{\beta}}^{L2} = \arg\min_{\beta} \left[-\mathbf{G}_{x,y}'\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}' \left(\mathbf{P}_x + \lambda\mathbf{I}\right)\boldsymbol{\beta} \right],$$

where $\mathbf{I}$ is a $p \times p$ identity matrix. Differentiating with respect to $\boldsymbol{\beta}$ and setting the derivatives equal to zero, we obtain the first-order conditions: $(\mathbf{P}_x + \lambda\mathbf{I})\hat{\boldsymbol{\beta}}^{L2} = \mathbf{G}_{x,y}$; therefore:

$$\hat{\boldsymbol{\beta}}^{L2} = \left(\mathbf{P}_x + \lambda\mathbf{I}\right)^{-1}\mathbf{G}_{x,y}$$

EN-PSI. The coefficients for the elastic-net family are obtained by considering an objective function as in Equation (2.3), with $J(\boldsymbol{\beta}) = 1/2(1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|$; therefore,

$$\hat{\boldsymbol{\beta}}^{EN} = \arg\min_{\beta} \left[-\mathbf{G}_{x,y}'\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}'\mathbf{P}_x\boldsymbol{\beta} + \lambda\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta} + \lambda\frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \lambda\alpha\sum_{j=1}^{p}|\beta_j| \right].$$

The L1-PSI and L2-PSI are particular cases corresponding to $\alpha = 1$ and $\alpha = 0$, respectively. When $\alpha = 0$ the solution has a closed-form (see L2-PSI above). If $\alpha > 0$, no closed-form solution exists; however, a solution can be obtained using the same iterative algorithms that are used to solve elastic-net regressions (e.g., LARS and coordinate descent (Hastie et al., 2009)). These algorithms can be implemented either by "partial residuals" or using "covariance updates" (Friedman et al., 2010). In our case, the objective function is entirely based on (co)variance terms. The objects $\mathbf{P}_x$ and $\mathbf{G}_{x,y}$ enter in the objective function of the PSI in the same way that $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\boldsymbol{y}$ enter in a standard elastic-net regression. Therefore, to obtain solutions, we implemented the standard LARS algorithm (e.g., Hastie et al., 2009) entirely based on covariance updates.

2.5.4 Data

The data set consists of 1,092 inbred wheat lines grouped into 39 trials and grown during the 2013-2014 season at the Norman Borlaug experimental research station in Ciudad Obregon, Sonora, Mexico. Each trial consisted of 28 breeding lines that were arranged in an alpha-lattice design with three replicates and six sub-blocks. The trials were grown in four different environments: *Flat-Drought* (sowing in flat with irrigation of 180 mm through drip system), *Bed-2IR* (sowing in bed with 2 irrigations approximately 250 mm), *Bed-EHeat* (bed sowing 30 days before optimal planting date with 5 normal irrigations approximately 500 mm), and *Bed-5IR* (bed sowing with 5 normal irrigations). In 2013, all the trials were planted by mid-November (optimal planting date), on the 21^{st} (*Bed-2IR* and *Bed-5IR*) and on the 26^{th} for *Flat-Drought*. Trials for *Bed-EHeat* were planted on October 30^{th}. Grain yield (ton ha^{-1}, total plot yield after maturity) was recorded.

Reflectance data were collected from the fields using both infrared (A600 series Infrared camera, FLIR, Wilsonville, OR) and hyper-spectral (A-series, Micro-Hyperspec, VNIR Headwall Photonics, Fitchburg, MA) cameras mounted on a Piper PA-16 Clipper aircraft on 9 different dates (time-points) between January 10^{th} and March 27^{th}, 2014. During each flight, data from $p = 250$ wavebands ranging from 392 to 850 nm were collected for each pixel in the pictures. Using ArcMap software (ESRI, CA), the average reflectance of all the pixels within each geo-referenced trial plot was calculated and reported as a single data-point for each genotype for each band. Days to heading were recorded as the number of days from the date of sowing/first irrigation until 50% of spike emergence in each plot. Heading of about 50-80% of the total number of plots was used as criterion to distinguish between vegetative (VEG) and grain filling (GF) stages. The crop was considered to be at maturity (MAT) stage when the average RNDVI decreased to ~ 0.4.

2.5.5 Phenotype pre-processing

Within each environment, grain yield phenotypic records were pre-adjusted by fitting the following mixed model,

$$y_{jklm} = \mu + g_j + t_k + r_{l(k)} + b_{m(kl)} + \varepsilon_{jklm}$$

where y_{jklm} is the grain yield phenotype value for the j^{th} genotype, k^{th} trial, l^{th} replicate (within trial), m^{th} sub-block (within trial and replicate), μ is the overall mean and g_j, t_k, $r_{l(k)}$, and $b_{m(kl)}$ are the genotype, trial, replicate, and sub-block effects, respectively (all assumed to be random) and ε_{jklm} is an error term. Random effects were assumed to be independently and identically distributed (*iid*) normal with null mean and effect-specific variances. Likewise, the error terms were assumed to be iid with null mean and common error variance.

Grain yield data were pre-adjusted by subtracting from the phenotypic record (y_{jklm}) the mean ($\hat{\mu}$) plus BLUPs of trial, replicate, and sub-block effects; this is

$$y^*_{jklm} = y_{jklm} - \hat{\mu} - \hat{t}_k - \hat{r}_{l(k)} - \hat{b}_{m(kl)} = \hat{g}_j + \hat{\varepsilon}_{jklm} \tag{2.4}$$

Reflectance data was pre-adjusted by fitting the above model, using reflectance at individual bands as phenotype expanded with the inclusion of a time-point effect. Separate models were fitted to each of the wavebands. As with grain yield, reflectance data were pre-adjusted by subtracting from the measured reflectance the estimated mean and predicted time-point, trial, replicate, and sub-block effects.

For quality control, pre-adjusted grain yield and reflectance phenotypes were removed for those grain yield scores lying beyond 3 times the inter-quantile region from the 0.25 and 0.75 quantiles. After pre-adjusting, all phenotypes were standardized (to have unit variance); for ease of exposition, hereinafter we refer to the adjusted-scaled phenotypes (including grain yield and the image data) simply as phenotypes.

2.5.6 Heritability estimation

After pre-adjusting standardization, we analyzed phenotypes using a mixed model of the form

$$y_{ij} = g_j + \varepsilon_{ij} \tag{2.5}$$

where y_{ij} is the phenotype for the i^{th} observation (i here is a single index for indices k, l, and m in Equation (2.4)) of the j^{th} genotype; the genetic values are $g_j \overset{iid}{\sim} N(0, \sigma^2_{g_y})$, where $\sigma^2_{g_y}$ is the

genetic variance; and the environmental terms are $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_{\varepsilon_y})$. Plot-basis heritability was calculated from variance components estimates using

$$h^2_y = \frac{\sigma^2_{g_y}}{\sigma^2_{g_y} + \sigma^2_{\varepsilon_y}}.$$

2.5.7 Training-testing partitions

The data set contains information from 39 trials with 84 observations each. To assess the accuracy of indirect selection, we randomly assigned complete trials to testing sets. The training set comprised all the data from the trials not assigned to the testing set. This approach guarantees that no data from a single trial is present in both training and testing sets. This approach aims at representing a situation where one has calibrated the coefficients of the index using historical trials and apply these coefficients to image data of future trials. A similar validation scheme has been used (using herd-year-season groups instead of trials) in validation problems in previous studies involving milk spectra data (Ferragina et al., 2015). In each training-testing partition, out of the 29 trials available, 26 trials ($n_{trn} \approx 2,184$ observations) were randomly assigned to the training set, and the data from the remaining 13 trials ($n_{tst} \approx 1,092$) was used for testing set. The regression coefficients of the indices (the $\boldsymbol{\beta}$'s for the standard SI, PSI, and PC-SI) were calculated using grain yield and reflectance data of the training set. Estimates of the coefficients and reflectance data were then used to calculate the SI, $\mathcal{I}_{ij} = \boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}$, for each observation i in the testing set ($i = 1, ..., n_{tst}$). The heritability of the index and the genetic correlation between the index and the selection goal were estimated in the testing set.

The training-testing procedure described above was repeated 100 times by randomly assigning trials to training and testing sets. From these analyses, we reported the mean of heritability, genetic correlation, and selection accuracy; and their standard deviation across training-testing partitions.

2.5.8 Estimation of phenotypic and genetic parameters

The population phenotypic (co)variance matrix $\mathbf{P}_x$ was estimated within the training set using the unbiased sample (co)variance matrix given by $\hat{\mathbf{P}}_x = \frac{1}{n-1}\sum_{i=1}^{n_{trn}}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'$, where $\overline{\boldsymbol{x}}$ is the vector containing the sample mean of each waveband. Since reflectance data are centered and standardized, this reduces to $\hat{\mathbf{P}}_x = \frac{1}{n-1}\mathbf{X}'\mathbf{X}$, where $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]'$ is the matrix containing all measured traits in the training set.

The genetic covariance ($\mathbf{G}_{x_j,y}$) between grain yield and the j^{th} measured trait ($j = 1, ..., p$) was estimated using a sequence of univariate genetic models as in Equation (2.5). We fitted that model with grain yield phenotypes as response, then for each of the reflectance bands and then for the sum of grain yield and each of the bands. The genetic covariances between the bands and grain yield were then estimated using

$$\hat{\mathbf{G}}_{y,x_j} = \frac{1}{2}\left(\hat{\sigma}^2_{gy+x_j} - \hat{\sigma}^2_{gy} - \hat{\sigma}^2_{gx_j}\right)$$

where $\hat{\sigma}^2_{gy}$, $\hat{\sigma}^2_{gx_j}$ and $\hat{\sigma}^2_{gy+x_j}$ are the estimated genetic variances for grain yield, the j^{th} band, and the sum of grain yield and the j^{th} band, respectively.

2.5.9 Estimation of the accuracy of indirect selection

To assess the accuracy of indirect selection we applied the regression coefficients derived in the training set to image data from the testing set to derive $\mathcal{I}_{ij} = \boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}$. Then, using a mixed model analysis like that described in the previous section we estimated the heritability of the SI ($h^2_{\mathcal{I}}$), the heritability of grain yield (h^2_y), and the genetic correlation between the SI and grain yield ($cor(g_{\mathcal{I}_i}, g_{y_i})$). From these estimates, we derived the accuracy of indirect selection, $Acc(\mathcal{I}) = h_{\mathcal{I}}cor(g_{\mathcal{I}_i}, g_{y_i})$, and the relative efficiency, $RE = \frac{h_{\mathcal{I}}}{h_y}cor(g_{\mathcal{I}_i}, g_{y_i})$.

2.5.10 Software

All the aforementioned analyses were implemented in the R software environment (R Core Team, 2019), version 3.5.1. Linear mixed models were implemented using the "lmer" function from the

LME4 (Bates et al., 2015) R-package. The software that implements the LARS and coordinate descent algorithms are available through the SFSI R-package (https://github.com/MarcooLopez/SFSI).

2.5.11 Materials and data availability

The data used in this study are publicly available by CIMMYT (https://www.cimmyt.org/) who owns all rights in the data. The data is also included in the SFSI R-package. The R-scripts needed to perform the analyses presented in this study can be found in the documentation of the SFSI R-package.

2.6 Acknowledgments

We acknowledge CIMMYT's Global Wheat Program that provided both experimental field and HTP data used in this work. M.L.C. was supported by the Monsanto's Beachell-Borlaug International Scholarship Program (MBBISP).

CHAPTER 3

OPTIMAL BREEDING VALUE PREDICTION USING A SPARSE "FAMILY" INDEX

[Material submitted to: *Genetics* (conditional accepted, revision pending)]

Marco Lopez-Cruz[1] and Gustavo de los Campos[2,3,4,*]

[1]Department of Plant, Soil and Microbial Sciences, Michigan State University, USA

[2]Department of Epidemiology and Biostatistics, Michigan State University, USA

[3]Institute for Quantitative Health Science and Engineering, Michigan State University, USA

[4]Department of Statistics and Probability, Michigan State University, USA

*Correspondence and request should be addressed to G.D.L.C. (e-mail: gustavoc@msu.edu).

3.1 Abstract

Genomic prediction uses DNA sequences and phenotypes to predict genetic values. In homogeneous populations, theory indicates that the accuracy of genomic prediction increases with sample size. However, differences in allele frequencies and in linkage disequilibrium patterns can lead to heterogeneity in SNP effects. In this context, calibrating genomic predictions using a large, potentially heterogeneous, training data may not lead to optimal prediction accuracy. Some studies tried to address this sample size/homogeneity trade-off by designing algorithms to identify an optimal training set; however, this approach assumes that a single training data set is optimum for all individuals in the prediction set. Here, we propose an approach that identifies, for each individual in the prediction set, a subset from the training data (i.e., a set of support points) from which predictions are derived. The methodology that we propose (which we label Sparse Selection Index, SSI) integrates traditional Selection Index methodology with sparsity-inducing techniques commonly used in high-dimensional regression settings. The sparsity of the resulting index is controlled by a regularization parameter (λ); the G-BLUP (the prediction method most commonly used in plant and animal breeding) appears as a special case which happen when $\lambda = 0$. In this study, we present the methodology and demonstrate, using two wheat data sets (a very large multi-generation breeding panel and a smaller, highly-structured, data set) with phenotypes collected in ten different environments, that the SSI can achieve significant (anywhere between 5-10%) gains in prediction accuracy relative to the G-BLUP.

3.2 Introduction

Selection decisions in plant and animal breeding rely on the predicted genetic merit of selection candidates. Early prediction methods were based either on phenotypes measured in the same selection candidates or on progeny testing (e.g., Lush, 1935). These methods were later extended into selection indices (Hazel, 1943; Smith, 1936) that can use information from various sources of correlated data, including secondary traits measured on the same individual, measurements of

the same phenotype collected from relatives, and combinations thereof (Lush, 1948). Henderson (1950) further extended the methodology by developing mixed-models that can include fixed and random effects.

The Best Linear Unbiased Predictor (BLUP) predicts breeding values by borrowing (i.e., averaging) information from multiple sources of correlated data. Pedigrees often trace back a limited number of generations (e.g., 5); therefore, in most animal and plant breeding data sets, pedigree relationships often define "families" and borrowing of information span within the scope of each family. However, this is not the case in genomic-BLUP (G-BLUP; VanRaden, 2008) because genomic relationships are not sparse as pedigree-derived relationships.

In the last two decades, genomic prediction (i.e., genomic selection, GS; Meuwissen et al., 2001) has become the method of choice for breeding value prediction. GS models predict breeding values using genome-wide markers and rely in the multi-locus linkage disequilibrium (LD) between SNPs and quantitative trait loci (QTL). However, it is also well-established that family relationships and population structure contribute to the accuracy of genomic prediction (Habier et al., 2007). In a Genomic relationship matrix (VanRaden, 2007) all individuals are related to some extent; therefore, every training data point contributes to the prediction of each individual in the testing set.

Genomic prediction models were originally developed with reference to a homogeneous population in which marker effects are assumed to be the same across subgroups of the data. However, several factors, including imperfect LD between markers and QTL and non-additive effects coupled with population structure and admixture can make marker effects vary across subgroups in the sample (de los Campos et al., 2015; Pritchard & Donnelly, 2001). All these factors can make the genomic relationships derived from markers inaccurate predictors of the genomic relationships realized at causal loci (e.g., de los Campos et al., 2013b). Therefore, the accuracy of G-BLUP may be sub-optimal when the training data consists of heterogeneous groups (e.g., multiple families or multiple strains or breeds) or even multi-generation data in which LD patterns may vary across distant generations.

Several authors have recognized the need to model heterogeneous SNP-effects in the context of

multi-breed (e.g., Hayes et al., 2009) and structured (e.g., de los Campos et al., 2015) data. Most of the existing methods model group-specific effects using either multivariate Gaussian models (e.g., Olson et al., 2012; Schulz-Streeck et al., 2012) or interaction models (e.g., de los Campos et al., 2015; Isidro et al., 2015; Veturi et al., 2019). However, these approaches can be difficult to use in the presence of cryptic genetic-heterogeneity patterns where no clear groups can be discerned.

Another line of research seeks to identify an optimal training set for a given prediction set. These optimal training sets often consist of individuals that are closely related to the individuals in the prediction set, i.e., the candidates of selection (Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015; Pszczola & Calus, 2016; Akdemir & Isidro-Sanchez, 2019). However, these methods assume that a single training set is optimal for all the individuals in the prediction set which is not necessarily the case. Therefore, in this study, we focus on developing a genomic prediction method that will identify, for each individual in a prediction set an optimal training set (i.e., a set of support points). Our approach achieves this goal by integrating sparsity (by adding an L1-penalty) into a selection index (SI) problem, we refer to the method as a sparse selection index (SSI).

3.3 Materials and Methods

A standard selection index ($\mathcal{I}_i$) predicts the breeding value of an individual (u_i) using a linear combination of the training phenotypes ($\boldsymbol{y} = (y_1, ..., y_n)'$): $\mathcal{I}_i = \boldsymbol{\beta}_i'\boldsymbol{y} = \sum_{j=1}^{n} \beta_{ij} y_j$. Here, phenotypes are assumed to be centered and corrected by non-genetic effects (e.g., experiment and block effects), and $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{in})'$ is a vector of weights which are obtained as the solution to the following optimization problem:

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\beta_i} \frac{1}{2}\mathbb{E}\left(u_i - \boldsymbol{\beta}_i'\boldsymbol{y}\right)^2$$

The right-hand side of the above problem expands to $\mathbb{E}(u_i^2) + \boldsymbol{\beta}_i'\mathbb{E}(\boldsymbol{y}\boldsymbol{y}')\boldsymbol{\beta}_i - 2\mathbb{E}(\boldsymbol{y} \times u_i)'\boldsymbol{\beta}_i$. Assuming that genetic (u_i) and non-genetic effects (ε_i) are independent, each with mean zero and (co)variance matrices $\text{var}(\boldsymbol{u}) = \sigma_u^2\mathbf{G}$ and $\text{var}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2\mathbf{I}$, we have that $\mathbb{E}(u_i - \boldsymbol{\beta}_i'\boldsymbol{y})^2 = \sigma_u^2 + \boldsymbol{\beta}_i'\mathbf{P}\boldsymbol{\beta}_i - 2\sigma_u^2\mathbf{G}_i'\boldsymbol{\beta}_i$, where σ_u^2 is a genetic variance parameter, $\mathbf{P} = \sigma_u^2\mathbf{G} + \sigma_\varepsilon^2\mathbf{I}$ is the phenotypic (co)variance

matrix of the training phenotypes, and $\mathbf{G}_i$ is a vector containing the genetic relationships between the i^{th} subject of the prediction set and each of the subjects in the training data. Since σ^2_u does not depend on $\boldsymbol{\beta}_i$, the aforementioned optimization problem can be reduced to

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\beta_i} \left[\frac{1}{2} \boldsymbol{\beta}_i' (\mathbf{G} + \lambda_0 \mathbf{I}) \boldsymbol{\beta}_i - \mathbf{G}_i' \boldsymbol{\beta} \right] \tag{3.1}$$

where $\lambda_0 = \frac{\sigma^2_\varepsilon}{\sigma^2_u} = \frac{1-h^2}{h^2}$ is the ratio of the error to the genetic variance, which can be expressed in terms of the heritability, h^2. The solution to the above problem can be shown to be

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{G} + \lambda_0 \mathbf{I})^{-1} \mathbf{G}_i \tag{3.2}$$

The vector $\hat{\boldsymbol{\beta}}_i$ can be shown to be the i^{th} row of the Hat matrix of the BLUPs of the genetic values of the individuals in the prediction set (see Appendix B.1 for a proof), thus, depending on whether $\mathbf{G}$ is a pedigree- or genomic-derived relationship matrix, the standard SI is equivalent to a pedigree- (Henderson, 1963) or genomic-BLUP, respectively.

When $\mathbf{G}$ is a pedigree-based relationship matrix the off-diagonal entries corresponding to pairs of subjects not connected through the pedigree are equal to zero. In that case, some of the entries of $\hat{\boldsymbol{\beta}}_i$ can also be equal to zero which implies that the corresponding predicted breeding value ($\hat{\mathcal{I}}_i = \hat{\boldsymbol{\beta}}_i' \boldsymbol{y}$) draws information from a subset of the training data. However, when $\mathbf{G}$ is a genomic relationship typically none of the off-diagonals are equal to zero; therefore, none of the entries of $\hat{\boldsymbol{\beta}}_i$ will be exactly equal to zero. This implies that all the observations in the training set contribute to some extent to predict the breeding values of all the individuals in the prediction set.

3.3.1 Sparse Selection Index Methodology

As noted earlier, there are several reasons (e.g., imperfect LD, effect heterogeneity) why borrowing of information between distantly related individuals may have a detrimental effect on prediction accuracy. Therefore, to achieve sparsity (and possibly differential shrinkage on the $\hat{\boldsymbol{\beta}}_i$) we considered adding an L1-penalty to the objective function in Equation (3.1); therefore,

$$\tilde{\boldsymbol{\beta}}_i = \arg\min_{\beta_i} \left[\frac{1}{2} \boldsymbol{\beta}_i' (\mathbf{G} + \lambda_0 \mathbf{I}) \boldsymbol{\beta}_i - \mathbf{G}_i' \boldsymbol{\beta} + \lambda \sum_{j=1}^{n} |\beta_{ij}| \right] \tag{3.3}$$

The above optimization problem does not have a closed-form solution; however, solutions can be obtained using a Coordinate Descent algorithm very similar to the one used to solve LASSO problems (see Lopez-Cruz et al., 2020). The regularization parameter λ controls how sparse $\tilde{\boldsymbol{\beta}}_i$ will be; this parameter is also expected to affect the accuracy of the index. Therefore, an optimal value of λ can be found by maximizing the accuracy of the resulting index.

3.3.2 Data

We used two wheat breeding data sets to evaluate and to compare the prediction performance of standard and sparse selection indices. The first data set (Wheat-large) is a multi-generation wheat breeding data set of a very large sample size ($n \sim 29,000$). The second one is (Wheat-599) is a small, structured data (see Supplementary Figure B.1).

The Wheat-large data set is from CIMMYT's Global Wheat Program and it includes phenotype data from 58, 798 wheat lines that were evaluated during five years (2009-2013) at the CIMMYT's experimental station in Ciudad Obregon, Mexico. Lines were evaluated under six environmental conditions (*B2I*, *B5I*, *MEL*, *LHT*, *DRB*, *EHT*) representing a combination of planting system (bed vs flat, the later referred to as melgas), number of irrigations (2, 5 irrigations or drip irrigation), and sowing date (optimum, late or early planting). Each year, grain yield trials were established in an α-lattice design with three replicates into incomplete blocks. Moisture-standardized grain yield (ton ha^{-1}) was measured at each plot. We used mixed-effects models with a ('fixed') intercept and the random effects of the trial, block (within trial) and replicate (within trial) to derive least-square means by line and environmental condition. Separate mixed models were fitted to data from each of the simulated environments. The average grain yield in this data set varied from 2.72 to 7.12 ton ha^{-1} (see Supplementary Figure B.2B for boxplots of grain yield) and the heritability of single-plot records varied between 0.23 and 0.57 (see Supplementary Table B.1 for a summary of the data). Only a subset of 29, 484 genotypes was genotyped using a GBS (Genotyping-by-sequencing) technology that yielded 42, 706 SNPs. We removed SNPs with more than 70% of missing values and those with minor allele frequency lower than 5%. After applying these filters, we retained

9,045 SNPs. The missing values at each SNP were imputed as the mean of the observed SNP data across genotypes. The data set has been previously described and analyzed by Pérez-Rodríguez et al. (2017).

The Wheat-599 data set is also from CIMMYT's Global Wheat Program and it is comprised of grain yield and genotype data for 599 historical inbred lines derived along 25 years. Lines were evaluated in the Elite Spring Wheat Yield Trials (ESWYT) that were grouped into four different mega-environments (*Env1*, ..., *Env4*). The available phenotypic values are least-square means from two replicates. The average grain yield in this data set ranged from 3.23 to 5.14 ton ha^{-1} (see Supplementary Figure B.2A for boxplots of grain yield data) with heritability estimates for the least-square means ranging between 0.43 to 0.50 (see Supplementary Table B.2). Each of the lines was genotyped for 1, 279 diversity array technology (DArT) markers. The data set is available with the BGLR R-package (Perez & de los Campos, 2014) and has been described and analyzed by previous authors (e.g., de los Campos et al., 2009b; Crossa et al., 2010).

3.3.3 Analyses

For each data set, we computed a genomic relationship matrix $\mathbf{G}$ using (centered and standardized) marker information, $\mathbf{X} = \{x_{im}\}$, as $\mathbf{G} = \mathbf{ZZ}'/p$, where p is the number of markers and $\mathbf{Z} = \{(x_{im} - \overline{x}_m)/sd_{x_m})\}$ is the matrix of centered and standardized markers obtained by subtracting from each marker entry the mean of each column ($\overline{x}_m$) followed by scaling by the standard deviation of the column (sd_{x_m}). The resulting matrix has an average of the diagonal elements equal to 1.

To quantify the prediction accuracy of each of the indices, we divided each data set into training (trn) and testing (tst) sets by randomly assigning 30% (70%) of the data points to testing (training). Predictions were derived by first using Equation (3.2):

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{G} + \lambda_0 \mathbf{I})^{-1} \mathbf{G}_i$$

(for the standard SI) and Equation (3.3):

$$\tilde{\boldsymbol{\beta}}(\lambda)_i = \arg\min_{\beta_i} \left[\frac{1}{2} \boldsymbol{\beta}_i' (\mathbf{G} + \lambda_0 \mathbf{I}) \boldsymbol{\beta}_i - \mathbf{G}_i' \boldsymbol{\beta} + \lambda \sum_{j=1}^{n} |\beta_{ij}| \right]$$

(for the SSI), with $\mathbf{G} = \mathbf{G}_{trn}$ representing the genomic matrix of the training data points (i.e., with dimensions $n_{trn} \times n_{trn}$, where $n_{trn} = 0.7n$), and $\mathbf{G}_i = \mathbf{G}_{trn,tst(i)}$ being the vector containing the genomic relationships between the i^{th} data-point of the testing set, with each of the individuals assigned to the training set (i.e., the dimensions of $\mathbf{G}_i$ are $n_{trn} \times 1$). This was repeated for each individual in the testing set ($i = 1, \ldots, n_{tst}$, where $n_{tst} = 0.3n$). Subsequently, predictions for each individual were obtained using $\hat{I}_i = \hat{\boldsymbol{\beta}}_i' \boldsymbol{y}_{trn}$ (for the standard SI) and $\hat{I}_i = \tilde{\boldsymbol{\beta}}(\lambda)_i' \boldsymbol{y}_{trn}$ (for the SSI) where $\boldsymbol{y}_{trn}$ is a $n_{trn} \times 1$ vector with the adjusted-centered phenotypes of the training set.

The implementation of the SI requires heritability estimates. We derived those by fitting a G-BLUP model of the form $y_i = \mu + u_i + \varepsilon_i$ with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2_\varepsilon)$ and $\boldsymbol{u} \sim N(\mathbf{0}, \sigma^2_u \mathbf{G})$. The model was fitted using the rrBLUP R-package, separate models were fitted to grain yield within each environment in each data set within the training set. We then used the variance parameters estimates to derive $h^2 = \sigma^2_u/(\sigma^2_u + \sigma^2_\varepsilon)$ for grain yield.

Prediction accuracy (ρ) was measured with the correlation between the phenotype and the index, divided by the square-root of the trait heritability of the trait, $\rho = Acc(\hat{I}) = cor(\hat{I}_i, y_i)/h$ (Dekkers, 2007).

For the SSI, we estimated accuracy over a grid of values of the regularization parameter ($\lambda = 0 < \lambda^{(1)} < \lambda^{(2)} < \cdots < \lambda_{max}$) where $\lambda_{max} = \underset{i}{max}\{\frac{|\mathbf{G}_i|}{diag(\mathbf{G})+\lambda_0}\}$. Here λ_{max} is the minimum value of λ that yields an SSI with no active predictors, and $\lambda = 0$ gives the weights of the standard SI. Following Friedman et al. (2010) we used a grid of values evenly spaced in the logarithm scale with a total of 100 values. Thus, for each value of λ in the grid, we had an estimate of the resulting accuracy of the SSI. This was used to profile accuracy as a function of the regularization parameter and also to choose an optimal value of λ.

To determine an optimal value of λ we implemented a calibration analysis using data from the training data only. Specifically, for each training set, we conducted an internal cross-validation (CV) as follows: (i) The training data set was partitioned into k subsets. (ii) SSIs were derived over a grid of values of λ using data from $k - 1$ folds for training and the data in the k^{th} fold as testing (i.e., for the estimation of accuracy, see the previous paragraph). (iii) The resulting curves

profiling accuracy (ρ) by values of λ were used to identify the value of λ ($\hat{\lambda}_{cv}$) that maximized accuracy. (iv) Finally, we used all the data from the training set to derive $\mathcal{I}_i(\hat{\lambda}_{cv})$ and evaluated the accuracy (ρ) of the resulting index in the left-out data from the testing set.

3.3.4 Software

All the analyses were performed in the R environment-language (R Core Team, 2019) version 3.5. The heritability of each of the traits was estimated using the rrBLUP R-package (Endelman, 2011). Sparse SIs were derived using the SSI function from the SFSI R-package that implements the Coordinate Descent algorithm described in Lopez-Cruz et al. (2020). The package is aided by ggplot2 (Hadley, 2016) and parallel (R Core Team, 2019) packages to visualize results and to speed computation. This package is available through the GitHub repository at https://github.com/MarcooLopez/SFSI. Scripts illustrating the use of this package using the Wheat-599 data set are presented in the Appendix B.2.

3.3.5 Data availability

Both phenotypic and marker data for the Wheat-large data set can be downloaded from CIMMYT's repository at http://genomics.cimmyt.org/wheat_50k/PG/ (accessed Sep 14th, 2020). The Wheat-599 data set can be downloaded from the BGLR R-package by calling "data(wheat)". All supplemental figures and tables are contained in the Appendix B. All supplemental files are available at Figshare at https://figshare.com/s/ce2258d168b16a86454d.

3.4 Results

3.4.1 Sparsity improves prediction accuracy

We assessed the effect of sparsity on the accuracy, by fitting the SSI for 100 values of λ ($0 < \lambda^{(1)} < \lambda^{(2)} < \cdots < \lambda_{max}$; the value $\lambda = 0$ produces the coefficients of the standard SI or G-BLUP). The results (averaged over 100 trn-tst partitions) are shown in Figure 3.1. The number of support points

(i.e., the number of training data points contributing to the prediction) was, as expected, inversely proportional to λ; therefore, to facilitate interpretation, the x-axis of Figure 3.1 is displayed as the average number of support points, which is more meaningful than the λ values. The accuracy of the G-BLUP is also shown at the rightmost side of the plot whose number of support points is equal to the size of the training data set. Intermediate values of λ led to sparse indices that, in most cases, achieved higher prediction accuracy than that of the G-BLUP (shaded "belly" area in Figure 3.1).

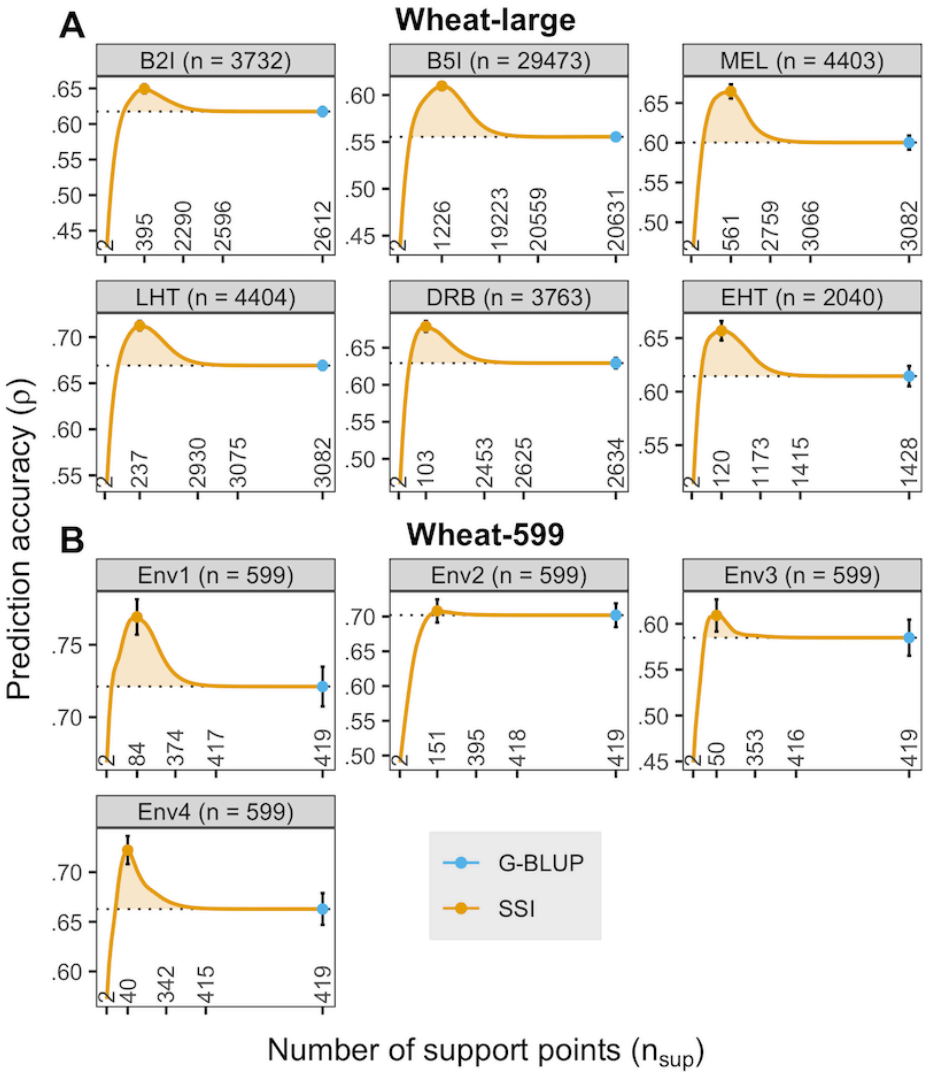

Figure 3.1: Prediction accuracy (average across 100 trn-tst partitions) of the SSI versus the (average) number of predictors in training set supporting the SSI of each individual in testing set (x-axis). Genomic-BLUP (blue rightmost point) appears as a special case of an SSI. Each panel represents one environment within data set. (A) Wheat-large data set. (B) Wheat-599 data set. Vertical bars represent a 95% confidence interval for the average.

The maximum accuracy in the environment *EHT* (see Figure 3.1A) was obtained with a penalization that leads to a sparse index with an average of 120 support points (n_{sup}). This predictive set of individuals represents around 8% of the total training set ($n_{trn} = 1,428$) available for prediction.

For the small data set (Figure 3.1B), the same "belly" pattern can be observed in all environments, except for environment 2. This case shows that the SSI does not always outperform the G-BLUP; however, the SSI achieves the prediction accuracy of the G-BLUP with a smaller support set ($n_{sup} \approx 151$ out of 419).

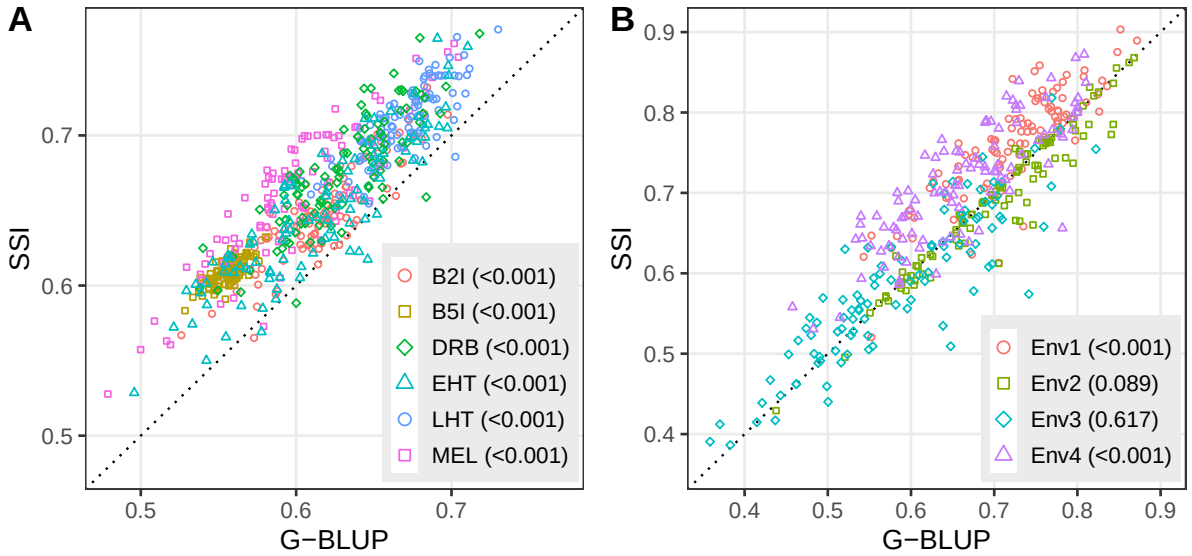

Figure 3.2: Prediction accuracy of the optimal sparse selection index (SSI) versus that of the G-BLUP. Each point represents a trn-tst partition (a total of 100 partitions were implemented), the point shape and color represent environments. (A) Wheat-large data set. (B) Wheat-599 data set. The value of λ in the SSI was estimated using 10 5-fold cross-validations conducted within the training data. In parenthesis, by the legend, is the p-value for the two-sided Sign (binomial) test for within-environment differences in accuracy between the SSI and the G-BLUP.

3.4.2 Using an internal cross-validation to achieve optimal sparsity

The results in Figure 3.1 suggest that one can find a value of λ that leads to an index with a predictive performance as least as high (and in most cases higher) as the G-BLUP. However, to obtain an unbiased estimate of the maximum accuracy that one could achieve with an SSI, one

should not use data from the testing set to select the optimal value of λ. Therefore, we repeated the analyses described above, this time performing the grid search for an optimal value of λ by implementing 10 5-fold CVs within each training data set. This CV was used to choose an optimal value of λ ($\hat{\lambda}_{cv}$). Then, we solved the SSI using $\hat{\lambda}_{cv}$ and all the training genotypes, and evaluated the accuracy of $\mathcal{I}_i(\hat{\lambda}_{cv})$ in a testing set that was not used to choose $\hat{\lambda}_{cv}$. This was repeated for 100 trn-tst partitions. Figure 3.2 shows the accuracy of $\mathcal{I}_i(\hat{\lambda}_{cv})$ versus that of the G-BLUP, each point in the plot represents a trn-tst partition.

Table 3.1: Prediction accuracy (average across 100 partitions) achieved by sparse selection indices (SSIs) and by the G-BLUP (standard SI), by data set and environmental condition.

Environment	$\boldsymbol{n_{tst}}$	$\boldsymbol{n_{trn}}$	**Method**	$\lambda_{cv}{}^a$	$\boldsymbol{n_{sup}}{}^b$	**Accuracy (SD)**^{c}	**Counts**^{d}
Wheat-large							
B2I	1,120	2,612	G-BLUP	0.0000	2,612	0.617 (0.031) b	97
			SSI	0.0135	434	0.648 (0.031) a	
B5I	8,842	20,631	G-BLUP	0.0000	20,631	0.555 (0.010) b	100
			SSI	0.0107	1,470	0.609 (0.009) a	
MEL	1,321	3,082	G-BLUP	0.0000	3,082	0.600 (0.045) b	99
			SSI	0.0131	524	0.661 (0.046) a	
LHT	1,322	3,082	G-BLUP	0.0000	3,082	0.669 (0.024) b	99
			SSI	0.0168	380	0.709 (0.025) a	
DRB	1,129	2,634	G-BLUP	0.0000	2,634	0.629 (0.035) b	98
			SSI	0.0322	136	0.675 (0.037) a	
EHT	612	1,428	G-BLUP	0.0000	1,428	0.614 (0.049) b	94
			SSI	0.0301	178	0.649 (0.047) a	
Wheat-599							
Env1	180	419	G-BLUP	0.0000	419	0.721 (0.070) b	87
			SSI	0.0413	78	0.760 (0.067) a	
Env2	180	419	G-BLUP	0.0000	419	0.702 (0.087) a	41
			SSI	0.0123	254	0.692 (0.085) a	
Env3	180	419	G-BLUP	0.0000	419	0.585 (0.101) a	53
			SSI	0.0613	84	0.586 (0.093) a	
Env4	180	419	G-BLUP	0.0000	419	0.663 (0.082) b	87
			SSI	0.0617	54	0.714 (0.075) a	

SD: Standard deviation across the 100 trn-tst partitions. G-BLUP model corresponds to an SSI where $\lambda = 0$. ^{a}Average value of λ estimated by cross-validating the training set. ^{b}Average number of individuals in training set supporting the prediction of individuals from testing set. ^{c}Models with the same letter are not significantly different from others (ANOVA followed by Tukey's HSD test, 5% significance level). ^{d}Number of times (out of the 100 partitions) that the SSI outperformed the G-BLUP in prediction accuracy.

In the Wheat-large data set, the optimal SSI outperformed the G-BLUP in 94% of the cases (Table 3.1). For this data set, the SSI offered significant (according to Tukey's Honest Significance Difference test, HSD) gains in accuracies across environments. These gains range from 5% (in the environment B2I) to 10% (in the environment *MELGAS*) in the correlation metric.

Similar patterns were observed with the Wheat-599 data set. In Environments 1 and 4 the SSI outperformed the G-BLUP in more than 80% of the trn-tst partitions (Table 3.1), with gains in accuracy ranging from 5-7%. However, in Environments 2 and 3, there were no statistically significant gains in accuracy (see Table 3.1).

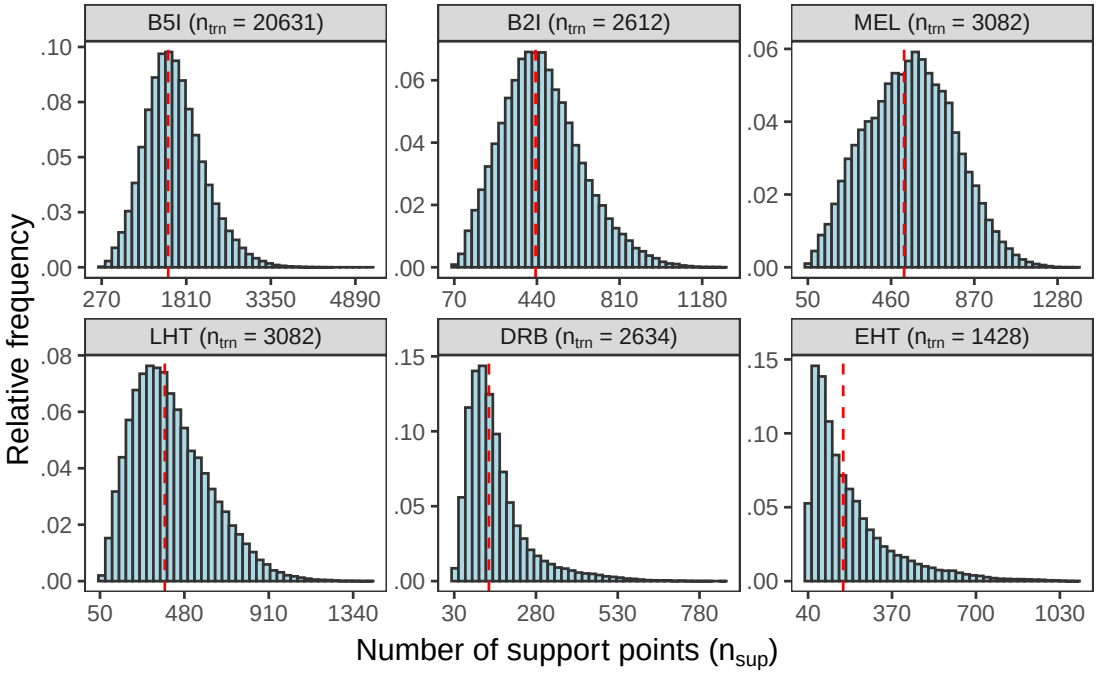

Figure 3.3: Distribution of the number of training support points (n_{sup}) in optimal sparse selection indices (results obtained over 100 trn-tst partitions; n_{trn}= size of the training data set), by environmental condition, Wheat-large data set.

3.4.3 Sparse Selection Indices build subject-specific training sets

For each individual in the prediction set, an SSI yields a set of support points in the training set consisting of the index of all the non-zero entries of $\tilde{\boldsymbol{\beta}}(\lambda)_i$. Figure 3.3 shows the distribution (across 100 trn-tst partitions) of the number of support points (n_{sup}) for $\hat{\lambda}_{cv}$ for each of the environments

of the Wheat-large data set. At $\hat{\lambda}_{cv}$, n_{sup} ranges from 30 to $\approx 5,000$. In 3 of the environments (*B2l*, *MELGAS*, and *LHT*) the average number of support points was $n_{sup} \approx 450$, that is $\sim 15 - 20\%$ of the size of the training set. In environment *B5l*, the proportion of active training support points was $\sim 5 - 10\%$. On the other hand, in environment *EHT* predictions relied on an average of $n_{sup} \approx 178$ (of $1,428$) individuals from training (Figure 3.3). Similar patterns were also observed in the Wheat-599 data (Supplementary Figure B.3); for instance, testing phenotypes from environment 1 were optimally predicted in average with $n_{sup} \approx 78$ (of 419); however, the relative sparsity (n_{sup}/n_{trn}) was smaller in the Wheat-large data set (5-17%) than in the Wheat-599 data set (12-60%).

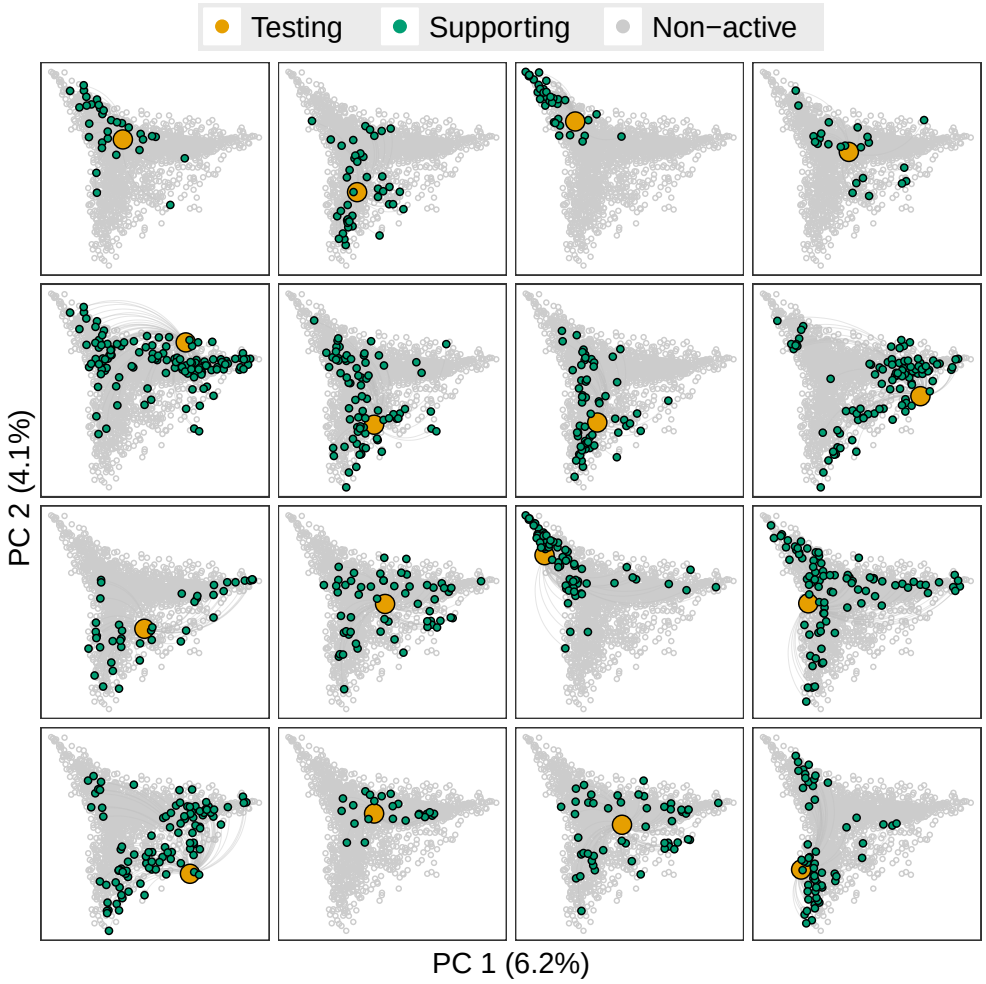

Figure 3.4: First two principal components coordinates for prediction points (yellow) and the corresponding support points (green). Grey points represent genotypes that did not contribute to the prediction of the genetic value of the genotype in yellow. All panels represent solutions for the environment *EHT*, Wheat-large data set.

Figure 3.4 shows (for selected testing genotypes) the coordinates on the 1^{st} and 2^{nd} PC of both the prediction point (yellow circle) and the training genotypes. Active training genotypes are represented in a green circle, and those non-active (i.e., with zero weight in the index) are represented in grey. In some cases, the support set includes training genotypes that are nearby (according to the coordinates on the 1^{st} 2 PCs) the prediction point. However, in other cases, the support set spanned outside the "neighborhood" of where the prediction point resides. This suggests that the SSI does not necessarily choose training points within the clusters. A similar plot for the Wheat-599 data set is presented in Supplementary Figure B.4.

3.4.4 Genomic relationships and weights in standard and sparse selection indices

Figure 3.5A shows the coefficients of the G-BLUP and of the SSI (i.e., the β_{ij}'s derived from Equations (3.2) and (3.3), respectively) versus the genomic relationship (g_{ij}, the ij entry of $\mathbf{G}$). In Figure 3.5A, the β_{ij}'s were derived for a training-testing partition with fixed heritability and λ chosen by CV conducted within the training set, for environment EHT from the Wheat-large data set. The weights used by the G-BLUP are, as expected, all different from zero and are positively associated with the genomic relationships (i.e., on average, training genotypes closely related to genotypes in the prediction set receive higher weight on the index). However, the points do not fall over a perfect line because the weight given to each of the training points depends not only on the relationship between the training point and the prediction point but also on the relationships among training points. On the other hand, as expected, the SSI zero-outs most of the weights. Interestingly, the SSI seems to zero-out most of the weights that are in the top left and lower-right quadrants (i.e., points that had a negative (positive) relationship and in the G-BLUP got positive (negative) weight, compare both plots in Figure 3.5A). Panel B in Figure 3.5 shows the proportion of coefficients that are zeroed-out by level of genomic relationship. Most of the coefficients corresponding to training genotypes with relationships with prediction points between -0.1 and 0.1 are zeroed-out; the proportion of coefficients that are zeroed decreases rapidly as g_{ij} increases; however, the decrease seems to be faster for the Wheat-large data set than for the Wheat-599 (Supplementary

Figure B.6). Interestingly, the proportion of coefficients zeroed-out also decreases for 'negative' genomic relationships, suggesting that the SSI does not use a 'local' support set; instead, the SSI seems to use informative support points. At least in the context of a 'linear' kernel as the one used here, a negative prior correlation (i.e., $g_{ij} < 0$) can be informative. The patterns observed in other environments of the Wheat-large data set and in the four environments of the Wheat-599 data set were conceptually similar to the ones presented in Figure 3.5 (see Supplementary Figures B.5 and B.6).

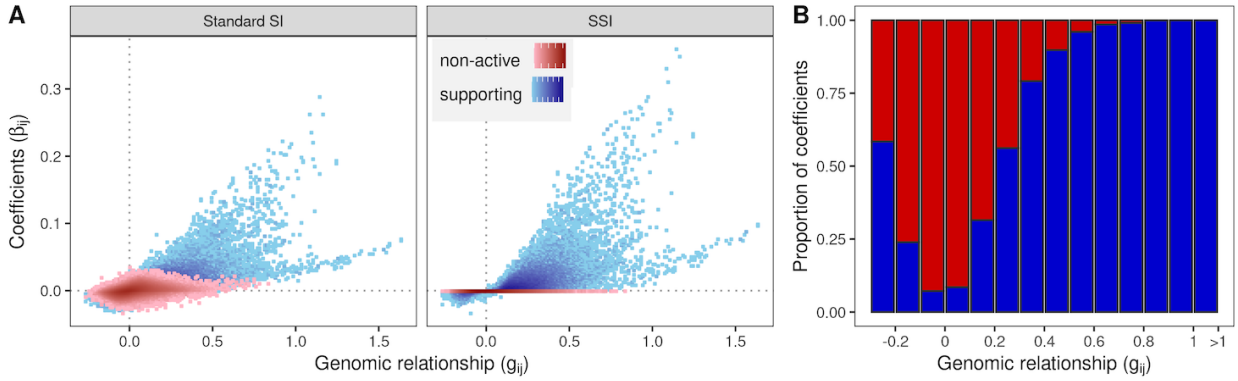

Figure 3.5: (A) Weights (β_{ij}) of a standard SI (G-BLUP) and of the optimal sparse selection index (SSI) versus the genomic relationship (g_{ij}). (B) Proportion of weights in the SSI that were zero (non-active) and non-zero (support points); Wheat-large data set, environment *EHT*.

3.5 Discussion

Sample size has been recognized as one of the main factors limiting prediction accuracy in genomic prediction (Lorenzana & Bernardo, 2009; de los Campos et al., 2013a; Habier et al., 2013). In un-structured populations, SNP effects can be assumed to be homogeneous and, therefore, genomic prediction accuracy increases with sample size (e.g., Daetwyler et al., 2008; de los Campos et al., 2013a). However, this is not necessarily the case in structured and admixed populations, in multi-family data (e.g., data from bi-parental families), or in multi-generation data. In those cases, differences in allele frequencies and in LD-patterns across may make SNP effects heterogenous across subgroups in the sample. In that context, a larger training data set may not translate into

higher prediction accuracy. This phenomenon has been recognized in both plant and animal breeding, as well as in complex trait prediction in humans.

For example, using data from a broiler breeding population, Wolc et al. (2016) showed that using training sets that included data from many generations led to slightly lower prediction accuracy than the one achieved when models were trained with data from just the last 3 generations. Likewise, Hayes et al. (2009) showed that the prediction accuracy for Holstein cattle was not improved by adding to the training set data from Jersey cattle. In plant breeding, using data from bi-parental families, Jacobson et al. (2014) reported that within family prediction accuracy could be increased by training models using only data from families that share at least one of the parents. Finally, in the context of human data, de los Campos et al. (2013b) noted that the accuracy of SNP-derived genomic relationships could be very low for distantly related individuals. Thus, combining family data with large volumes of data from distantly related individuals may not improve (or may even reduce) prediction accuracy relative to models trained with family data only.

Thus, when data originates from heterogeneous sources there may be trade-offs between sample size and the possibility of having a homogenous data set in which SNP can be conceived as homogenous within the training data and between training and testing sets. The recognition that in genomic prediction 'bigger is not always better' led to the development of several models and model-training strategies aiming to improve prediction accuracy. One line of research attempts to model effect heterogeneity using group-specific effects (e.g., Veturi et al., 2019; Rio et al., 2020). However, this approach is useful when individuals cluster in a small number (e.g., 2 or 3) of well-defined clusters, and becomes less useful and difficult to apply when data is characterized by either a large number of groups (e.g., bi-parental families) or when groups overlap in cryptic manners (e.g., admixed populations or partially-overlapping-multi-generation data). Another line of research seeks to identify an "optimal training set" by either selecting data from individuals that are closely related to the prediction set (e.g., Rincent et al., 2012; Jacobson et al., 2014; Wolc et al., 2016) or by using more sophisticated optimization algorithms (e.g., Akdemir et al., 2015). However, these approaches assume that it is possible to build a training set that is optimal for all

the genotypes in the prediction set. Our approach differs from these ones in that we developed a methodology that builds subject-specific training sets. Indeed, the SSI selects, for each individual in the prediction set, a custom training set (or support points) from which predictions are derived. Our approach builds on selection index methodology by adding an L1- (sparsity-inducing-) penalty into the optimization problem. The result is an index in which only a subset of the training points contributes to prediction accuracy.

When the training data consists of disconnected families, pedigree BLUP equations can also be sparse. However, this is not the case of the G-BLUP because genomic relationship matrices are dense. The SSI brings back sparsity into genomic prediction. The level of sparsity is largely controlled by the penalization parameter (λ). This parameter can be tuned using cross-validation within the training data. As with any other parameter, the value of λ that maximizes accuracy may change slightly between trn-tst partitions; however, in our experience, using a few (e.g., 10) training-testing partitions are enough to obtain an accurate estimate of the value of the regularization parameter that maximizes accuracy.

As noted, an SSI identifies, for each individual in the prediction set, a network of genotypes in the training data set (see Figure 3.4 and Supplementary Figure B.4) that contribute to the prediction. At first glance, this appears similar to the approach used in a k-nearest neighbor (KNN) regression (Cover & Hart, 1967). In KNN, the k genetically closest individuals (neighbors) predict each selection candidate, and predictions are derived using an average of the phenotypes in the neighborhood. There are important differences between the KNN and the SSI. First, the KNN uses only marginal similarities/distances between a prediction point and the points in the training data to build a 'neighborhood'; the SSI, however, also considers the correlations (i.e., redundancies) between points within the training data. As a consequence, the optimal support set of the SSI may include some distantly related individuals (see Figure 3.4 and Supplementary Figure B.4). Second, while in the standard KNN predictions are simply the arithmetic mean of the phenotypes in the neighborhood, in the SSI each training point contributes differently with weights (the β_{ij}'s) that reflect both the correlation of the training point with the prediction point as well as correlations

among points in the training set.

Best Linear Unbiased Prediction (BLUP) methods are equivalent to L2-penalized regressions. In BLUP, shrinkage is controlled by the noise and signal variances ($\lambda_0 = \sigma^2_\varepsilon/\sigma^2_u$, see Equation (3.2). We added to the optimization problem an L1-penality; thus, the SSI uses both L1 and L2 (which is intrinsically built in the SI) penalties. Therefore, the SSI can be seen as being a type of Elastic-Net (Zou & Hastie, 2005) regression. However, in the SSI the weight on the L2-penalty is determined by the ratio of variance components ($\lambda_0 = \sigma^2_\varepsilon/\sigma^2_u$) which may or may not be an optimal choice from a prediction perspective (particularly if the underlying assumptions of the BLUP method, e.g., homogeneity of effects, do not hold). Therefore, to add flexibility to the SSI we considered explicitly adding L1- and L2-penalties, and searching for an optimal combination, using cross-validation, of the relative weights of the penalization parameters of the Elastic-Net (α and λ) optimization problem:

$$\tilde{\boldsymbol{\beta}}(\alpha,\lambda)_i = \arg\min_{\beta_i}\left[\frac{1}{2}\boldsymbol{\beta}_i'(\mathbf{G}+\lambda_0\mathbf{I})\boldsymbol{\beta}_i - \mathbf{G}_i'\boldsymbol{\beta} + \lambda\frac{1}{2}(1-\alpha)\sum_{j=1}^{n}\beta_{ij}^2 + \lambda\alpha\sum_{j=1}^{n}|\beta_{ij}|\right]$$

To avoid too-much penalization, we decreased the weight of the initial L2-penalty to $0.5\lambda_0$. We found that this practice could increase prediction accuracy by a small factor (2-3.5%, see Supplementary Table B.3 for the Wheat-large data set) relative to the original SSI method (Equation (3.3)). However, this practice did not provide any advantage over the original SSI in the Wheat-599 data set (see Supplementary Table B.4).

In conclusion, we presented a novel prediction method that combines in a single framework, selection index methodology with sparsity-inducing methods. The resulting SSI identifies optimal training sets for each point in the prediction set. The method can be useful for multiple applications, including the use in genomic prediction of data from structured populations, bi-parental families, and the analyses of very large multi-generation data sets.

3.6 Acknowledgments

We are grateful to CIMMYT's Global Wheat Program that provided both the experimental field and marker data used in this study. M.L.C. was supported by the Monsanto's Beachell-Borlaug International Scholarship Program (MBBISP) and by the Dissertation Completion Fellowship funded by the Michigan State University Graduate School.

CHAPTER 4

GENOMIC PREDICTION IN MULTI-GENERATIONAL MAIZE HYBRIDS USING SPARSE KERNEL MODELS

Marco Lopez-Cruz[1], Yoseph Beyene[2], Manje Gowda[2], Jose Crossa[3],

and Gustavo de los Campos[4,5,6]

[1]Department of Plant, Soil and Microbial Sciences, Michigan State University, USA

[2]Global Maize Program, International Maize and Wheat Improvement Center (CIMMYT), Kenya

[3]Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Mexico

[4]Department of Epidemiology and Biostatistics, Michigan State University, USA

[5]Institute for Quantitative Health Science and Engineering, Michigan State University, USA

[6]Department of Statistics and Probability, Michigan State University, USA

4.1 Abstract

There has been much interest in the use of large historical data to calibrate more accurate prediction models to improve current breeding programs. However, multi-generational data often comes with increased heterogeneity that might include complex admixture patterns, in which genetic relationships are reduced as generations advance. It has been recognized that differences in heterogeneity patterns between the training and the prediction set, or including in the training set individuals that are distantly related to the prediction set can reduce the prediction accuracy. Most of the research in this sense focuses on designing optimal training sets that include only a few previous generations or a group of genotypes that are more closely related to the current prediction set. However, some training individuals can be optimal for some, but not all individuals in the prediction set. The sparse selection index (SSI) determines, for each individual in the prediction set, a customized optimal training set. Using additive genomic relationships, the SSI can provide an increased accuracy relative to the standard G-BLUP. Models with Gaussian kernels (K-BLUP) have been shown to yield a higher accuracy by maximizing the covariance between closely related genotypes. We studied whether the SSI using Gaussian kernels can provide increased accuracies. Using a three-generation doubled haploid maize data set from the International Maize and Wheat Improvement Center (CIMMYT), we show that the standard K-BLUP outperformed the G-BLUP. Also, we found that using an SSI with additive genomic relationships (sparse G-BLUP) leads to gains in accuracy between 5%-20%, relative to the standard G-BLUP. For the K-BLUP, the gains obtained by adding sparsity were smaller and not always significant.

4.2 Introduction

Almost two decades have passed since Genomic Selection (GS) was first proposed by (Meuwissen et al., 2001). This groundbreaking idea was quickly adopted for breeding dairy cattle (Hayes et al., 2009), beef cattle (Garrick, 2011), broilers (Wolc et al., 2016), maize (Bernardo & Yu, 2007), wheat (Poland et al., 2012), and many other animal species and crops (Xu et al., 2020). Over time,

investments by public and private organizations led to the development of large genomic data sets comprising DNA-sequences and phenotypes. The large sample size of modern genomic data sets has increased our ability to train high-dimensional genomic prediction equations accurately.

However, the larger sample size often comes with an increased genetic heterogeneity, including many generations of data and often complex admixture patterns. Moreover, there have been some signs that in genomic prediction, 'bigger may not always be better'. For example, Wolc et al. (2016) reported that the accuracy of genomic prediction in a broiler breeding program was higher when using data from the last three generations relative to prediction equations trained using data from the last five generations. Likewise, Riedelsheimer et al. (2013) and Jacobson et al. (2014) reported that the prediction accuracy was higher when models were trained using data from biparental families that shared at least one parent relative to training using data from all the available biparental families.

Early work by Habier et al. (2010) showed that family relationships have an important impact on prediction accuracy, and many studies have demonstrated that distantly related individuals make a small (sometimes negligible) contribution to the prediction accuracy. However, as noted above, some evidence suggests that using training sets formed by individuals distantly related to the genotypes of the prediction set may actually have a negative impact on the prediction accuracy (e.g., Lorenz & Smith, 2015). This may happen if, for example, heterogeneity in allele frequency and in linkage disequilibrium (LD) patterns between the training and prediction set lead to SNP-effect heterogeneity.

Issues related to data- and effect-heterogeneity have spawned multiple research efforts. One line of research models effect-heterogeneity explicitly using SNP-by-group interaction models or multivariate models ('multi-breed genomic prediction') in which effects are assumed to be correlated among groups (e.g., Olson et al., 2012; Lehermeier et al., 2015; Rio et al., 2020). This approach has shown promise, but it is only adequate when genotypes can be clustered into clearly disjoint groups.

Another line of research seeks to increase prediction accuracy by optimal design of training

data sets. The methods proposed and used to identify an optimal training set span from simple threshold-based methods (e.g., Clark et al., 2012; Lorenz & Smith, 2015) to more sophisticated algorithms that seek to minimize prediction error variance and functions thereof (Rincent et al., 2012; Akdemir & Isidro-Sanchez, 2019; Roth et al., 2020). A main assumption of these training set optimization methods is that a single training set is optimal for all individuals in the prediction set. But this may not be the case if some genotypes in the training set can improve prediction accuracy for some of the candidates of selection and reduce it for others.

To address the limitations of existing methods, in Chapter 3 of this dissertation, we developed a prediction method (sparse selection index, SSI) that identifies, for each individual in the prediction set, a customized training set. Our approach integrates into the selection index methodology (Smith, 1936; Hazel, 1943), a sparsity-inducing penalty that leads to sparse selection indices.

In Chapter 3, we used the SSI methodology to predict grain yield in two wheat data sets. The application presented in that chapter used additive genomic relationships, and the results showed that the SSI outperformed the genomic-BLUP (G-BLUP; VanRaden, 2008) by 5-10% in the correlation scale.

Reproducing Kernel Hilbert Spaces (RKHS) regression has shown good predictive performance in many genomic applications (e.g., de los Campos et al., 2010; González-Camacho et al., 2012). The G-BLUP is a special case of RKHS regression in which a linear kernel (additive genomic relationships) is used to describe the genetic similarity between genotypes. However, several studies (e.g., Crossa et al., 2010; Morota & Gianola, 2014) have suggested that using non-linear kernels (e.g., Gaussian kernels) may lead to a higher genomic prediction accuracy. In a Gaussian kernel, the covariance between genetic values is higher for closely related individuals and drops as two genotypes become increasingly distant. The rate at which the prior covariance between genetic values drops is controlled by a bandwidth parameter. Large bandwidth parameter values (that lead to highly local covariances) can be used to derive predictions which are largely dependent on closely related individuals. Thus, there is a clear link between RKHS with Gaussian kernels and the SSI methodology presented in Chapter 3. However, the Gaussian kernel does not yield strictly

sparse prediction equations.

Therefore, to further advance our research in the use of SSIs, in this chapter, we study whether the SSI can also improve the prediction performance of RKHS regressions with non-linear kernels. In this chapter we evaluate the performance of the SSI using additive (sparse G-BLUP) and non-additive (sparse K-BLUP) kernels using a three-generation DH (doubled haploid) maize data set from the International Maize and Wheat Improvement Center (CIMMYT). For several scenarios of training set composition, we show that the standard RKHS regression with a Gaussian kernel outperformed the additive G-BLUP. In agreement with what we report in Chapter 3, we found that the use of an SSI with additive genomic relationships (sparse G-BLUP) leads to gains in prediction accuracy between 5%-20%, relative to the standard (non-sparse) G-BLUP. For the K-BLUP, the gains obtained by adding sparsity were smaller and not always significant.

4.3 Materials and Methods

4.3.1 Genotypes and phenotypic data

The genotypes used in the study consist of 3068 DH lines derived from 54 biparental families. The DH lines were developed in 2017, 2018, and 2019 at CIMMYT's Maize DH facility at the Agricultural & Livestock Research Organization (KALRO) in Kiboko, Kenya. The biparental families were obtained by crossing elite inbred lines with drought-tolerant lines. Seeds from each of the families were collected and submitted for DH induction. The 3068 DH lines were selected from a larger population for stage I multi-location yield trials in 2017-2019, based on the results of evaluating germination, good stand establishment, plant type, low ear placement, and well-filled ears.

Each year, the selected DH lines were crossed with a single-cross tester from the complementary heterotic group and evaluated under well-watered (denoted as OPT) and drought (denoted as DRT) environmental conditions. The number of hybrids (trials) planted in 2017, 2018, and 2019 were 923 (14), 1423 (34), and 722 (17), respectively; trials were connected by a common check and three

to six commercial checks, planted in an alpha-lattice design with two replications, and evaluated in two well-watered locations and one managed drought stress locations during the 2017, 2018, and 2019 growing seasons. The OPT experiments were conducted during the rainy season, applying supplemental irrigation as needed. The DRT experiments were conducted during the dry (rain-free) season and irrigation was suspended 2 weeks before flowering and until harvest. Entries were planted in two-row plots, 5 m long, with 0.75 m spacing between rows and 0.25 m between hills. Two seeds per hill were initially planted and then, three weeks after emergence, one plant per hill was maintained to obtain a final plant density of 53333 plants/ha. Fertilizers were applied at the rate of 60 kg N and 60 kg P_2O_5/ha, as recommended for the area. Nitrogen was applied twice: at planting and 6 weeks after emergence. Fields were kept free of weeds by hand weeding. Grain yield (GY, tons ha^{-1}), anthesis date (AD, days), plant height (PH, cm) traits were recorded. Plots were manually harvested and GY was corrected to a 12.5% moisture. AD was measured from planting to when 50% of the plants shed pollen, and PH was measured from the soil surface to the flag leaf collar on five representative plants within each plot.

Leaf samples were taken from each of the 3068 DH lines and sent to Intertek, Sweden, for DNA extraction. The DNA sample plates were forwarded to the Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA, for genotyping with repetitive sequences (rAmpSeq) as described by Buckler et al. (2016). A total of 5465 markers coded as 0 (absence) and 1 (presence) were filtered by minor allele frequency (MAF<0.05) from which only 5173 were kept for analyses.

Further information about the 2017 and 2018 data can be found in Beyene et al. (2019) and Atanda et al. (2020), who have previously described and analyzed these data sets.

4.3.2 Phenotypes pre-processing

Adjusted means of GY, PH, AD were obtained using mixed effects model fitted separately for each trait-year-environmental-condition combination. The Best Linear Unbiased Estimates (BLUE) of genotypes across locations for the OPT experiments were estimated using the META-R software

(Alvarado et al., 2020) following the linear mixed:

$$Y_{ijkl} = \mu + g_i + L_j + R_{k(j)} + B_{l(kj)} + (g \times L)_{ij} + e_{ijkl}$$

where Y_{ijkl} is the phenotypic record of genotype i at location j in replicate k within the block l, μ is the overall mean, L_j is the fixed effect of the location j, $R_{k(j)}$ is the fixed effect of the replicate k within location j, $B_{l(kj)}$ is the random effect of the incomplete block l within replicate k and location j assumed to be independently and identically distributed (iid) normal with mean zero and variance σ_b^2, g_i is the fixed effect of genotype i, $(g \times L)_{ij}$ is the fixed effect of the genotype $\times$ location interaction, and e_{ijkl} is the random error assumed to be iid normal with mean zero and variance σ_e^2. After fitting the model just described, adjusted phenotypes (y_i) were obtained by subtracting from phenotypic records (GY, PH, and AD), the estimated effects of location, replicate, incomplete block, genotype $\times$ location interaction, and error. Likewise, within each year, the BLUE for each trait for the single-location DRT experiment was obtained through the linear model

$$Y_{ikl} = \mu + g_i + R_k + B_{l(k)} + e_{ikl}$$

where R_k is the fixed effect of the replicate k, $B_{l(k)}$ is the random effect of the incomplete block l within replicate k assumed to be iid normal with mean zero and variance σ_b^2, and the remaining factors are as before. The adjusted phenotypes were obtained by subtracting from the phenotypic records, the estimated effects of replicate, incomplete block, and error.

After phenotypes pre-processing, a total of $n = 3039$ lines containing marker information and that were observed in all environments for all traits were kept for GS models. The final number of lines in each year is as follows: $n_1 = 901$ lines in 2017, $n_2 = 1417$ in 2018, and $n_3 = 721$ in 2019.

4.3.3 Genomic selection models

We considered four different models: genomic-BLUP (G-BLUP) using additive genomic relationships (VanRaden, 2008), Reproducing Kernel Hilbert Spaces (RKHS) regression (Gianola et al., 2006) which is equivalent to a G-BLUP with a non-linear kernel, and sparse selection indices (SSI)

obtained by imposing an L1-penalty on the G-BLUP (sparse G-BLUP) and on the RKHS (sparse K-BLUP). In what follows we describe each of these models; for simplicity, since all phenotypes were centered, we present models without intercept nor fixed effects.

G-BLUP: In this model, the data-equation takes the form

$$\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{\varepsilon} \tag{4.1}$$

where $\boldsymbol{y} = (y_1, ..., y_n)'$, $\boldsymbol{u} = (u_1, ..., u_n)'$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$ are the vectors of adjusted phenotypes, breeding values (BV), and environmental error terms, respectively. Breeding values and errors are assumed to be normally distributed $\boldsymbol{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{G})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$, where σ_u^2 and σ_ε^2 are the genetic and error variances, $\mathbf{G}$ is the additive genetic relationship matrix (GRM), and $\mathbf{I}$ is an identity matrix.

The genomic relationship matrix $\mathbf{G} = \{G_{ij}\}$ was derived from markers, $\mathbf{X} = \{x_{im}\}$, using $\mathbf{G} = \mathbf{ZZ}'/p$, where $p = 5173$ is the number of markers and $\mathbf{Z} = \{(x_{im} - \overline{x}_m)/sd_{x_m}\}$ is the matrix of centered and scaled markers obtained by subtracting from each marker entry the mean of the corresponding column followed by scaling by the standard deviation of the column. The resulting matrix has an average of the diagonal elements equal to one.

The predicted BVs ($\hat{\boldsymbol{u}}_{PS}$) for the individuals in the prediction set (PS) are then linear combinations of the phenotypes ($\boldsymbol{y}_{TS}$) of the subjects in the training set (TS), this is (e.g., Searle et al., 1992)

$$\hat{\boldsymbol{u}}_{PS} = \mathbf{B}_G\boldsymbol{y}_{TS} \tag{4.2}$$

where $\mathbf{B}_G = \mathbf{G}_{PS,TS}(\mathbf{G}_{TS} + \lambda_0\mathbf{I})^{-1}$ is a Hat matrix (i.e., coefficients of regression of BVs on phenotypes), $\lambda_0 = \sigma_\varepsilon^2/\sigma_u^2$ is the ratio between residual and genetic variances, $\mathbf{G}_{PS,TS}$ is a matrix containing the additive genetic relationships between the data points in the prediction set with those in the training set, and $\mathbf{G}_{TS}$ represents the additive GRM of the training data points.

RKHS regression: In a RKHS regression, the vector of genomic predictions ($\boldsymbol{u}$) in Equation (4.1) are obtained as linear combinations on kernel evaluations as $\boldsymbol{u} = \mathbf{K}\boldsymbol{\alpha}$, where $\mathbf{K}$ is an $n \times n$ matrix of kernel evaluations on markers genotypes, and $\boldsymbol{\alpha}$ is a vector of regression coefficients. As

shown in de los Campos et al. (2009a), if the vector $\boldsymbol{\alpha}$ is assumed to be distributed $\boldsymbol{\alpha} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{K}^{-1})$, it follows that

$$\boldsymbol{u} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{K}). \tag{4.3}$$

Therefore, the vector of genomic predictions for the individuals in the prediction set are

$$\hat{\boldsymbol{u}}_{PS} = \mathbf{B}_K \boldsymbol{y}_{TS} \tag{4.4}$$

where the Hat matrix is now $\mathbf{B}_K = \mathbf{K}_{PS,TS}(\mathbf{K}_{TS} + \lambda_0 \mathbf{I})^{-1}$, $\lambda_0 = \sigma_\varepsilon^2/\sigma_a^2$, $\mathbf{K}_{PS,TS}$ is the matrix containing the evaluation of the Gaussian kernel for pairs of training-prediction genotypes, and $\mathbf{K}_{TS}$ is the kernel GRM of the training data.

Finding the solution of an RKHS model is therefore equivalent to finding the solution for the G-BLUP model but using $\mathbf{K}$ instead of the additive matrix $\mathbf{G}$. In this sense, we refer to the RKHS regression as to K-BLUP model.

We considered K-BLUP models using Gaussian kernel matrices $\mathbf{K} = \{K_{ij}\}$, $i, j = 1, ..., n$, given by $K_{ij}(\theta) = \exp(-\theta \tilde{d}_{ij}^2)$, where θ is a bandwidth parameter and $\tilde{d}_{ij}^2$ is the scaled squared Euclidean distance between individuals i and j given by their markers genotypes, obtained by dividing the distance $d_{ij}^2 = \sum_{m=1}^{p}(x_{im} - x_{jm})^2$ by the average distance $\overline{d} = \frac{1}{n^2}\sum_i \sum_j d_{ij}^2$. Three extreme Gaussian kernels were used as defined by González-Camacho et al. (2012), namely $\mathbf{K}_1 = \{K_{ij}(\theta_1)\}$, $\mathbf{K}_2 = \{K_{ij}(\theta_2)\}$, and $\mathbf{K}_3 = \{K_{ij}(\theta_3)\}$, where $\theta_1 = 0.2$, $\theta_2 = 1$, and $\theta_3 = 5$. See Supplementary Figure C.1 for pairwise comparisons of kernel (K_{ij}) versus additive (G_{ij}) genomic relationships.

In addition, kernel averaging (KA) was also implemented as described in de los Campos et al. (2010) using the three kernels $\mathbf{K}_k$, $k = 1, 2, 3$. Briefly, the three kernels are considered to jointly contribute to the prediction as $\boldsymbol{u} = \sum_{k=1}^{3} \mathbf{K}_k \boldsymbol{\alpha}_k$, where each summand has its own distribution (as in Equation (4.3)) as $\mathbf{K}_k \boldsymbol{\alpha}_k = \boldsymbol{u}_k \sim N(\mathbf{0}, \sigma_{a_k}^2 \mathbf{K}_k)$. This KA-BLUP model is fitted in a Bayesian fashion; however, it can be rewritten as a single random effect (as in Equation (4.1)) by making $\boldsymbol{u} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{K}_A)$, where $\sigma_a^2 = \sigma_{a_1}^2 + \sigma_{a_2}^2 + \sigma_{a_3}^2$ is the total kernel variance and $\mathbf{K}_A$ is an average kernel GRM given by

$$\mathbf{K}_A = \frac{\sigma_{a_1}^2}{\sigma_a^2}\mathbf{K}_1 + \frac{\sigma_{a_2}^2}{\sigma_a^2}\mathbf{K}_2 + \frac{\sigma_{a_3}^2}{\sigma_a^2}\mathbf{K}_3. \tag{4.5}$$

Sparse BLUP models: Sparsity was incorporated into the G-BLUP and K-BLUP models as described in Chapter 3. Briefly, in a SSI, the weights of the selection index for the i^{th} individual in the prediction set was obtained from the following L1-penalized optimization problem

$$\tilde{\boldsymbol{b}}(\lambda)_i = \arg\min_{b_i} \left[\frac{1}{2} \boldsymbol{b}_i' (\mathbf{G}_{TS} + \lambda_0 \mathbf{I}) \boldsymbol{b}_i - \mathbf{G}_i' \boldsymbol{b}_i + \lambda \sum_{j=1}^{n} |b_{ij}| \right] \tag{4.6}$$

where $\mathbf{G}_i' = \mathbf{G}_{PS(i),TS}$ is the vector containing the additive relationships between the i^{th} subject in the prediction set and each of the subjects in the training set, λ is a parameter controlling the degree of sparsity of $\tilde{\boldsymbol{b}}(\lambda)_i$, and $\sum_{j=1}^{n} |b_{ij}|$ is a penalty on the L1-norm of $\boldsymbol{b}_i$. A (sparse) Hat matrix for the SSI, $\tilde{\mathbf{B}}(\lambda)_G$, contains in each row the solutions to Equation (4.6), obtained for each testing genotype. A value of $\lambda = 0$ yields the same (non-sparse) Hat matrix of the standard G-BLUP in Equation (4.2). For the sparse K-BLUP models we used Equation (4.6) with the Gaussian kernel (either $\mathbf{K}_1$, $\mathbf{K}_2$, $\mathbf{K}_3$, or $\mathbf{K}_A$) instead of additive relationship matrices ($\mathbf{G}$). Although optimization problem in Equation (4.6) does not have a closed-form, solutions for it can be derived using a coordinate descent algorithm (see Chapter 3 for further details). Finally, an optimal value of λ can be obtained using cross-validation within the training set (more details in Chapter 3).

4.3.4 Variance components

The implementation of G-BLUP, K-BLUP, and the corresponding sparse versions of these models require estimates of variance components. We obtained these estimates by fitting Bayesian genomic models within each trait-environment combination. These analyses were performed using the BGLR R-package, with the default setting for hyper-parameters (Perez & de los Campos, 2014). After fitting the models, posterior means of the variance components were obtained. For the standard KA-BLUP, the model was fitted with the three kernels together to estimate the kernel-specific variances and then used to derive σ_a^2 and the kernel $\mathbf{K}_A$ (Equation (4.5)). A heritability estimate for the G-BLUP model was derived as $h^2 = \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2)$. For the K-BLUP models, the heritability was obtained by replacing the estimate of σ_u^2 by the kernel genetic variance estimate σ_a^2. All these models were fitted using data from the training set only.

4.3.5 Assessment of prediction accuracy

Variance components estimates and the corresponding GRM ($\mathbf{G}$, $\mathbf{K}_1$, $\mathbf{K}_2$, $\mathbf{K}_3$, or $\mathbf{K}_A$) were used to derive the non-sparse ($\mathbf{B}_G$ or $\mathbf{B}_K$ for the standard BLUP) and the sparse ($\tilde{\mathbf{B}}(\lambda)_G = \{\tilde{\boldsymbol{b}}(\lambda)'_{i_G}\}$ or $\tilde{\mathbf{B}}(\lambda)_K = \{\tilde{\boldsymbol{b}}(\lambda)'_{i_K}\}$) Hat matrices. (Note that in the SSI, the rows of the sparse Hat matrix are simply the solutions to Equation (4.6), obtained for each testing genotype.) The predictions ($\hat{\boldsymbol{u}}_{PS}$) were derived (as in Equation (4.2) and (4.4)) as the product of the (non-sparse or sparse) Hat matrix times the vector of phenotypes in the training set. Prediction accuracy was measured as the correlation between observed and predicted values in the prediction set, i.e., $\rho = cor(\boldsymbol{y}_{PS}, \hat{\boldsymbol{u}}_{PS})$.

Prediction accuracy was assessed for different prediction scenarios using cycle 2019 as the prediction set with different training set compositions, as follows: (*i*) the data from the 2019 cycle was randomly partitioned into 85%-15% (i.e., 612 and 109 individuals), (*ii*) the 85%-set ($n_{PS} = 612$) from the year 2019 was predicted using data of the earlier generations 2017 ($n_{TS} = 901$), 2018 ($n_{TS} = 1417$), and 2017+2018 combined ($n_{TS} = 2318$) as training set, (*iii*) the prediction of the 612 individuals was also performed using the same training sets but augmented by progressively including the remaining 15%-set from 2019, first 37 (5%), then 73 (10%), and lastly 109 (15%) individuals. See Table 4.1 for a summary of all the training set compositions. All predictions were performed 100 times using different random partitions of the 2019 data.

4.3.6 Software

All the aforementioned analyses were performed in the R environment-language (R Core Team, 2019). All standard Bayesian G-BLUP and K-BLUP models were fitted using the BGLR R-package (Perez & de los Campos, 2014) to estimate variance components. The sparse Hat matrices ($\tilde{\mathbf{B}}(\lambda)_G$ or $\tilde{\mathbf{B}}(\lambda)_K$) were obtained with the SFSI R-package (Lopez-Cruz et al., 2020). For each trait-environment-partition, an optimal value of λ was obtained using 10-fold cross-validation within the training set.

Table 4.1: Training set (TS) composition used in each prediction scenario. (The prediction set was the same for all training scenarios and consisted of 612 randomly chosen individuals from 2019).

Data from previous years (n)	**% of 2019 data used for training (n)**	**Total training size (n_{TS})**
2017 (901)	0 (0)	901
	5 (37)	938
	10 (73)	978
	15 (109)	1010
2018 (1417)	0 (0)	1417
	5 (37)	1454
	10 (73)	1490
	15 (109)	1526
2017+2018 (2318)	0 (0)	2318
	5 (37)	2355
	10 (73)	2391
	15 (109)	2427

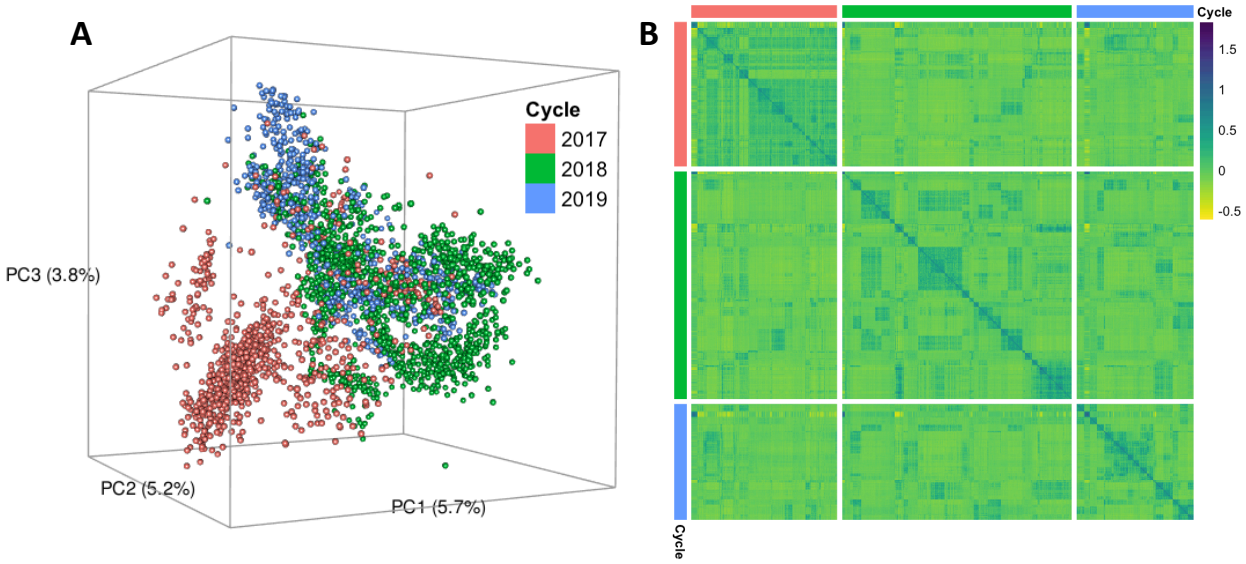

Figure 4.1: (A) First 3 principal components of the additive genomic relationships matrix, **G**. Points represent individuals that are color separated according to cycle (2017, 2018, or 2019). (B) Heatmap of the genomic relationships matrix.

4.4 Results

The germplasm used in this study is derived from different biparental families across years. This richness of the data is reflected in a high population heterogeneity in which individuals cluster into groups within and across generations (Figure 4.1). However, the crosses performed prevented the formation of a clear structure (e.g., 2 clusters); instead the population shows a more cryptic substructure with varying degrees of admixture between families. The intermixing between generations that is observed in Figure 4.1 can be attributed to allele sharing as only alleles from the selected elite parents in one generation are passed to the next generation.

4.4.1 Prediction accuracy comparison of G-BLUP and K-BLUP models

Figure 4.2 shows the accuracy of prediction (averaged across all 100 partitions) for GY-OPT using all standard BLUP models for all different training set compositions representing a combination of previous cycles (2017, 2018, or 2017+2018) plus the inclusion of 0, 5, 10, and 15% (i.e., 0, 37, 73, and 109 subjects) of the total individuals from the same 2019 cycle (see Table 4.1). As expected, the inclusion of individuals from the same cycle increases the prediction accuracy across all models and training set composition. For instance, using the 2017 cycle as training set, the accuracy of the G-BLUP was increased by 100% when adding 73 individuals; however, using 2018 as training set showed a 60% increase, and when using 2017+2018, a gain of 27% was observed.

Likewise, prediction accuracy was higher when combining data from 2017+2018 as training set compared with models using data from 2018 or 2017 alone for training (see top-left panel in Figure 4.2). However, the accuracy was not always increased when the training set is augmented to include 2017+2018 together (see bottom panels and top-right panel in Figure 4.2), and, in some cases, was even lowered (see Supplementary Figure C.3 for trait PH). Contrastingly, the prediction using individuals from 2017+2018 was sometimes equally performed (see the bottom-left panel in Figure 4.2) and in some cases outperformed (see Supplementary Figure C.2 for GY-DRT and Supplementary Figure C.3 for PH) when using only data from the 2017 cycle.

In general, kernel-based models achieved higher prediction accuracy than the standard G-BLUP. Although the kernels $\mathbf{K}_1$, $\mathbf{K}_2$, and $\mathbf{K}_3$ are ranked differently across training set compositions, models with $\mathbf{K}_A$ seems to be more stable across all scenarios performing similar to the best of the three kernels $\mathbf{K}_1$, $\mathbf{K}_2$, or $\mathbf{K}_3$. This result is in agreement with the findings in de los Campos et al. (2010).

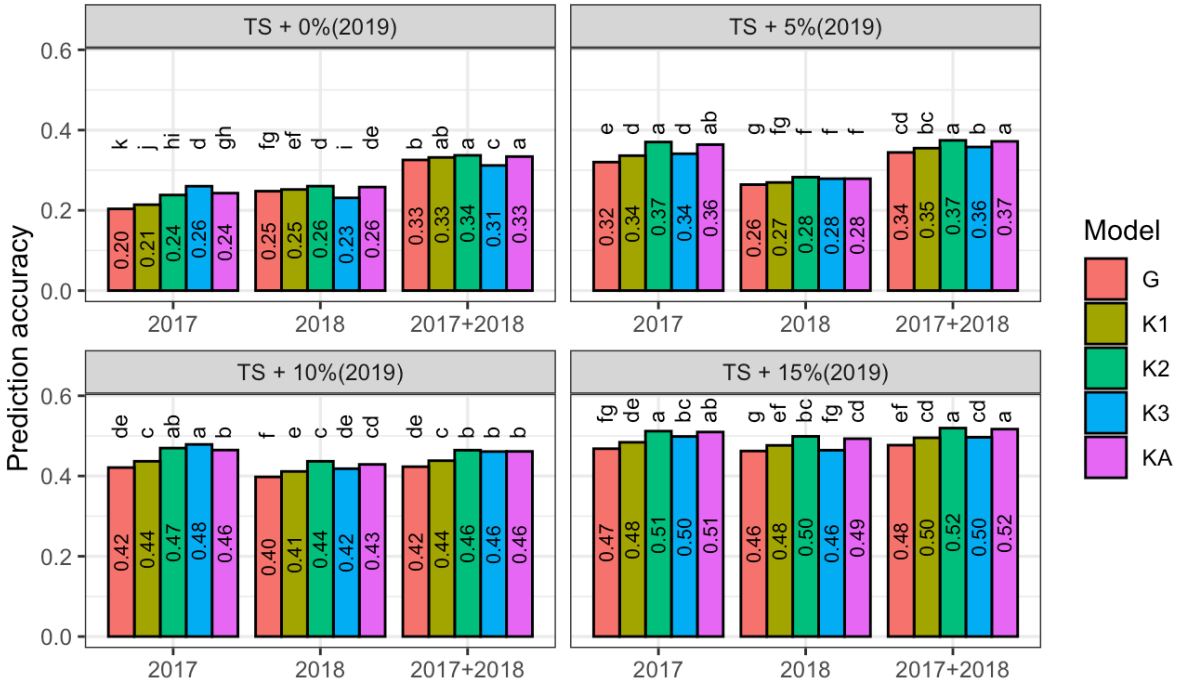

Figure 4.2: Prediction accuracy by model and training set (TS). TSs consisted on all the data from the 2017, 2018, or 2017+2018 cycles alone (top-left panel), or in combination with a proportion (5%, 10%, 15%) of the data from the 2019 cycle. The prediction set consisted of 612 genotypes from the 2019 cycle that were not used for model training. Models with the same letter within panel indicate no significant difference from each other ($\alpha = 0.05$, ANOVA followed by Tukey test). GY-OPT trait-environment combination.

4.4.2 Effect of sparsity on prediction accuracy

The same partitions of training-prediction sets used to obtain the results for the standard models were used to evaluate the prediction accuracy of SSIs (sparse models). A cross-validated value λ_{CV} was found within the training set to calculate an optimal sparse BLUP model. Table 4.2 contains the results of the predictions of the GY-OPT trait-environment combination for the scenario where 15% of data from 2019 is included in the training set (2017, 2018, or 2017+2018). Results for

the cases when adding 0%, 5%, and 10% of the 2019 data are presented in Supplementary Table C.1. With this training set composition, the accuracy of the standard G-BLUP models was between 0.46-0.48. K-BLUP (standard or sparse) and sparse G-BLUP models achieved higher prediction accuracy than the standard G-BLUP, with gains in accuracy (relative to G-BLUP) ranging from minimal (1%) to substantial (12%).

Table 4.2: Heritability and accuracy of prediction for each training set (TS) composition (including 15% of subjects from the 2019 cycle), GY-OPT trait-environment combination.

					Accuracy (SD)		% Gain	
TS (n_{TS})	**GRM**	$\lambda_{CV}{}^a$	n_{sup} **(RS)**^{b}	h^2	**Standard**	**Sparse**	**I**^{c}	**II**^{d}
2017 (1010)	G	0.0181	171 (17)	0.52	0.47 (0.031)	0.51 (0.030)	-	9
	K1	0.0037	188 (19)	0.87	0.48 (0.030)	0.50 (0.034)	3	4
	K2	0.0060	219 (22)	0.75	0.51 (0.028)	0.52 (0.027)	9	2
	K3	0.0000	1007 (100)	0.92	0.50 (0.029)	0.50 (0.029)	7	0
	KA	0.0041	240 (24)	0.84	0.51 (0.028)	0.52 (0.029)	9	2
2018 (1526)	G	0.0112	337 (22)	0.61	0.46 (0.026)	0.48 (0.028)	-	5
	K1	0.0016	480 (31)	0.91	0.48 (0.026)	0.49 (0.027)	3	3
	K2	0.0023	684 (45)	0.80	0.50 (0.026)	0.50 (0.027)	8	0
	K3	0.0000	1526 (100)	0.90	0.46 (0.029)	0.46 (0.029)	0	0
	KA	0.0015	683 (45)	0.87	0.49 (0.027)	0.49 (0.028)	7	0
2017+18 (2427)	G	0.0188	268 (11)	0.53	0.48 (0.025)	0.51 (0.027)	-	7
	K1	0.0033	350 (14)	0.88	0.50 (0.025)	0.51 (0.026)	4	2
	K2	0.0031	750 (31)	0.77	0.52 (0.025)	0.52 (0.026)	9	0
	K3	0.0000	2427 (100)	0.87	0.50 (0.027)	0.50 (0.027)	4	0
	KA	0.0020	922 (38)	0.83	0.52 (0.026)	0.52 (0.027)	8	0

GRM: Genetic relationship matrix. SD: standard deviation. ^{a}Penalization parameter in Equation (4.6) found by cross-validating the TS. $^b n_{sup}$=average number of individuals from the TS with a non-zero coefficient in the sparse Hat matrix (support set). RS: relative sparsity ($100 n_{TS}/n_{sup}$). In the standard models λ_{CV} is equal to zero and n_{sup} is equal to the total TS size. Within each TS cycle, percentage of gain in accuracy of the ^{c}standard K-BLUP relative to the standard G-BLUP, and ^{d}sparse *-BLUP relative to the standard *-BLUP (*=G- or K-).

The gains in prediction accuracy are more evident when the accuracy of the non-sparse G-BLUP models was lower (i.e., when fewer individuals from the 2019 cycle are included in the training set). For instance, when the accuracy of the G-BLUP is as low as 0.2 (the case where no data from 2019 is included in the training set), SSIs yielded gains in accuracy (relative to the standard G-BLUP) of up to 28% (Supplementary Table C.1). It was only in these low-accuracy situations

that, in some cases, the use of a standard K-BLUP model with $\mathbf{K}_3$ resulted in $\sim$ 6% loss of accuracy (relative to the standard G-BLUP) and that using sparse models, the accuracy lost was 3-10% (see Supplementary Table C.1). The advantage in accuracy of a sparse model over its standard version was more marked for the G-BLUP (i.e., when using additive relationships matrices, 23%); however, the same gains were smaller for the RKHS regressions using Gaussian kernels. ($\sim 2-8\%$ gain in accuracy, Supplementary Table C.1). No significant difference between sparse and non-sparse model was observed when using the large-bandwidth kernel $\mathbf{K}_3$. Similar results can be found for trait GY-DRT (see Supplementary Table C.2).

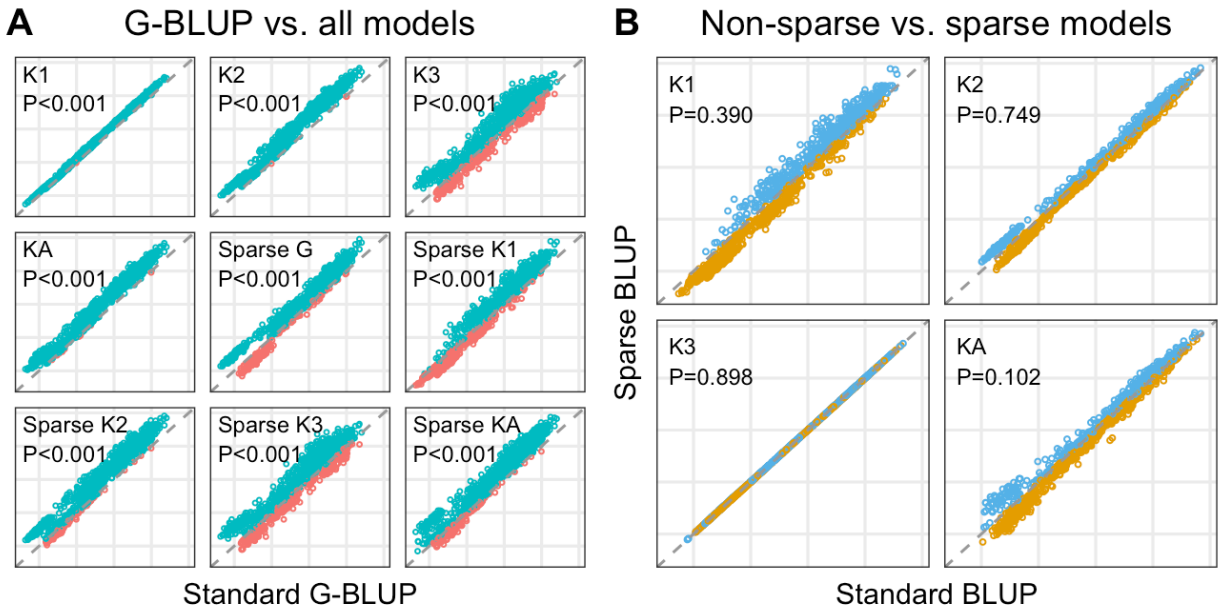

Figure 4.3: (A) Prediction accuracy of the standard (non-sparse) G-BLUP model (horizontal axis) versus the prediction accuracy of all other models (vertical axis of each panel). (B) Prediction accuracy of the standard *-BLUP model (horizontal axis) versus the prediction accuracy of its sparse version (vertical axis), by type of kernel used in panels. Each point represent a training-testing partition within each training set composition. Colored points above (below) the 45 degree line represent cases for which one model outperformed the other model. P: p-value for the test (from ANOVA) for differences in accuracy between the two models. Trait GY, environment OPT.

For the PH-OPT trait-environment combination, when models were trained using 15% of the 2019 data, the gains in accuracy obtained with the sparse G-BLUP, relative to the non-sparse G-BLUP, were as high as 11%, and up to 18% with a sparse K-BLUP model (Table 4.3). These gains in accuracy were very notable (> 100%) when adding to the training set 10% or fewer of

the individuals from the 2019 cycle (see Supplementary Table C.3). Results for the PH-DRT trait-environment are reported in Supplementary Table C.4 where similar patterns are observed.

Table 4.3: Heritability and accuracy of prediction for each training set (TS) composition (including 15% of subjects from the 2019 cycle), PH-OPT trait-environment combination.

					Accuracy (SD)		% Gain	
TS (n_{TS})	**GRM**	λ_{CV}[a]	n_{sup} **(RS)**[b]	h^2	**Standard**	**Sparse**	**I**[c]	**II**[d]
2017 (1010)	G	0.0112	506 (50)	0.56	0.49 (0.037)	0.51 (0.042)	0.0	3.4
	K1	0.0008	726 (72)	0.89	0.51 (0.036)	0.51 (0.039)	3.1	1.2
	K2	0.0007	773 (77)	0.77	0.54 (0.034)	0.54 (0.034)	9.1	0.1
	K3	0.0000	1010 (100)	0.92	0.51 (0.036)	0.50 (0.036)	2.4	-0.1
	KA	0.0003	862 (85)	0.87	0.53 (0.035)	0.53 (0.035)	7.3	-0.7
2018 (1526)	G	0.0161	208 (14)	0.71	0.50 (0.038)	0.56 (0.033)	0.0	10.9
	K1	0.0020	350 (23)	0.95	0.53 (0.036)	0.56 (0.035)	5.6	4.6
	K2	0.0021	627 (41)	0.88	0.56 (0.033)	0.57 (0.031)	11.3	1.4
	K3	0.0000	1526 (100)	0.92	0.53 (0.032)	0.53 (0.032)	5.4	0.0
	KA	0.0015	647 (42)	0.92	0.56 (0.033)	0.56 (0.033)	10.2	1.4
2017+18 (2427)	G	0.0132	356 (15)	0.64	0.47 (0.038)	0.52 (0.039)	0.0	10.8
	K1	0.0015	696 (29)	0.93	0.52 (0.036)	0.54 (0.037)	9.3	4.2
	K2	0.0007	1728 (71)	0.85	0.56 (0.034)	0.56 (0.034)	17.7	-0.2
	K3	0.0000	2427 (100)	0.92	0.51 (0.036)	0.51 (0.036)	8.3	0.1
	KA	0.0004	1896 (78)	0.89	0.55 (0.034)	0.55 (0.035)	16.4	-0.8

GRM: Genetic relationship matrix. SD: standard deviation. [a]Penalization parameter in Equation (4.6) found by cross-validating the TS. [b]n_{sup}: average number of individuals from the TS with a non-zero coefficient in the sparse Hat matrix (support set). RS: relative sparsity ($100n_{TS}/n_{sup}$). In the standard models λ_{CV} is equal to zero and n_{sup} is equal to the total TS size. Within each TS cycle, percentage of gain in accuracy of the [c]standard K-BLUP relative to the standard G-BLUP, and [d]sparse *-BLUP relative to the standard *-BLUP (*=G- or K-).

Across all scenarios, the standard G-BLUP showed the lowest accuracy among all models (SSI and standard K-BLUP, see Figure 4.3A for GY-OPT). This inferiority of the standard G-BLUP was also observed for GY-DRT and PH (see Supplementary Figure C.4A and Supplementary Figure C.5A). The addition of sparsity to the K-BLUP models resulted sometimes in an extra advantage in accuracy when using a kernel with a small bandwidth ($\mathbf{K}_1$ with $\theta = 0.2$, and $\mathbf{K}_2$ with $\theta = 1$) or averaged across extreme kernels ($\mathbf{K}_A$) for GY (Figure 4.3B and Supplementary Figure C.4B) and PH (Supplementary Figure C.5B).

4.4.3 Automatic training-sample selection

Tables 4.2 and 4.3 (and Supplementary Tables C.1-C.4) show the optimal value of the penalization parameter λ and the degree of sparsity of the resulting index, measured by the average number of subjects from the training set in the support set (n_{sup}, subjects with a non-zero coefficient in the estimated Hat matrix) of each predicted genotype. The degree of sparsity varied across models. For the GY-OPT trait-environment combination, across all training set compositions, the strongest sparsity was achieved using the genomic matrix G with a relative sparsity (n_{sup}/n_{TS}) of 11-33% (Table 4.2 and Supplementary Table C.1) while the relative sparsity with kernels increases as the bandwidth parameter θ increases (relative sparsity of 14-47% for $\mathbf{K}_1$ and 22-59% for $\mathbf{K}_2$). The relative sparsity achieved when using the $\mathbf{K}_A$ kernel (25-62%) was similar to that of the $\mathbf{K}_2$ kernel (see Table 4.2 and Supplementary Table C.1). The fact that no difference in accuracy was observed between standard and sparse $\mathbf{K}_3$-BLUP model is due that the optimal λ_{CV} was zero; thus, the sparse model was equivalent to the standard model.

Figure 4.4 displays a heatmap of the sparse Hat matrix ($\tilde{\mathbf{B}}(\lambda)_G$) of the sparse G-BLUP model. Individuals in the training set (2017+2018 plus 15% from 2019) appear in columns and those in the prediction set are shown in rows. Individuals from the training set that did not contribute to each index (i.e., those with zero weight in the index) are displayed in grey. Those with a non-zero coefficient (support set) are shown in a yellow-blue (logarithm) scale. The heatmap makes evident how SSIs select custom training sets for each genotype in the prediction set. Individual genotypes in the training set supports the prediction of some but not all the genotypes in the prediction set. The solution for the Hat matrix in Figure 4.4 is very sparse, with varying number of support points by testing genotype. Predictions of each of the 612 testing genotypes was performed using phenotypes from, on average, 268 (out of 2427 training genotypes, i.e., 11%, see Table 4.2) training genotypes. For the same prediction scenario, a heatmap for the sparse $\mathbf{K}_A$-BLUP model (showing a 38% of sparsity) is provided in Supplementary Figure C.6.

Figure 4.5 shows, for each of the sparse models, the proportion of the training individuals from each cycle (2017, 2018, or 2019) that contributed to the prediction (within cycle support set) of the

testing individuals. Each panel represents the different training sets composed of 2017+2018 data plus the addition of either 0%, 5%, 10%, or 15% of the 2019 data.

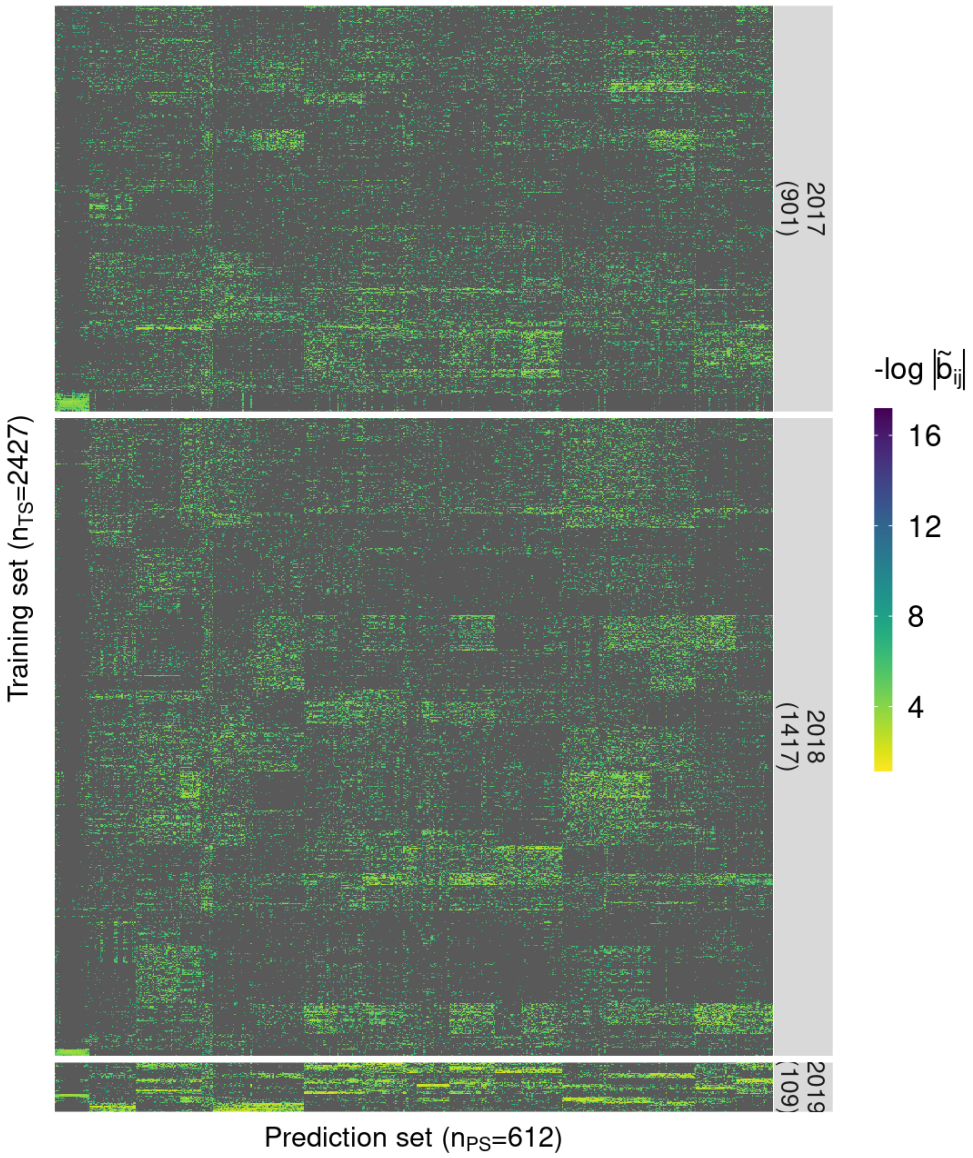

Figure 4.4: Heatmap of the coefficients in the Hat matrix ($\tilde{\mathbf{B}}(\lambda)_G$) of the sparse G-BLUP model for one training-prediction (TS-PS) partition in the prediction of $n_{PS} = 612$ individuals from 2019 using $n_{TS} = 2427$ individuals (2017+2018 plus 15% of the 2019 set). Predicted individuals are presented in columns and training individuals are presented in rows separated by cycle and number of individuals in parentheses. The value of λ was obtained by cross-validation. Each column represents values of the vector $\tilde{\mathbf{b}}(\lambda)_{i_G} = \{\tilde{b}_{ij}\}$, $j = 1, ..., 2427$ (Equation (4.6)). Individuals no contributing to the prediction have a coefficient $\tilde{b}_{ij} = 0$ represented in grey color. Individuals with a non-zero coefficient are shown in a yellow-blue logarithm scale (in the original scale, yellow indicates large values and blue indicates small value). GY-OPT trait-environment combination.

As expected, training individuals that belong to the same group as the testing individuals are more likely to be included in the support set. For example, using a sparse G-BLUP model trained with 2017+2018 plus 5% from the 2019 data (see the top-right panel in Figure 4.5), on average, 41% of the 2019 genotypes of the 37 included in the training set contributed to the prediction of the testing individuals. Although more abundant, a smaller portion of the total individuals from previous cycles (19% of the 901 subjects from 2017 and 18% of the 1417 from 2018) are also contributing to the prediction. With a smaller degree of sparsity, similar patterns were also observed for the sparse K-BLUP models (see Figure 4.5) except with $\mathbf{K}_3$ that did not render sparsity at all (not shown in the figure). Plots showing the within cycle sparsity patterns for GY-DRT and PH (OPT and DRT) are shown in Supplementary Figures C.7 and C.8.

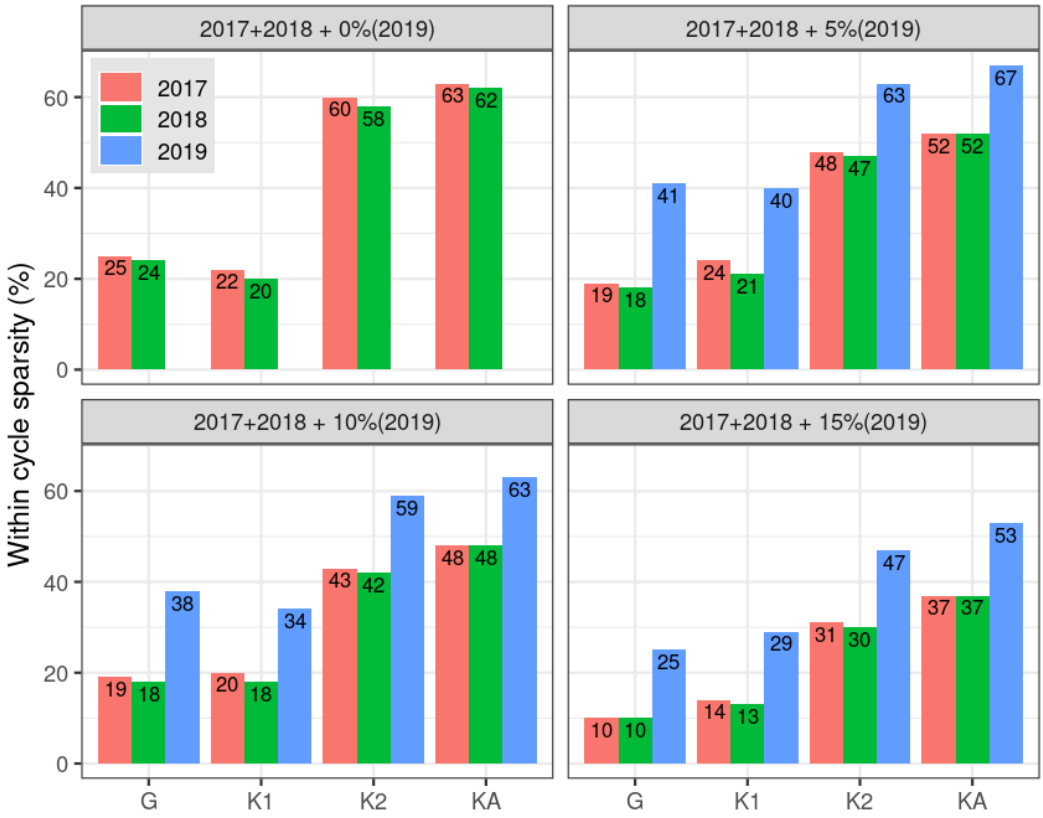

Figure 4.5: Proportion of the training individuals from each cycle that contributed to the prediction of the 612 testing genotypes from 2019, using sparse models with different relationship matrices (horizontal axis): $\mathbf{G}$, $\mathbf{K}_1$, $\mathbf{K}_2$, or $\mathbf{K}_A$. The training set was composed by individuals from 2017 (n = 901) and 2018 (n = 1417) alone (top-left panel) or in combination with a proportion (5%, 10%, 15%) of the data from the 2019 cycle. GY-OPT trait-environment combination.

As more individuals from the same cycle are added to the training set, fewer individuals from previous generations become less frequent in the support set. For instance, performing the prediction with a sparse G-BLUP using 2017+2018 data including 15% of the 2019 data (bottom-right panel in Figure 4.5), yielded a smaller support set with only 10% (90 of 901) of the 2017 data and 10% (142 of 1417) of the data from 2018.

4.5 Discussion

Multiple factors affect the predictive performance of GS models, including sample size, trait heritability, the extent of linkage disequilibrium (LD) between markers and quantitative trait loci (QTL), and the relationships between training and testing genotypes (Daetwyler et al., 2008; Heffner et al., 2009; Lorenzana & Bernardo, 2009; Combs & Bernardo, 2013).

General guidelines suggest that prediction accuracy is maximized when the training set includes a sufficient large number of individuals which are distantly related to each other (Rincent et al., 2012) and are closely related to the subjects in the prediction set (Habier et al., 2010; Clark et al., 2012). On the other hand, there is evidence suggesting that increasing the training set size by including individuals that are genetically distant to those in the prediction set does not necessarily increase and might even decrease the prediction accuracy (e.g., Lorenz & Smith, 2015).

Each cycle of a breeding program produces a new batch of genomic and phenotypic data; therefore, after many years of adopting GS, the data available for model training is typically multi-generational and may often include complex patterns of pedigree relationships within and between generations. There is clear evidence that a GS model needs to be re-trained every cycle (Wolc et al., 2011; Pszczola & Calus, 2016; Wientjes et al., 2013). When re-training models, breeding organizations face many challenges. Should all the available data be used for model training? Should instead one restrict the training data to only include genotype/phenotypes from recent generations? Should one exclude data from genotypes distantly related to the current set of candidates of selection?

Some evidence suggests that in genomic prediction, 'bigger is not necessarily better'. For

instance, using historic wheat data generated over 17 years, Dawson et al. (2013) observed that the accuracy of year-to-year predictions using training sets composed of all previous years was approximately the same as when considering only three years back. Likewise, in a broiler breeding population, Wolc et al. (2016) found that the maximum accuracy was accomplished when the training set was composed of the three most recent generations.

The SSI (sparse G-BLUP) methodology presented in Chapter 3 offers a framework to identify, for each individual in the prediction set, a customized training set (or support points) from which the predictions are derived. This methodology considers both the relationships between the candidate of selection and each training genotype as well as relationships among training genotypes (Equation (4.6)). Therefore, in this chapter, we propose that the sparse selection indices presented in Chapter 3 can be used to address the problem of training set optimization with multi-generation data. We used a multi-generation data originated from more than 50 biparental families to measure the impact of sparsity using SSIs formed using additive and non-additive kernels.

Our results confirmed that an SSI based on additive relationships yields a higher prediction accuracy than the standard additive G-BLUP. When a non-additive kernel was used, we found that sparsity improved prediction accuracy provided that the kernel used was not already a 'local' kernel, that is, a kernel in which genetic covariances are positive only for closely related individuals. For local kernels ($\mathbf{K}_3$), adding sparsity did not improve prediction accuracy in a clear and systematic way. Therefore, our results confirm the benefit of 'local predictions', which can be obtained either by using an RKHS with local kernels or with an SSI applied to additive genomic relationships.

Both the SSI and the Kernels regression require optimizing a parameter that controls how local predictions are. The SSI requires optimizing the penalization parameter (λ), which can be done by cross-validation within the training data set. On the other hand, the RKHS regression requires tuning the bandwidth parameter which controls how fast covariances drop with genetic distance. This can be done either by comparing multiple kernels using cross-validation or by using multiple kernels with 'kernel averaging' as discussed in de los Campos et al. (2010).

Standard training-optimization methods assume that a single training set is optimal for all

candidates of selection. The SSI does not make this assumption. Our results show clearly that each SSI picks a particular set of support points and that the optimal training set varies from genotype to genotype. The inspection of the Hat matrix of the SSI makes it clear that in prediction, *one-size-does-not-fit-all* candidates of selection. Likewise, the inspection of the Hat matrix shows that optimizing training sets by restricting the training data to recent generations may also not be optimal. Indeed, most of the SSIs picked information from all the generations available, with varying levels of sparsity.

In **conclusion**: SSIs can be used to optimize prediction accuracy when the training data exhibit complex relationship patterns. In this context, differences in allele frequencies and in LD-patters may make SNP effect heterogenous across families and sub-families, thus making the standard G-BLUP sub-optimal. Both local kernels and SSIs can be used to optimize prediction accuracy in such data sets.

CHAPTER 5

CONCLUDING REMARKS AND FUTURE DIRECTIONS

The core of this dissertation has a methodology point of view, presenting an innovative procedure that combines two well-established approaches: selection index and penalized regression. The former was developed for breeding value prediction using different sources of correlated information. The latter is commonly used in statistics and machine learning for variable selection to prevent overfitting in situations where there are more variables than observations. This novel approach, which we named sparse selection index (SSI), offers opportunities such as integrating high-throughput phenotypes in genetic evaluations and solutions for training set optimization in genomic selection with highly heterogeneous data.

We made an effort to present our SSI methodology in deep detail, develop software for its implementation, and empirically validate it (with cross-validation) with real data using several data sets with genomic and phenotypic information. As a brand-new method, no exhaustive evaluation of the SSI was possible to be presented in this dissertation at this stage; however, the results obtained with these data sets are very promising, performing at least as good as the standard methods.

The SSI applications presented in this dissertation are of the type single-trait model; they might be feasibly extended to multi-trait models allowing the borrowing of information within and between individuals at the same time. Multi-trait models have the potential to increase prediction accuracy; therefore, further research is required to investigate whether the use of a multi-trait SSI can also be advantageous.

The scope of the SSI can go beyond the breeding values prediction only. This method leaves the door open to other genetic research areas (e.g., genome-wide association analysis on high-dimensional phenotypes and network analysis from gene expression). Therefore, more research is needed to explore the full potential of the SSI procedure.

APPENDICES

APPENDIX A

SUPPLEMENTARY FIGURES AND TABLES FROM CHAPTER 2

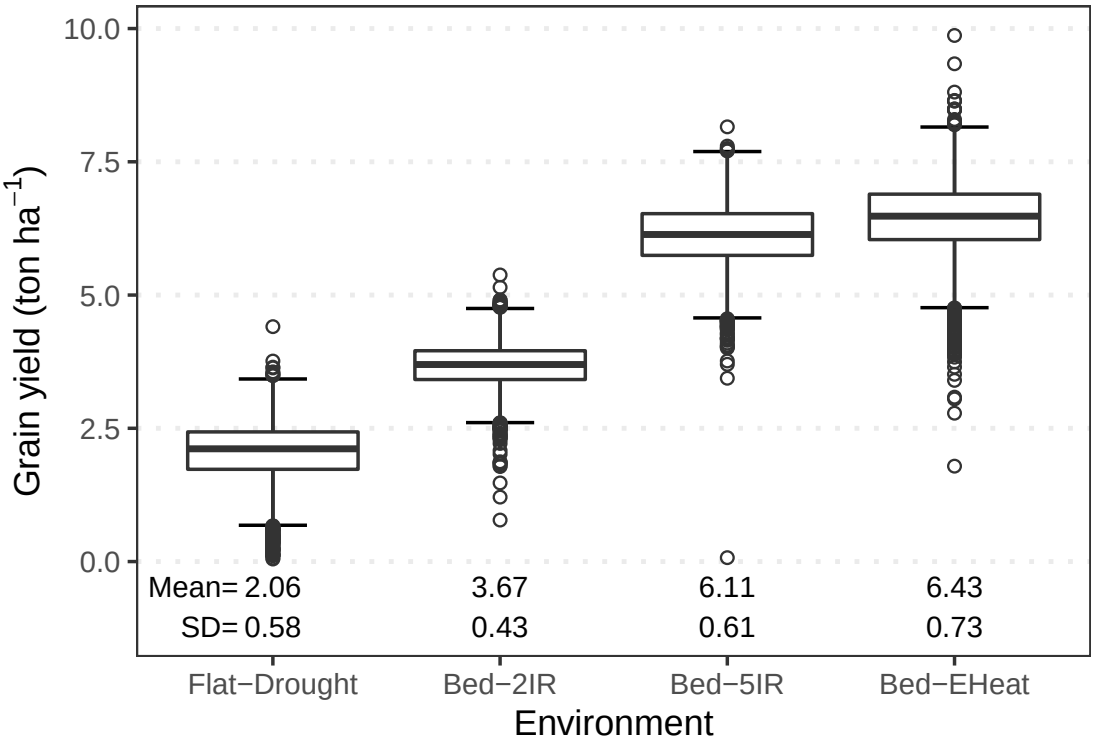

Figure A.1: Box-plot of grain yield phenotypic records by environmental condition. $n \approx 3200$ observations within environment. SD: standard deviation.

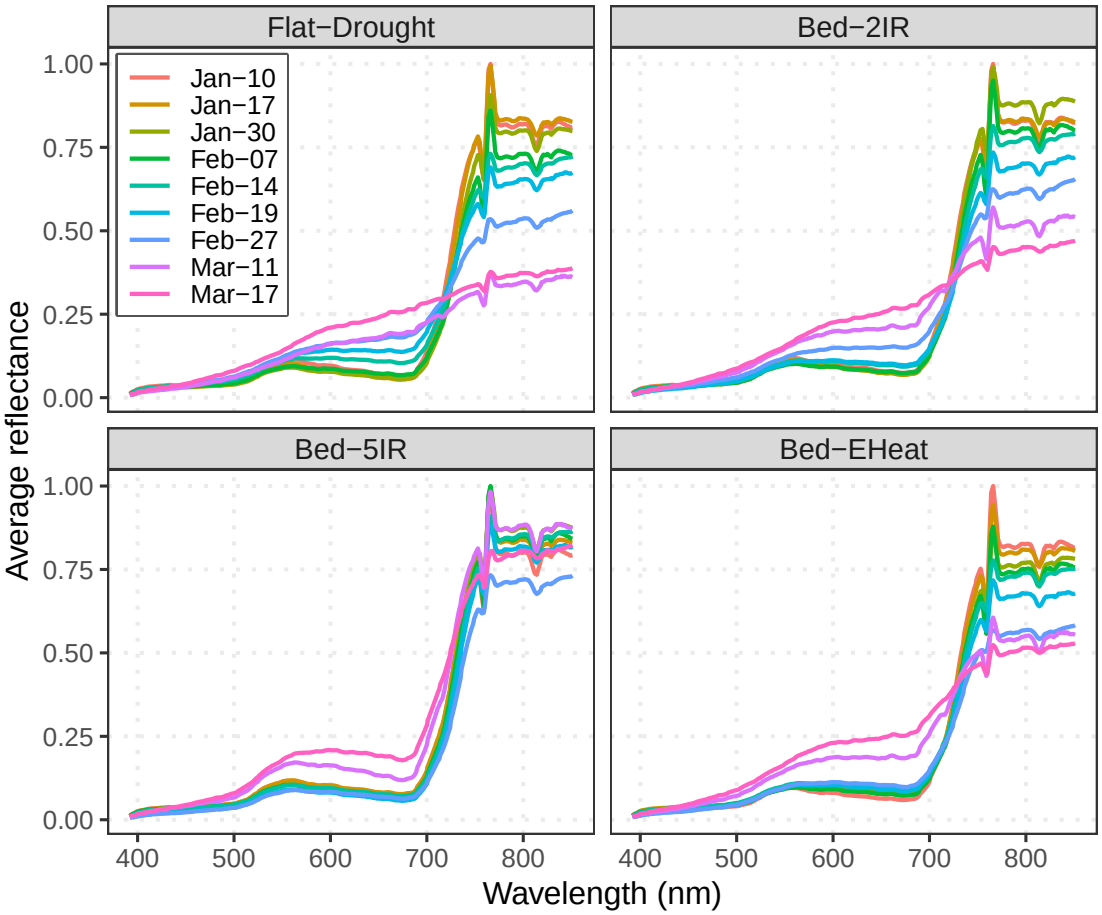

Figure A.2: Light reflectance patterns as function of the wavelength. Each line represents the mean (across $n \approx 3200$ observations) reflectance for each waveband, within time-point (flight date). Within each environment, means were scaled to lie within 0 and 1 by dividing them by the maximum average.

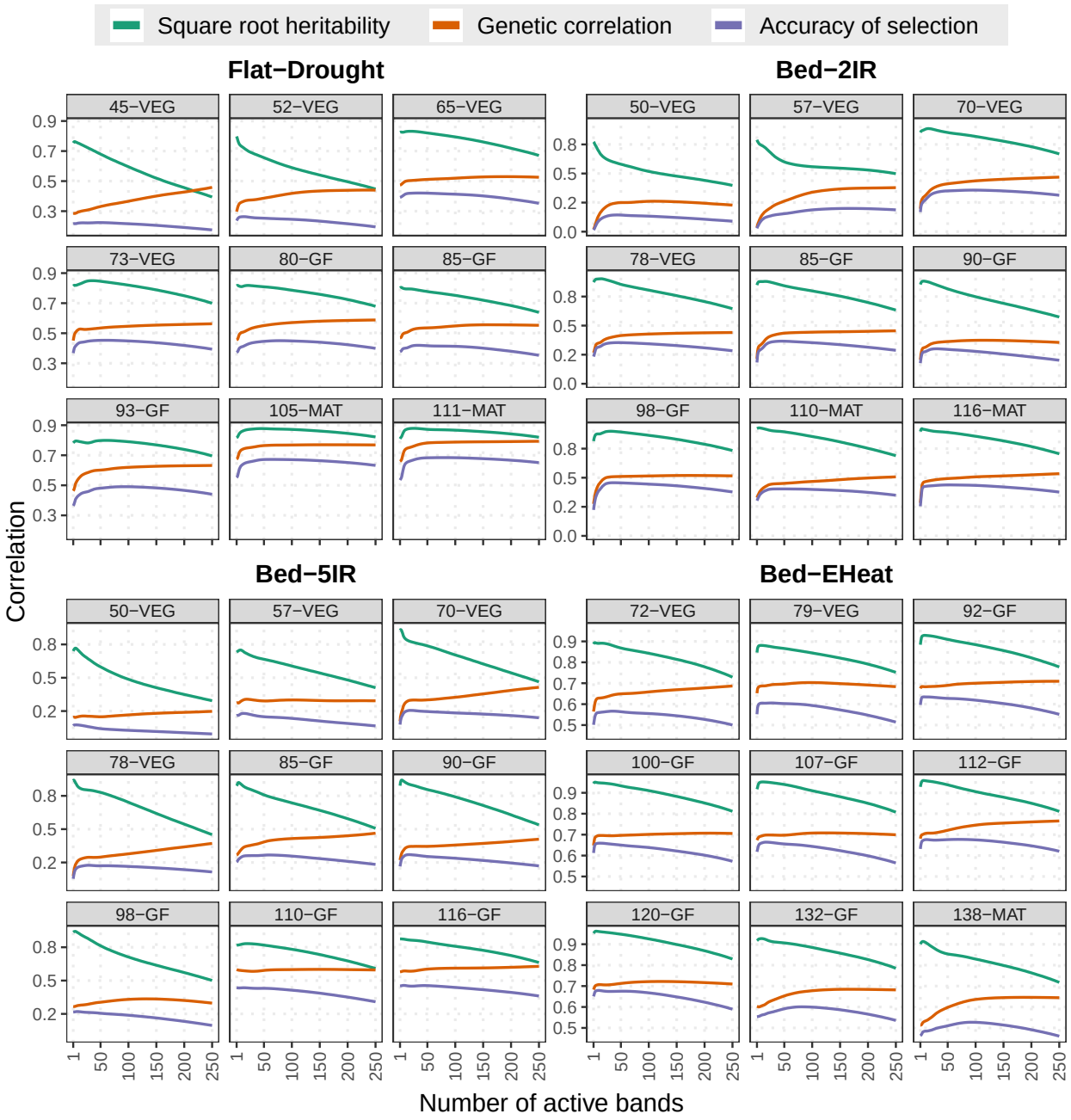

Figure A.3: Accuracy of indirect selection of L1-PSI and its components. Square root heritability, genetic correlation and accuracy of indirect selection, all averaged over 100 training-testing partitions versus the number of bands entering in the index; by time-point (DAS=days after sowing, Stage: VEG=vegetative, GF=grain filling, or MAT=maturity) within environment.

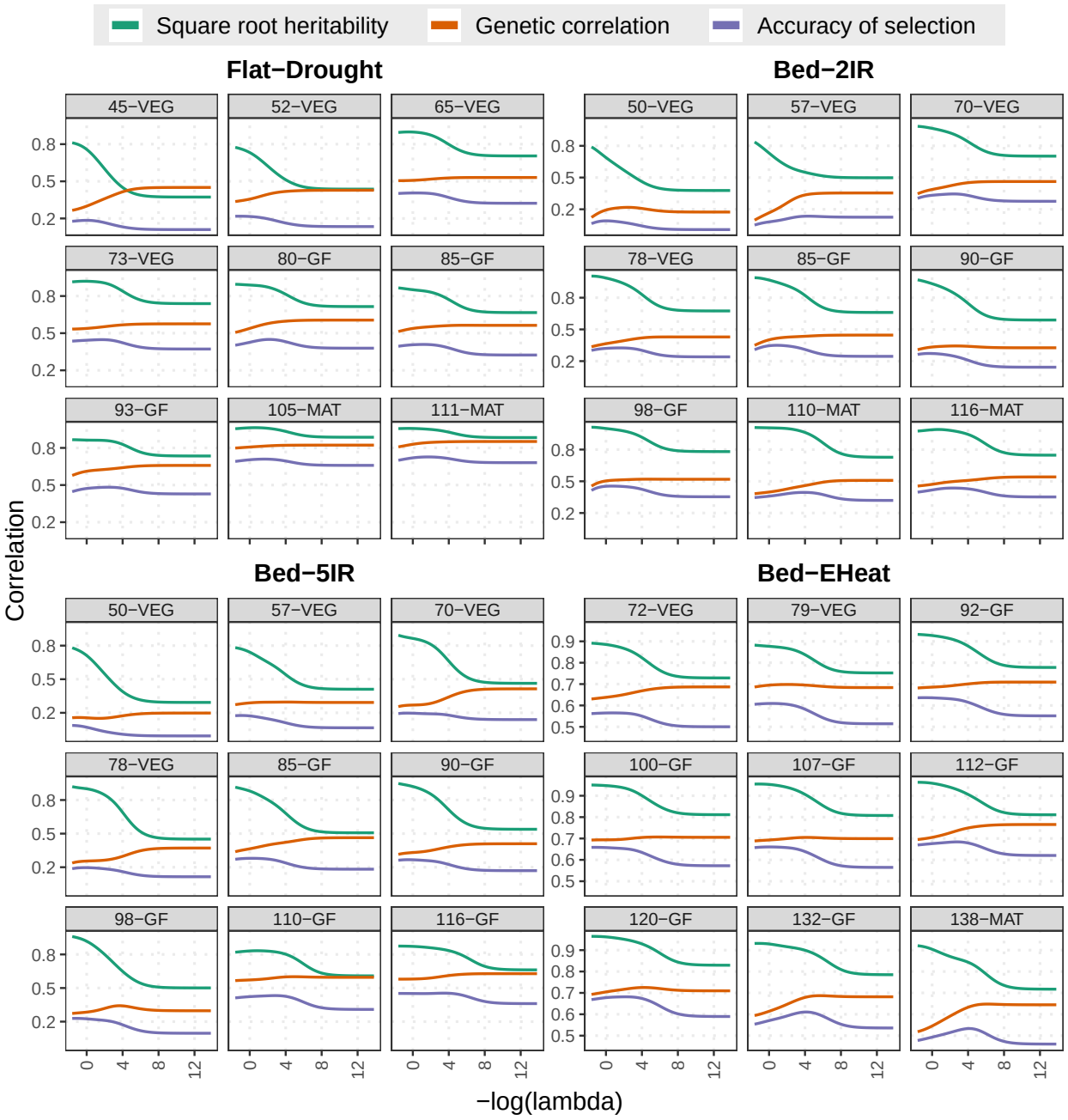

Figure A.4: Accuracy of indirect selection of L2-PSI and its components. Square root heritability, genetic correlation and accuracy of indirect selection, all averaged over 100 training-testing partitions versus the penalization parameter (λ, logarithm scale) used to build the index; by time-point (DAS=days after sowing, Stage: VEG=vegetative, GF=grain filling, or MAT=maturity) within environment.

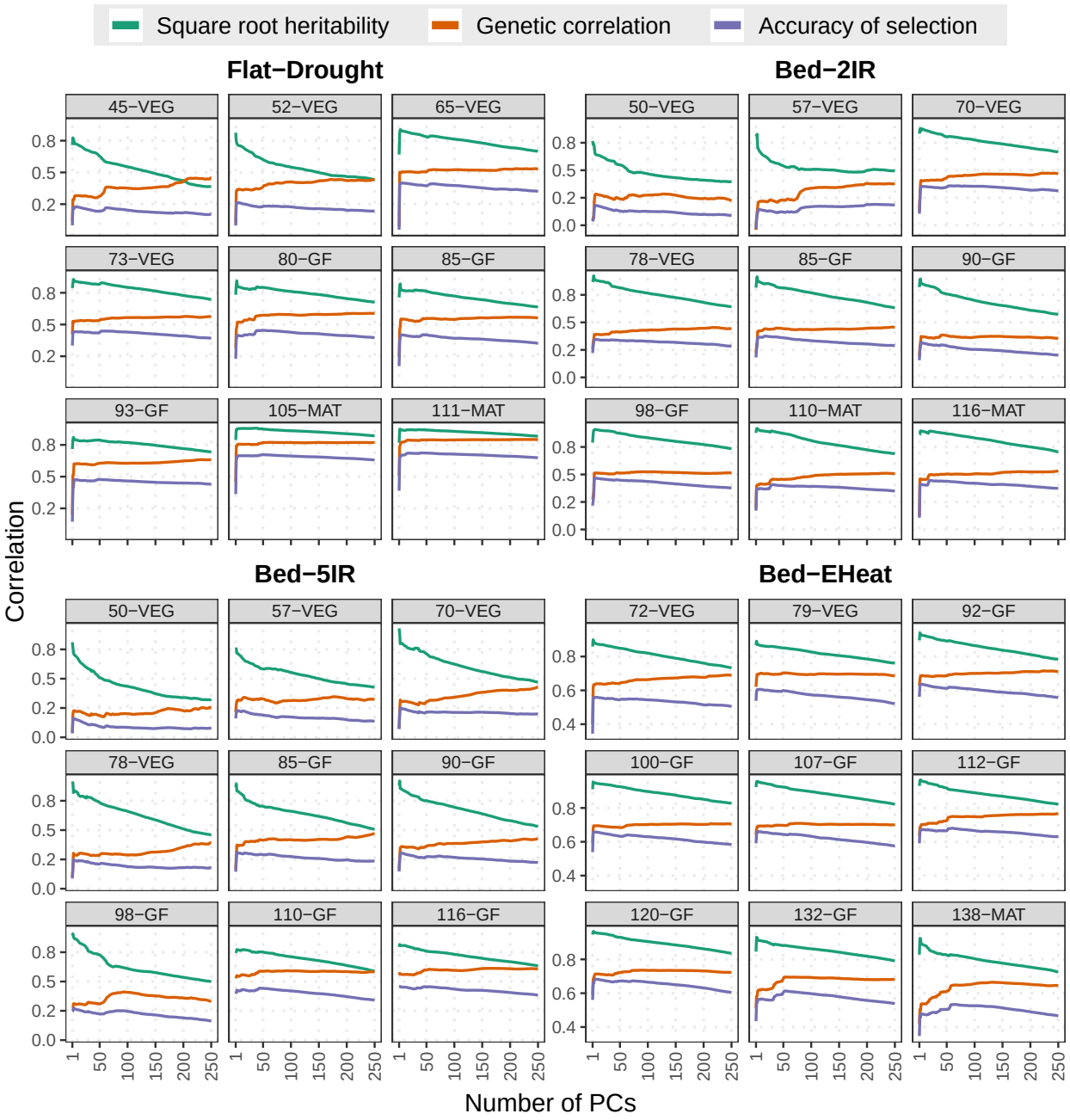

Figure A.5: Accuracy of indirect selection of PC-SI and its components. Square root heritability, genetic correlation and accuracy of indirect selection, all averaged over 100 training-testing partitions versus the number of principal components used to build the index; by time-point (DAS=days after sowing, Stage: VEG=vegetative, GF=grain filling, or MAT=maturity) within environment.

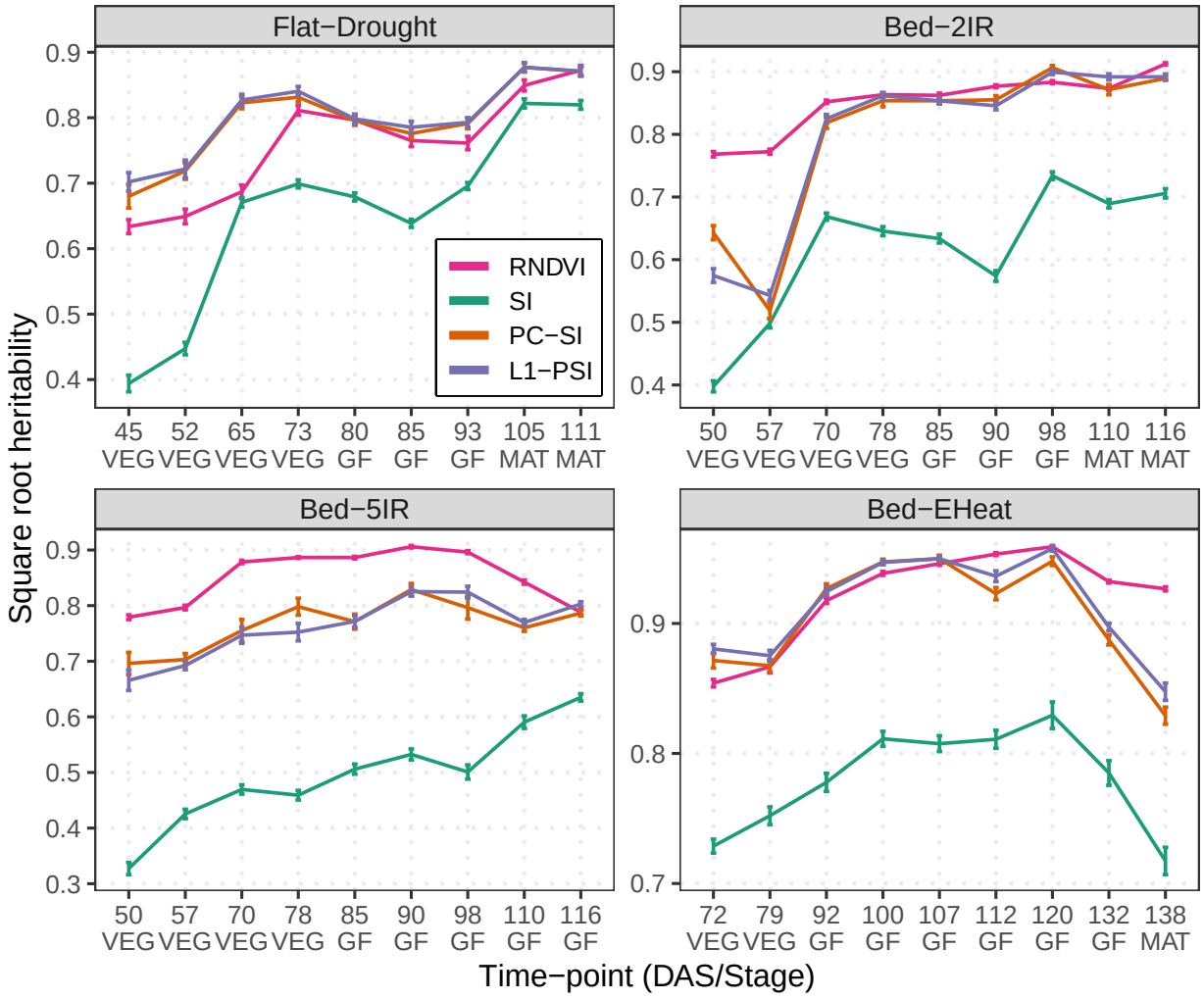

Figure A.6: Square root of heritability of the standard (SI), of the regularized (PC-SI and L1-PSI) selection indices, and of the RNDVI. The lines provide the average square root heritability over 100 training-testing partitions. Vertical lines represent a 95% CI for the average. The horizontal axis give the time-point at which images were collected and are expressed in both days after sowing (DAS) and stages (VEG=vegetative, GF=grain filling, MAT=maturity).

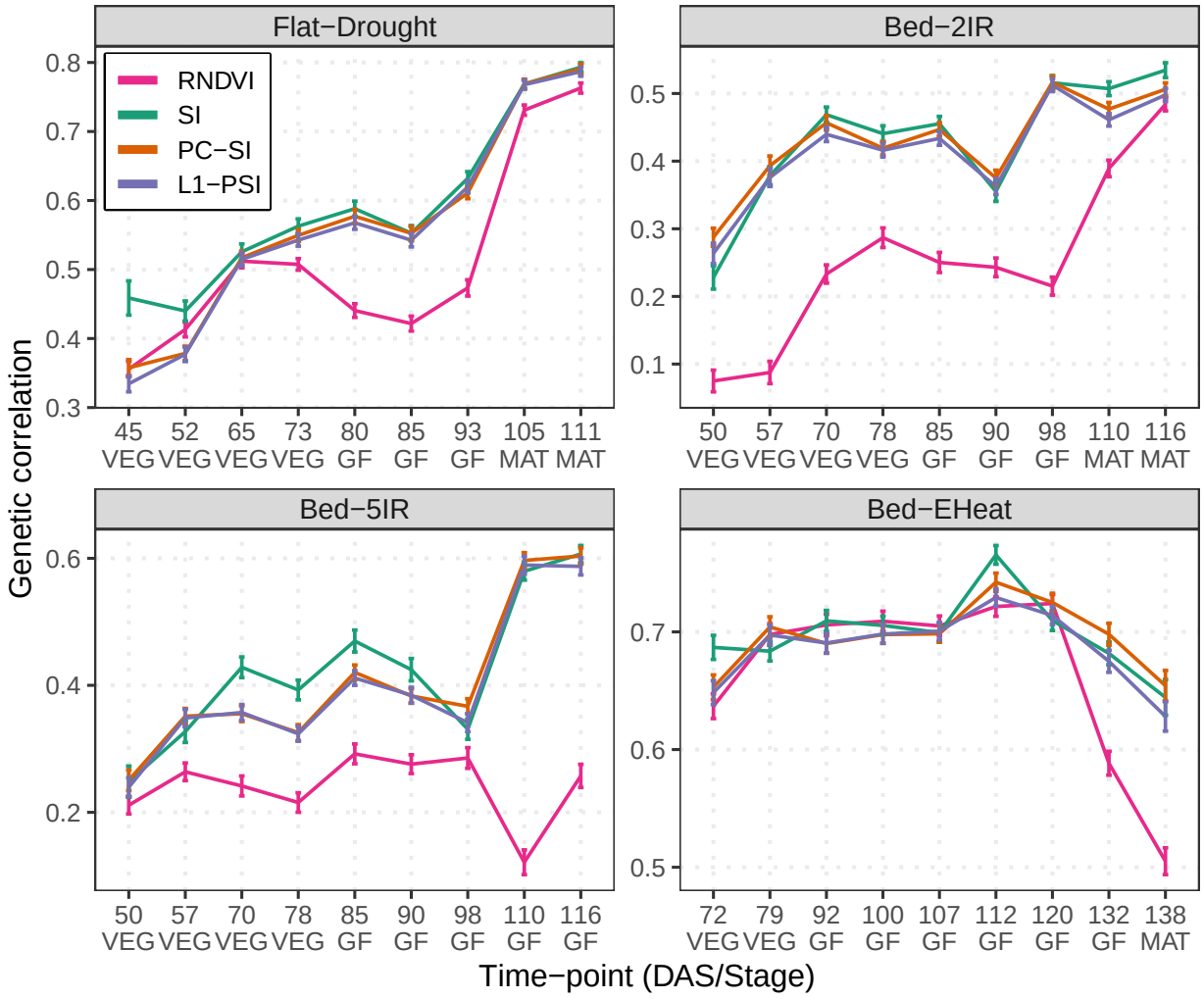

Figure A.7: Genetic correlation between grain yield and all: the standard (SI), the regularized (PC-SI and L1-PSI) selection indices, and the RNDVI. The lines provide the average genetic correlation over 100 training-testing partitions. Vertical lines represent a 95% CI for the average. The horizontal axis give the time-point at which images were collected and are expressed in both days after sowing (DAS) and stages (VEG=vegetative, GF=grain filling, MAT=maturity).

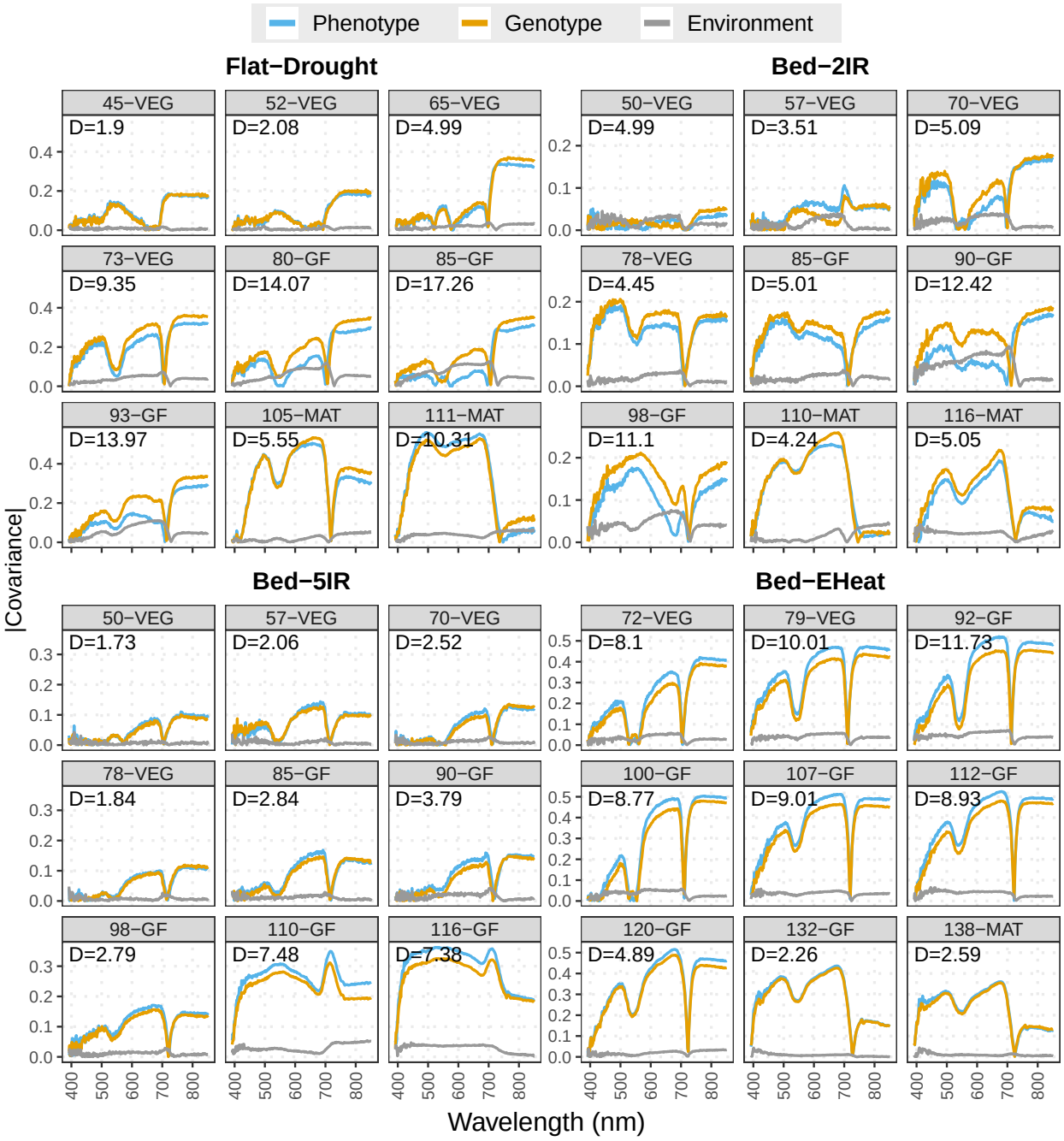

Figure A.8: Phenotypic, genetic, and environmental covariances (absolute value) between wavebands and grain yield. 'D': discrepancy between phenotypic and genetic covariances as measured by the sum of the absolute differences; by time-point (DAS: days after sowing, Stage: VEG=vegetative, GF=grain filling, MAT=maturity) within environment.

Table A.1: Accuracy of indirect selection (average over 100 training-testing partitions) for best phenotypic prediction (principal components (PCR), L1-penalized prediction (L1-Phen), RNDVI, and GNDVI) and for best genotypic prediction (standard SI, optimal PC-SI, L1-PSI, and L2-PSI).

		Phenotypic prediction				**Genotypic prediction**			
Env	**TP***	**PCR**	**L1-Phen**	**RNDVI**	**GNDVI**	**SI**	**PC-SI**	**L1-PSI**	**L2-PSI**
Flat-Drought	45	0.24 a	0.23 a	0.23 a	0.21 b	0.18 c	0.24 a	0.23 a	0.24 a
	52	0.27 ab	0.27 ab	0.27 ab	0.25 b	0.20 c	0.27 a	0.27 a	0.27 ab
	65	0.42 a	0.42 a	0.35 b	0.35 b	0.35 b	0.43 a	0.43 a	0.42 a
	73	0.45 ab	0.45 ab	0.41 cd	0.43 bc	0.39 d	0.46 a	0.46 a	0.46 a
	80	0.44 bc	0.43 c	0.35 e	0.39 d	0.40 d	0.46 a	0.45 ab	0.46 a
	85	0.41 abc	0.40 cd	0.32 f	0.39 d	0.35 e	0.43 a	0.43 ab	0.43 a
	93	0.46 bc	0.47 abc	0.36 e	0.45 cd	0.44 d	0.48 ab	0.49 a	0.49 a
	105	0.67 a	0.67 a	0.62 b	0.64 b	0.63 b	0.68 a	0.67 a	0.68 a
	111	0.68 ab	0.68 ab	0.67 bc	0.64 d	0.65 cd	0.69 ab	0.69 ab	0.69 a
	Multi	0.68 cd	0.68 bcd	0.68 d	0.65 e	0.00 f	0.70 ab	0.70 abc	0.70 a
Bed-2IR	50	0.18 a	0.14 cd	0.00 f	0.12 d	0.09 e	0.18 a	0.15 bc	0.16 ab
	57	0.19 a	0.19 a	0.00 c	0.03 b	0.19 a	0.20 a	0.20 a	0.20 a
	70	0.37 a	0.36 a	0.20 c	0.21 c	0.31 b	0.37 a	0.36 a	0.38 a
	78	0.35 a	0.35 a	0.25 c	0.30 b	0.28 b	0.36 a	0.36 a	0.36 a
	85	0.37 a	0.36 a	0.22 c	0.30 b	0.29 b	0.38 a	0.37 a	0.38 a
	90	0.30 abcd	0.29 cd	0.21 e	0.28 d	0.20 e	0.32 a	0.31 abc	0.32 ab
	98	0.45 a	0.46 a	0.18 d	0.35 c	0.38 b	0.47 a	0.46 a	0.46 a
	110	0.40 abc	0.39 bc	0.34 d	0.39 c	0.35 d	0.42 a	0.41 ab	0.42 a
	116	0.44 a	0.44 a	0.44 a	0.39 b	0.38 b	0.45 a	0.44 a	0.45 a
	Multi	0.53 cd	0.53 d	0.46 e	0.40 f	0.01 g	0.55 ab	0.54 bc	0.56 a
Bed-5IR	50	0.18 a	0.17 ab	0.16 ab	0.15 b	0.08 c	0.17 ab	0.16 ab	0.16 ab
	57	0.25 a	0.25 a	0.21 c	0.21 bc	0.14 d	0.25 a	0.24 a	0.24 ab
	70	0.27 a	0.26 a	0.21 b	0.19 b	0.20 b	0.27 a	0.27 a	0.26 a
	78	0.26 a	0.24 a	0.19 b	0.19 b	0.18 b	0.26 a	0.24 a	0.26 a
	85	0.32 a	0.32 a	0.26 b	0.25 b	0.24 b	0.32 a	0.32 a	0.33 a
	90	0.31 a	0.31 a	0.25 c	0.28 b	0.22 d	0.32 a	0.32 a	0.32 a
	98	0.30 a	0.29 a	0.26 b	0.25 b	0.16 c	0.29 a	0.28 a	0.28 a
	110	0.46 a	0.45 a	0.10 d	0.22 c	0.34 b	0.45 a	0.45 a	0.45 a
	116	0.47 a	0.47 a	0.20 d	0.34 c	0.38 b	0.47 a	0.47 a	0.47 a
	Multi	0.54 a	0.54 a	0.32 c	0.37 b	0.00 d	0.54 a	0.55 a	0.55 a
Bed-Heat	72	0.57 a	0.57 a	0.54 b	0.53 b	0.50 c	0.57 a	0.57 a	0.57 a
	79	0.61 a	0.61 a	0.60 a	0.58 b	0.51 c	0.61 a	0.61 a	0.61 a
	92	0.64 a	0.64 a	0.65 a	0.63 a	0.55 b	0.64 a	0.64 a	0.64 a
	100	0.66 a	0.66 a	0.67 a	0.65 a	0.57 b	0.66 a	0.66 a	0.66 a
	107	0.66 a	0.66 a	0.67 a	0.66 a	0.56 b	0.66 a	0.67 a	0.66 a
	112	0.68 ab	0.68 ab	0.69 a	0.66 b	0.62 c	0.68 a	0.68 a	0.69 a
	120	0.69 a	0.68 a	0.69 a	0.66 b	0.59 c	0.69 a	0.68 a	0.68 a
	132	0.62 a	0.61 a	0.55 b	0.54 b	0.54 b	0.62 a	0.61 a	0.61 a
	138	0.54 a	0.53 a	0.47 b	0.46 b	0.46 b	0.54 a	0.53 a	0.54 a

Table A.1 (cont'd)

Multi	0.71 a	0.70 a	0.70 a	0.67 b	0.00 c	0.71 a	0.71 a	0.72 a

*Each row contains results for each environment and time-point (DAS: days after sowing). Models with the same letter (within each row) are not significantly different from each other (α = 0.05, ANOVA followed by Tukey test).

APPENDIX B

SUPPLEMENTARY MATERIAL FROM CHAPTER 3

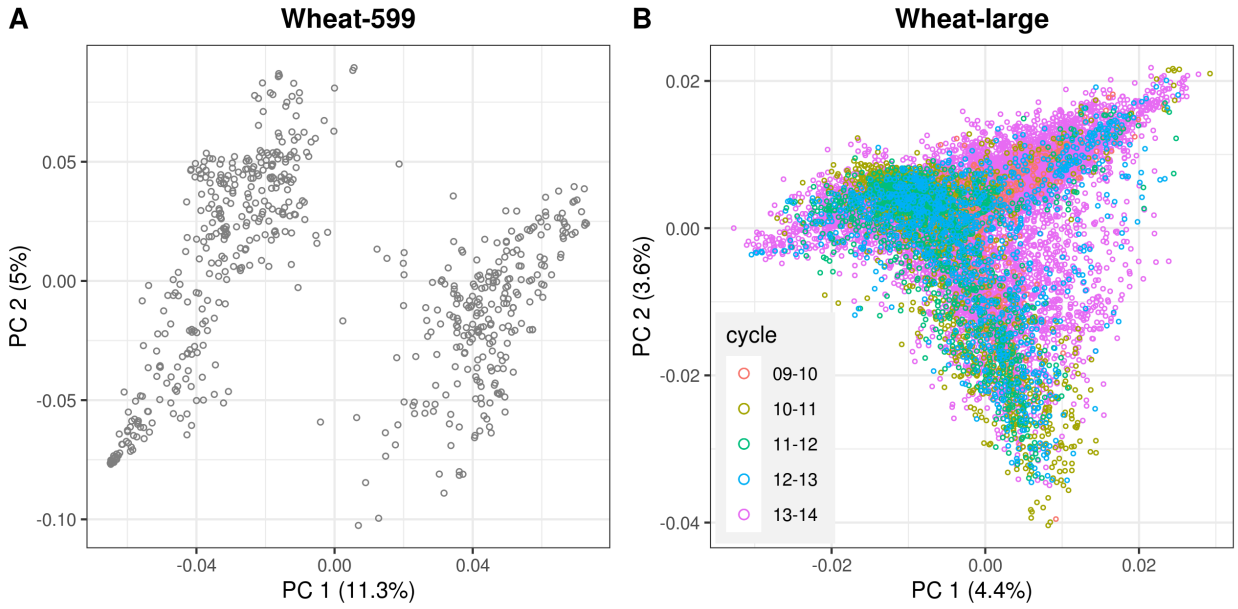

Figure B.1: Top two principal components of the genomic relationship matrix, **G**, for each data set. Each point represent individuals. (A) Wheat-599 data set. (B) Wheat-large data set. Individuals are color-grouped by the cycle (sowing-harvest year).

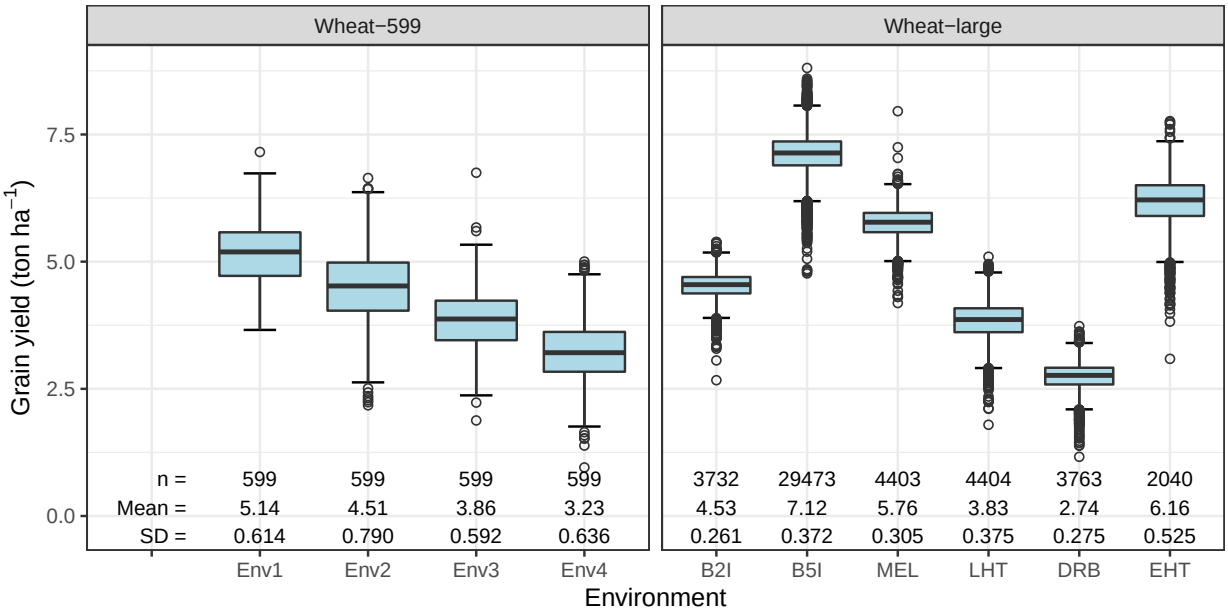

Figure B.2: Boxplot of grain yield phenotypic records (in ton ha^{-1}) by environmental condition for both Wheat-599 and Wheat-large data sets. SD standard deviation.

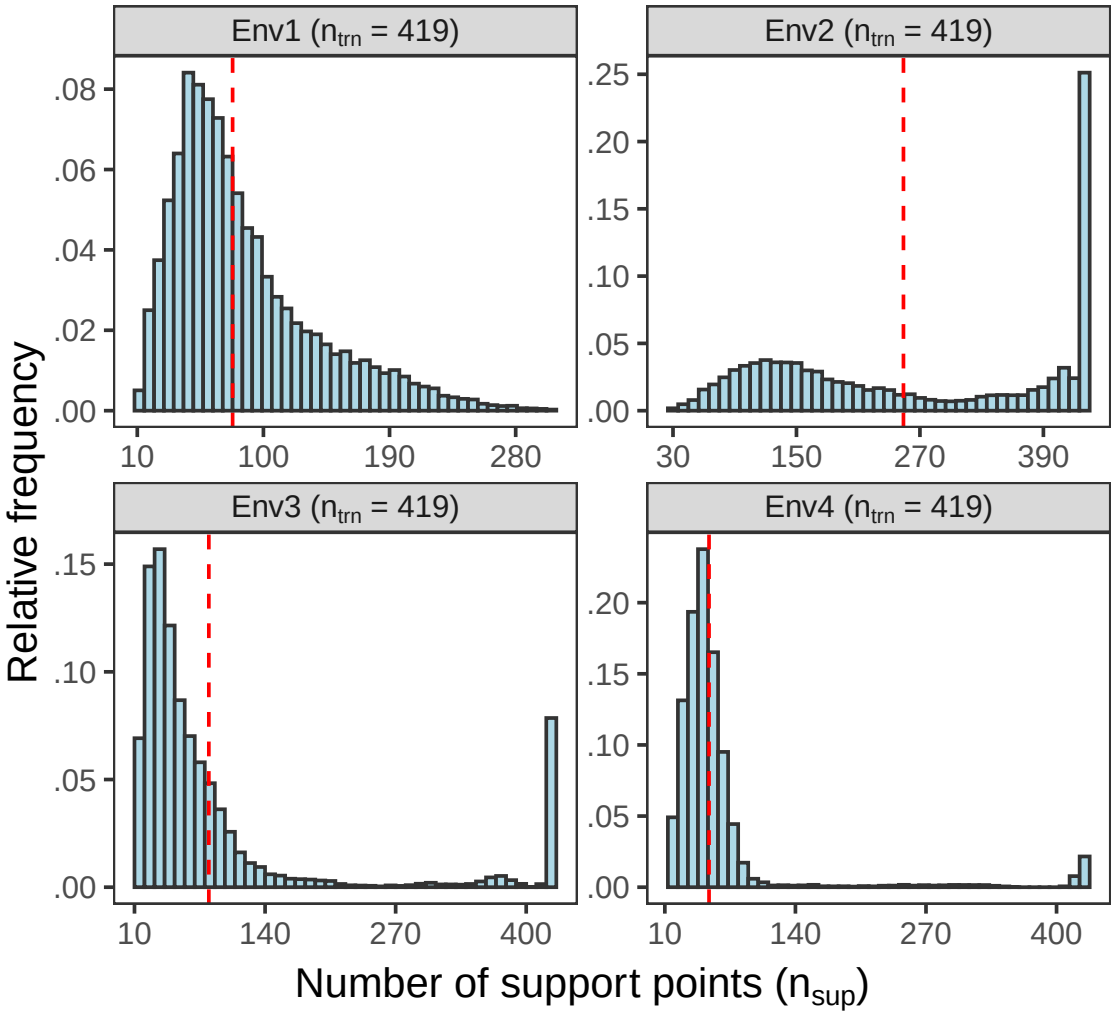

Figure B.3: Distribution of the number of training support points (n_{sup}) in optimal sparse selection indices (results obtained over 100 trn-tst partitions; n_{trn}= size of the training data set), by environmental condition, Wheat-599 data set.

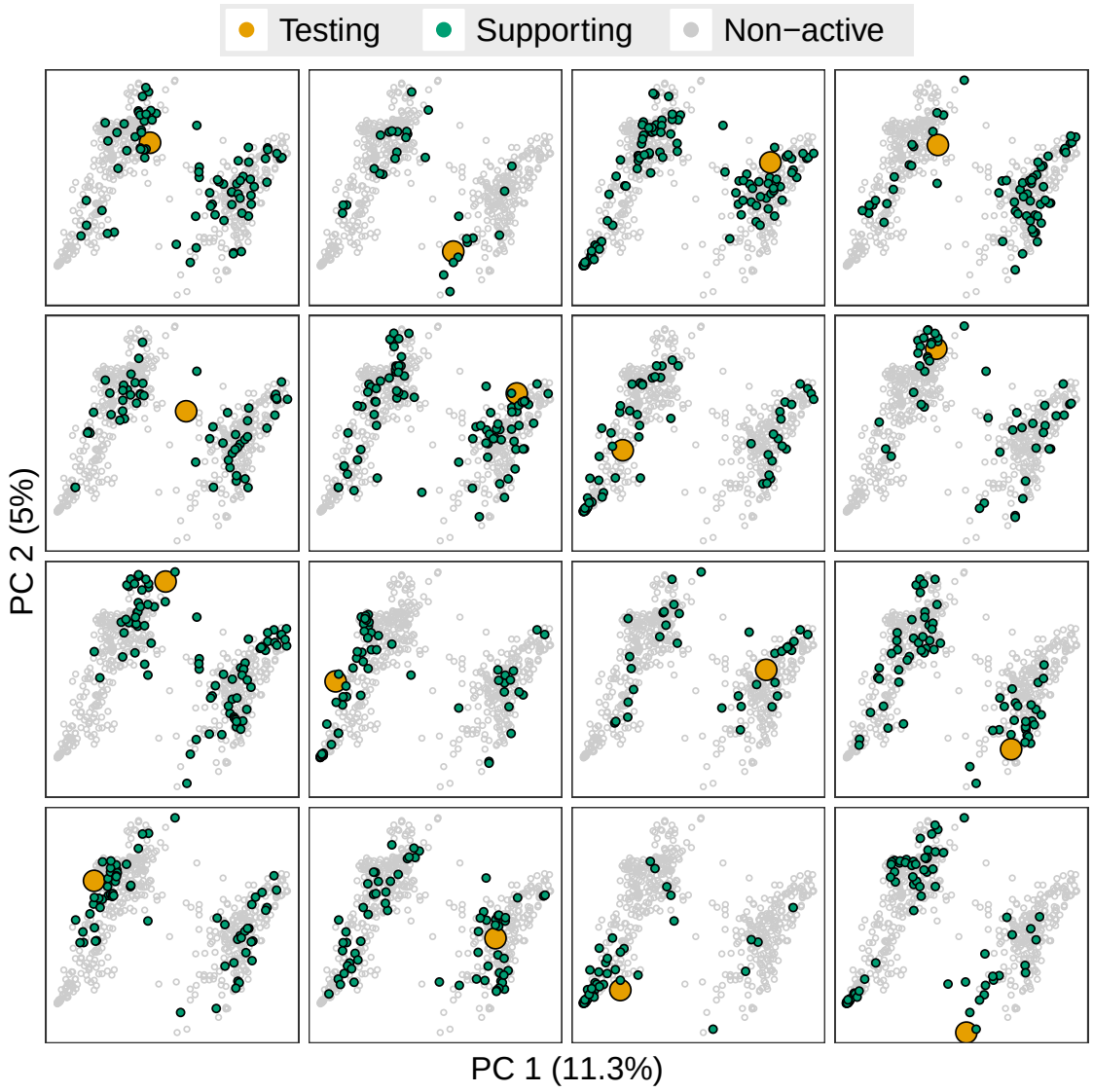

Figure B.4: First two principal components coordinates for prediction points (yellow) and the corresponding support points (green). Grey points represent genotypes that did not contribute to the prediction of the genetic value of the genotype in yellow. All panels represent solutions for the environment 1, Wheat-599 data set.

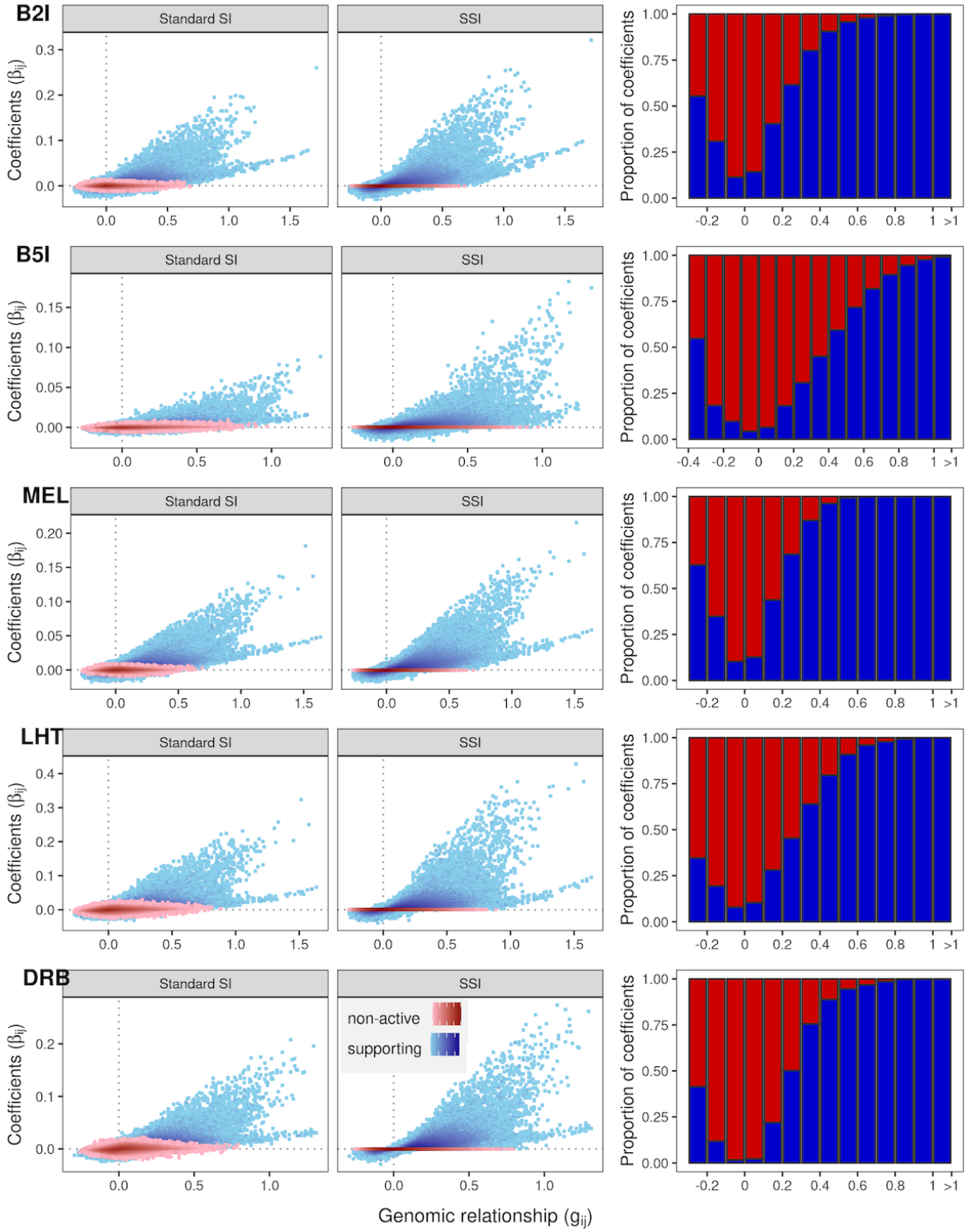

Figure B.5: (left and center) Weights (β_{ij}) of a standard SI (G-BLUP) and of the optimal sparse selection index (SSI) versus the genomic relationship (g_{ij}), and (right) proportion of weights in the SSI that belonged to either the supporting or non-active sets, by genomic-relationship; by environment, Wheat-large data set.

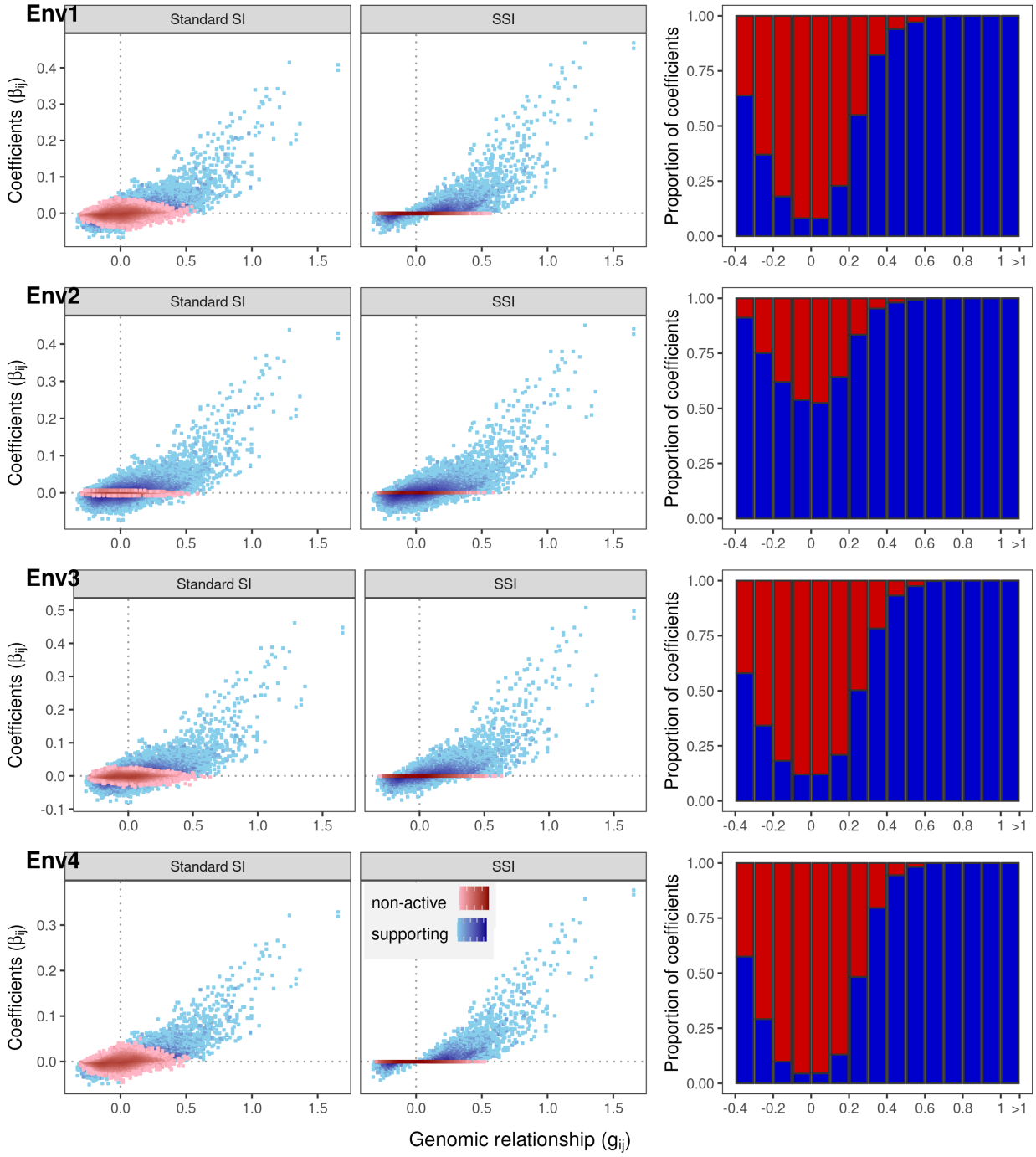

Figure B.6: (left and center) Weights (β_{ij}) of a standard SI (G-BLUP) and of the optimal sparse selection index (SSI) versus the genomic relationship (g_{ij}), and (right) proportion of weights in the SSI that belonged to either the supporting or non-active sets, by genomic-relationship; by environment, Wheat-599 data set.

Table B.1: Number of available observations, average grain yield, and heritability by environmental condition for the Wheat-large data set.

Planting conditions		Number of irrigations	Name	n	Average (SD) Yield	Heritability (SD)[a]
Date	System					
Optimum	Bed	2	B2I	3,732	4.53 (0.261)	0.41 (0.029)
Optimum	Bed	5	B5I	29,473	7.12 (0.372)	0.57 (0.025)
Optimum	Flat	5	MEL	4,403	5.76 (0.305)	0.23 (0.025)
Late	Bed	5	LHT	4,404	3.83 (0.375)	0.51 (0.025)
Optimum	Bed	Minimal	DRB	3,763	2.74 (0.275)	0.38 (0.029)
Early	Bed	5	EHT	2,040	6.16 (0.525)	0.41 (0.038)

SD. Standard deviation. [a]Posterior mean and SD obtained across 10,000 Monte Carlo replicates using Gibbs sampling.

Table B.2: Number of available observations, average grain yield, and heritability by environmental condition for the Wheat-599 data set.

Moisture regime	Temperature	Name	n	Average (SD) Yield	Heritability (SD)[a]
Optimal irrigation, low rainfall	Optimal	Env1	599	5.14 (0.614)	0.50 (0.054)
High rainfall	Optimal	Env2	599	4.51 (0.790)	0.46 (0.056)
Low rainfall	High drought	Env3	599	3.86 (0.592)	0.43 (0.062)
Irrigation or rainfall	Hot, low humidity	Env4	599	3.23 (0.636)	0.44 (0.061)

SD. Standard deviation. [a]Posterior mean and SD obtained across 10,000 Monte Carlo replicates using Gibbs sampling.

Table B.3: Maximum prediction accuracy (average across 100 partitions) achieved by the SSI for different values of the parameter α of an Elastic-Net-type SSI, by environmental condition for the Wheat-large data set.

Environment	$\lambda_0{}^a$	α	$\lambda_{opt}{}^b$	$\boldsymbol{n}_{sup}{}^c$	**Accuracy (SD)**^{d}
B2I	1.5320	1.00	0.0141	395	0.649 (0.032) b
	0.7660	0.25	0.0667	320	0.663 (0.032) a
		0.50	0.0320	330	0.664 (0.032) a
		0.75	0.0233	290	0.664 (0.032) a
		1.00	0.0155	338	0.664 (0.032) a
B5I	1.8412	1.00	0.0119	1,226	0.610 (0.009) b
	0.9215	0.25	0.0460	1,187	0.630 (0.009) a
		0.50	0.0223	1,203	0.631 (0.009) a
		0.75	0.0164	1,044	0.631 (0.009) a
		1.00	0.0132	943	0.631 (0.009) a
MEL	3.7934	1.00	0.0116	561	0.665 (0.046) b
	1.8967	0.25	0.0705	338	0.685 (0.045) a
		0.50	0.0406	270	0.686 (0.045) a
		0.75	0.0294	236	0.687 (0.045) a
		1.00	0.0195	282	0.687 (0.045) a
LHT	0.9841	1.00	0.0218	237	0.712 (0.026) b
	0.4921	0.25	0.0854	248	0.727 (0.025) a
		0.50	0.0491	194	0.729 (0.025) a
		0.75	0.0295	223	0.729 (0.025) a
		1.00	0.0235	202	0.730 (0.025) a
DRB	1.7555	1.00	0.0344	103	0.679 (0.038) b
	0.8778	0.25	0.1461	102	0.694 (0.039) a
		0.50	0.0823	79	0.696 (0.039) a
		0.75	0.0588	69	0.697 (0.040) a
		1.00	0.0386	85	0.697 (0.040) a
EHT	1.5514	1.00	0.0284	120	0.657 (0.046) a
	0.7757	0.25	0.0970	159	0.670 (0.048) a
		0.50	0.0554	126	0.672 (0.048) a
		0.75	0.0399	110	0.672 (0.048) a
		1.00	0.0264	133	0.673 (0.048) a

SD: Standard deviation across the 100 trn-tst partitions. ^{a}Shrinkage factor involved in the standard SI (Equation (3.2)). Within environment, in the top row a value of λ_0 was used as in the G-BLUP and in rows below, λ_0 was reduced to half. ^{b}Optimal value of λ that yielded an SSI with the maximum accuracy among all indices obtained for a grid of 100 values of λ. ^{c}Average number of individuals in supporting the prediction of individuals from testing set. ^{d}Models with the same letter are not significantly different from others (ANOVA followed by Tukey's HSD test, 5% significance level.

Table B.4: Maximum prediction accuracy (average across 100 partitions) achieved by the SSI for different values of the parameter α of an Elastic-Net-type SSI, by environmental condition for the Wheat-599 data set.

Environment	$\lambda_0{}^a$	α	$\lambda_{opt}{}^b$	$\boldsymbol{n}_{sup}{}^c$	**Accuracy (SD)**^{d}
Env1	1.2101	1.00	0.0314	84	0.769 (0.062) a
	0.5061	0.25	0.1042	99	0.772 (0.063) a
		0.50	0.0492	101	0.773 (0.063) a
		0.75	0.0296	110	0.773 (0.063) a
		1.00	0.0236	103	0.773 (0.063) a
Env2	1.3034	1.00	0.0175	151	0.708 (0.085) a
	0.6517	0.25	0.0686	147	0.711 (0.086) a
		0.50	0.0397	126	0.710 (0.086) a
		0.75	0.0240	136	0.710 (0.086) a
		1.00	0.0192	129	0.710 (0.086) a
Env3	1.4084	1.00	0.0514	50	0.609 (0.090) a
	0.7042	0.25	0.2213	48	0.611 (0.089) a
		0.50	0.1017	48	0.610 (0.090) a
		0.75	0.0601	54	0.609 (0.091) a
		1.00	0.0474	50	0.609 (0.091) a
Env4	1.4380	1.00	0.0615	40	0.722 (0.073) a
	0.7190	0.25	0.2689	39	0.727 (0.074) a
		0.50	0.1483	31	0.727 (0.074) a
		0.75	0.0870	35	0.728 (0.075) a
		1.00	0.0681	32	0.728 (0.075) a

SD: Standard deviation across the 100 trn-tst partitions. ^{a}Shrinkage factor involved in the standard SI (Equation (3.2)). Within environment, in the top row a value of λ_0 was used as in the G-BLUP and in rows below, λ_0 was reduced to half. ^{b}Optimal value of λ that yielded an SSI with the maximum accuracy among all indices obtained for a grid of 100 values of λ. ^{c}Average number of individuals in supporting the prediction of individuals from testing set. ^{d}Models with the same letter are not significantly different from others (ANOVA followed by Tukey's HSD test, 5% significance level.

B.1 Equivalence between Standard Selection Index and BLUP

Consider a standard single-trait model of the form

$$\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y} = (y_1, ..., y_n)'$, $\boldsymbol{u} = (u_1, ..., u_n)'$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$ are vectors of phenotypes, genetic, and environmental effects, respectively. Here, for simplicity we assume that all these vectors have zero-mean.

In a standard G-BLUP model, $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are assumed to be independent (i.e., $cov(\boldsymbol{u}, \boldsymbol{\varepsilon}') = \mathbf{0}$), both have null means (i.e., $\mathbb{E}(\boldsymbol{u}) = \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$), and (co)variance matrices $var(\boldsymbol{u}) = \sigma^2_u \mathbf{G}$ and $var(\boldsymbol{\varepsilon}) = \sigma^2_\varepsilon \mathbf{I}$, respectively; here $\mathbf{G}$ is a relationship matrix that could be derived from a pedigree or from DNA sequences.

Consider now a partition of each of the data in into a training (trn) and a testing (tst) set. The objective is to predict the genetic values of the individuals in the testing set ($\boldsymbol{u}_{tst}$) using the phenotype data available from the training set ($\boldsymbol{y}_{trn}$). The (co)variance matrix of the vector of breeding values can be partitioned as follows

$$var\left(\begin{bmatrix} \boldsymbol{u}_{trn} \\ \boldsymbol{u}_{tst} \end{bmatrix}\right) = \sigma^2_u \begin{bmatrix} \mathbf{G}_{trn} & \mathbf{G}_{trn,tst} \\ \mathbf{G}'_{trn,tst} & \mathbf{G}_{tst} \end{bmatrix}$$

where $\mathbf{G}_{trn}$ and $\mathbf{G}_{tst}$ are the genetic relationship submatrices for the training and testing data points, respectively, and $\mathbf{G}_{trn,tst}$ is the genetic relationship submatrix between training and testing subjects. The Best Linear Predictor (BLP) of $\boldsymbol{u}_{tst}$ ($\hat{\boldsymbol{u}}_{tst}$) takes the form (e.g., Searle et al., 1992):

$$\begin{aligned} \mathbb{E}(\boldsymbol{u}_{tst}|\boldsymbol{y}_{trn}) &= \mathbb{E}(\boldsymbol{u}_{tst}) + cov(\boldsymbol{u}_{tst}, \boldsymbol{y}'_{trn}) \left[var(\boldsymbol{y}_{trn})\right]^{-1} \left(\boldsymbol{y}_{trn} - \mathbb{E}(\boldsymbol{y}_{trn})\right) \\ &= \mathbf{G}'_{trn,tst} \left(\mathbf{G}_{trn} + \lambda_0 \mathbf{I}\right)^{-1} \boldsymbol{y}_{trn}. \end{aligned}$$

Alternatively, one can write $\hat{\boldsymbol{u}}_{tst} = \mathbf{H} \cdot \boldsymbol{y}_{trn}$, where $\mathbf{H} = \mathbf{G}'_{trn,tst}(\mathbf{G}_{trn} + \lambda_0 \mathbf{I})^{-1}$ is a "Hat" matrix. Thus, the BLUP of the genetic value of the i^{th} testing individual is $\hat{\boldsymbol{u}}_{tst(i)} = \mathbf{H}'_i \boldsymbol{y}_{trn}$ where $\mathbf{H}'_i$ is the i^{th} row of $\mathbf{H}$, that is $\mathbf{H}'_i = \mathbf{G}'_i(\mathbf{G}_{trn} + \lambda_0 \mathbf{I})^{-1}$ which is equal to the weights of the standard selection index, $\hat{\boldsymbol{\beta}}'_i = \mathbf{G}'_i(\mathbf{G}_{trn} + \lambda_0 \mathbf{I})^{-1}$ (see Equation (3.2) in the manuscript).

B.2 Sparse Selection Indices (SSI) using the SFSI R-package

In this section, we use data from the Wheat-599 data set used in the manuscript to illustrate how to fit Sparse Selection Indices using the SFSI R-package (Lopez-Cruz et al., 2020).

B.2.1 Installing the package from GitHub

The following snippet shows how to install the package from GitHub.

```
rm(list = ls())

# Install devtools package first
install.packages('devtools', repos='https://cran.r-project.org/')

# Install SFSI package from GitHub
devtools::install_git('https://github.com/MarcooLopez/SFSI')
library(SFSI)    # Load the package

# Install BGLR package (needed to download the data)
install.packages('BGLR', repos='https://cran.r-project.org/')
```

B.2.2 Data preparation

To illustrate the use of the software we will use data from the Wheat-599 data set which is available with the BGLR R-package (Perez & de los Campos, 2014). The following code shows how to prepare data for environment 1; all the analyses hereinafter are based on this data.

```
data(wheat, package="BGLR")     # Load data from the BGLR package

# Select the environment 1 to work with
y <- as.vector(scale(wheat.Y[,1]))

# Calculate G matrix
G <- tcrossprod(scale(wheat.X))/ncol(wheat.X)

# Create a directory and save data
dir.create("data", recursive=TRUE)
save(y, G, file="data/geno_pheno.RData")
```

B.2.3 Heritability and variance components

Implementing the SSI requires an estimate of the heritability. We obtain this using a G-BLUP model $y_i = \mu + u_i + \varepsilon_i$ with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2_\varepsilon)$ and $\boldsymbol{u} \sim N(\mathbf{0}, \sigma^2_u \mathbf{G})$. This model can be fitted with the 'fitBLUP' function included in the SFSI R-package. The BGLR R-package can be also used to fit a Bayesian version of the model. The code below illustrates how to estimate heritability using the 'fitBLUP' function.

```
load("data/geno_pheno.RData") # Load data

# Fit model
fm0 <- fitBLUP(y, K=G)
fm0$h2 <- fm0$varU/(fm0$varU+fm0$varE)  # Estimate heritability
c(fm0$varU,fm0$varE,fm0$h2)             # Variance components (varU,varE,h2)

# Create a directory and save data
dir.create("output", recursive=TRUE)
save(fm0, file="output/varComps.RData")
```

B.2.4 Training-testing partitions

The code below produces training (trn, 70%) and testing (tst, 30%) partitions. The parameter 'nPart' defines the number of partitions. The output is a matrix with 'nPart' columns and 180 rows containing indices representing the observations that are assigned to the testing sets. The object is saved in the file 'partitions.RData' and will be used in later analyses.

```
nPart <- 5                        # Number of partitions
load("data/geno_pheno.RData")     # Load data
nTST <- ceiling(0.3*length(y))    # Number of elements in TST set
partitions <- matrix(NA,nrow=nTST,ncol=nPart)   # Matrix to store partitions
seeds <- round(seq(1E3, .Machine$integer.max, length=nPart))

for(k in 1:nPart)
{   set.seed(seeds[k])
    partitions[,k] <- sample(1:length(y), nTST, replace=FALSE)
}
save(partitions, file="output/partitions.RData")    # Save partitions
```

B.2.5 Accuracy of the G-BLUP and of the Sparse SI

The following script shows how to derive SSIs using the partitions above created. The weights of the SSI are computed using the 'SSI' function for 'nLambda=100' values of λ. The G-BLUP model is fitted for comparison using the 'fitBLUP' function. Estimates of μ and h^2 are computed internally in the 'SSI' function when these are not provided. These estimates obtained from the G-BLUP model will be passed to the 'SSI' function to save time. Indices denoting training and testing sets are passed through the 'trn' and 'tst' parameters, respectively. The accuracy of the G-BLUP and SSI models are stored in the object 'accSSI', and saved in the file 'results_accuracy.RData'.

```
# Load data
load("data/geno_pheno.RData")
load("output/varComps.RData")
load("output/partitions.RData")

accSSI <- mu <- h2 <- c()       # Objects to store results

for(k in 1:ncol(partitions))
{ cat("  partition = ",k,"\n")
  tst <- partitions[,k]
  trn <- (1:length(y))[-tst]
  yNA <- y;   yNA[tst] <- NA

  # G-BLUP model
  fm1 <- fitBLUP(yNA, K=G)
  mu[k] <- fm1$b        # Retrieve mu estimate
  h2[k] <- fm1$h2       # Retrieve h2 estimate

  # Sparse SI
  fm2 <- SSI(y,K=G,b=mu[k],h2=h2[k],trn=trn,tst=tst,mc.cores=1,nLambda=100)
  fm3 <- summary(fm2)   # Useful function to get results

  accuracy <- c(GBLUP=cor(fm1$u[tst],y[tst]), fm3$accuracy)/sqrt(fm0$h2)
  lambda <- c(min(fm3$lambda),fm3$lambda)
  df <- c(max(fm3$df),fm3$df)
  accSSI <- rbind(accSSI,data.frame(rep=k,SSI=names(accuracy),accuracy,lambda,df))
}
save(mu,h2,accSSI,file="output/results_accuracy.RData")
```

B.2.6 Displaying Results

The following code creates a plot (as in Figure 3.1 in the manuscript) showing the estimated genetic prediction accuracy by values of the penalty parameter (in logarithmic scale). The rightmost point in the plot corresponds to the G-BLUP model (obtained when $\lambda = 0$). The point at the peak denotes the maximum accuracy that was obtained by the SSI.

```
load("output/results_accuracy.RData")

dat <- data.frame(do.call(rbind,lapply(split(accSSI,accSSI$SSI),
         function(x) apply(x[,-c(1:2)],2,mean))))
dat$Model <- unlist(lapply(strsplit(rownames(dat),"\\."),function(x)x[1]))

dat2 <- do.call(rbind,lapply(split(dat,dat$Mod),function(x)x[which.max(x$acc),]))
ggplot(dat[dat$df>1,],aes(-log(lambda),accuracy)) +
   geom_hline(yintercept=dat["GBLUP",]$accuracy, linetype="dashed") +
   geom_line(aes(color=Model),size=1.1) + theme_bw() +
   geom_point(data=dat2,aes(color=Model),size=2.5)
```

B.2.7 Cross-validating to obtain an optimal penalization

The snippet below can be used to perform, within each trn-tst partition, k-folds CV to get an 'optimal' value of λ within the training data, and then used to fit an SSI for the testing set. The CV is implemented using the 'SSI_CV' function from the SFSI R-package for one 5-folds CV, this can be set by changing the 'nCV' and 'nFolds' parameters.

```
load("data/geno_pheno.RData");     load("output/varComps.RData")
load("output/partitions.RData");   load("output/results_accuracy.RData")

lambdaCV <- accSSI_CV <- dfCV <- c()   # Objects to store results

for(k in 1:ncol(partitions))
{   cat("  partition = ",k,"\n")
    tst <- partitions[,k]
    trn <- (1:length(y))[-tst]

    # Cross-validating the training set
    fm1 <- SSI_CV(y,K=G,trn.CV=trn,nLambda=100,mc.cores=1,nFolds=5,nCV=1)
    lambdaCV[k] <- summary(fm1)$optCOR["mean","lambda"]

    # Fit a SSI with the estimated lambda
    fm2 <- SSI(y,K=G,b=mu[k],h2=h2[k],trn=trn,tst=tst,lambda=lambdaCV[k])

    accSSI_CV[k] <- summary(fm2)$accuracy/sqrt(fm0$h2)
    dfCV <- cbind(dfCV, fm2$df)
}
save(accSSI_CV,lambdaCV,dfCV,file="output/results_accuracyCV.RData")
```

After running the above analysis, the following snippet can be run to create a plot (as in Figure 3.2 in the manuscript) comparing partition-wise the accuracy of the optimal SSI with that of the G-BLUP. The average accuracies are also shown in the plot.

```
load("output/results_accuracy.RData")
load("output/results_accuracyCV.RData")

dat <- data.frame(GBLUP=accSSI[accSSI$SSI=="GBLUP",]$acc,SSI=accSSI_CV)
rg <- range(dat)
tmp <- c(mean(rg),diff(rg)*0.4)

ggplot(dat,aes(GBLUP,SSI)) + geom_abline(slope=1,linetype="dotted") +
  geom_point(shape=21,color="orange") + xlim(rg) + ylim(rg) +
  annotate("text",tmp[1],tmp[1]-tmp[2],label=round(mean(dat$GBLUP),3)) +
  annotate("text",tmp[1]-tmp[2],tmp[1],label=round(mean(dat$SSI),3))
```

The code below creates a plot (as in Figure 3.3 in the manuscript) showing the distribution of the number of points in the support set for the SSI, across all partitions.

```
load("output/results_accuracyCV.RData")

dat <- data.frame(df=as.vector(dfCV))

bw <- round(diff(range(dat$df))/40)
ggplot(data=dat,aes(df,stat(count)/length(dfCV))) +  theme_bw() +
    geom_histogram(color="gray20",fill="lightblue",binwidth=bw) +
    labs(x=bquote("Support set size(" *n[sup]*")"),y="Frequency")
```

B.2.8 Subject-specific training sets

The next script can be used to create a plot (as in Figure 3.4 in the manuscript) showing (for a single trn-tst partition) the subset of points in the support set, for each individual being predicted. This plot can be made through the 'plotNet' function from the SFSI package.

```
# Load data
load("data/geno_pheno.RData")
load("output/partitions.RData")
load("output/results_accuracyCV.RData")

part <- 1      # Choose any partition from 1,...,nPart
tst <- partitions[,part]
trn <- (1:length(y))[-tst]

# Fit SSI with lambda previously estimated using CV
fm <- SSI(y,K=G,trn=trn,tst=tst,lambda=lambdaCV[part])

plotNet(fm,K=G,tst=fm$tst[1:25],single=FALSE,title=NULL,bg.col="white")
```

APPENDIX C

SUPPLEMENTARY FIGURES AND TABLES FROM CHAPTER 4

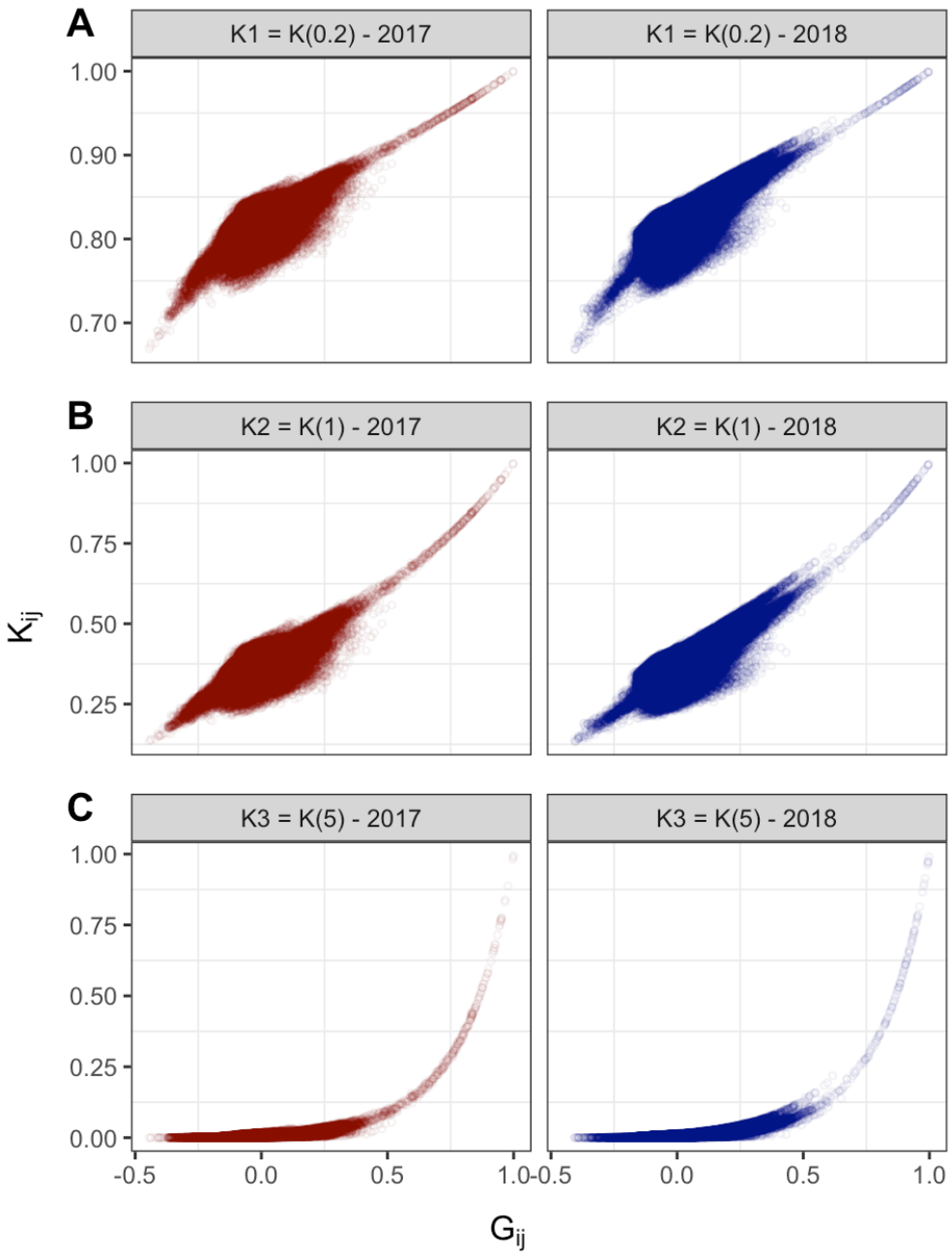

Figure C.1: Genomic relationships (G_{ij}) versus kernel relationships (K_{ij}) of individuals in cycle 2019 with those in cycles 2017 (left) and 2018 (right). G_ij and K_ij are the ij^{th} element of $\mathbf{G}$ and $\mathbf{K}(\theta)$, respectively. (A) $\mathbf{K}_1 = \mathbf{K}(0.2)$. (B) $\mathbf{K}_2 = \mathbf{K}(1)$. (C) $\mathbf{K}_3 = \mathbf{K}(5)$.

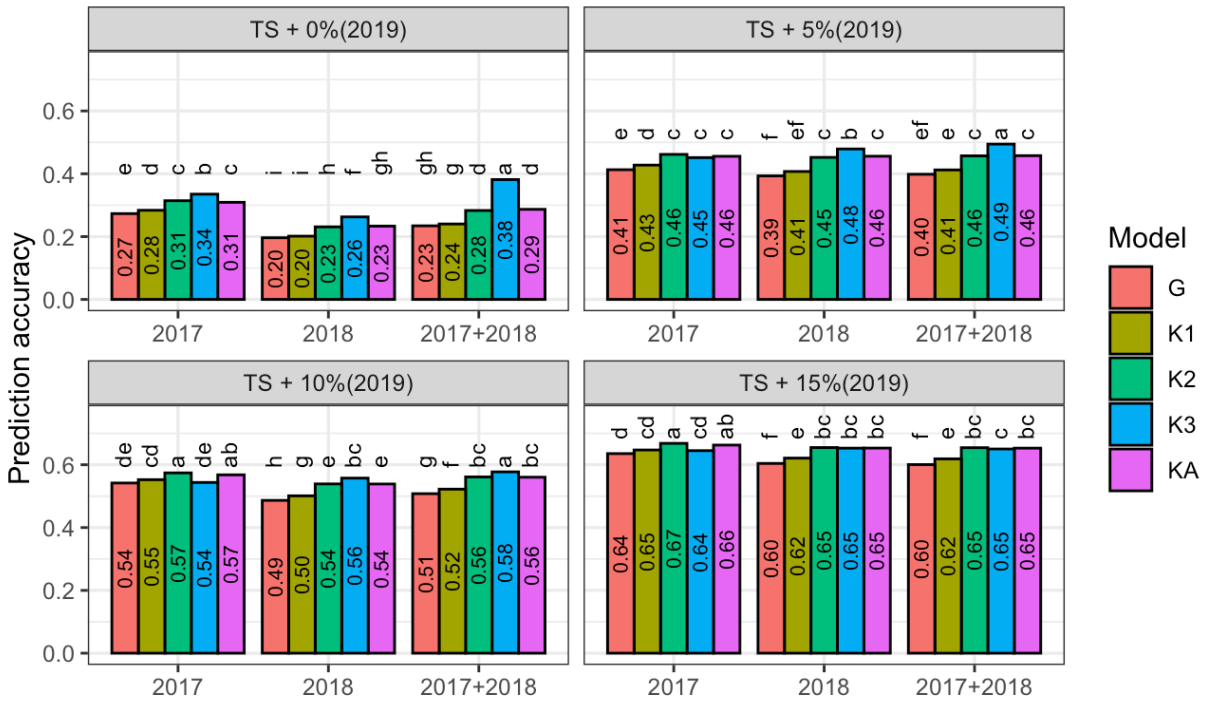

Figure C.2: Prediction accuracy by model and training set (TS). TSs consisted on all the data from the 2017, 2018, or 2017+2018 cycles alone (top-left panel), or in combination with a proportion (5%, 10%, 15%) of the data from the 2019 cycle. The prediction set consisted of 612 genotypes from the 2019 cycle that were not used for model training. Models with the same letter within panel indicate no significant difference from each other (α = 0.05, ANOVA followed by Tukey test). GY-DRT trait-environment combination.

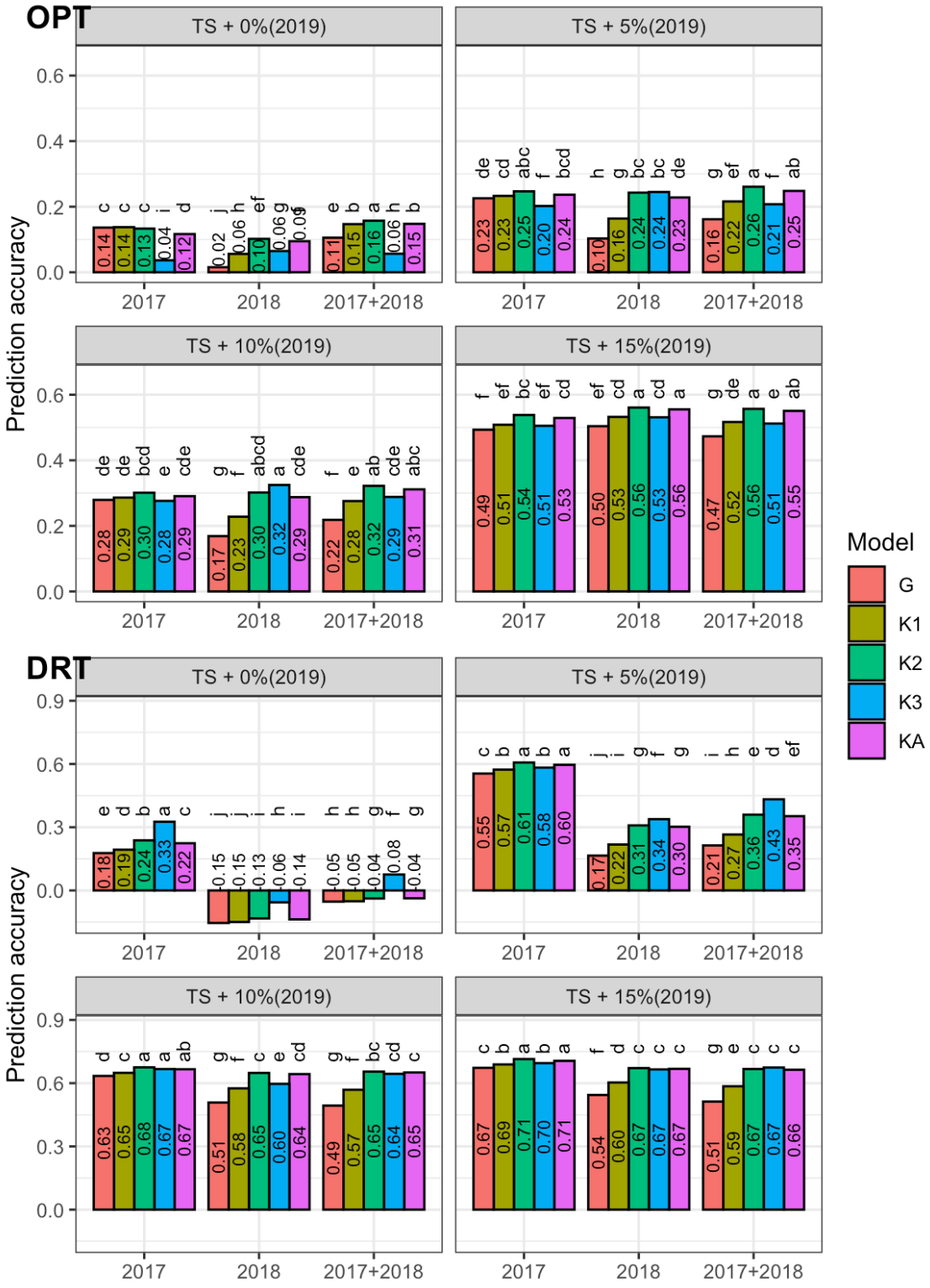

Figure C.3: Prediction accuracy by model and training set (TS). TSs consisted on all the data from the 2017, 2018, or 2017+2018 cycles alone (top-left panel), or in combination with a proportion (5%, 10%, 15%) of the data from the 2019 cycle. The prediction set consisted of 612 genotypes from the 2019 cycle that were not used for model training. Models with the same letter within panel indicate no significant difference from each other (α = 0.05, ANOVA followed by Tukey test). Trait PH (OPT and DRT environments).

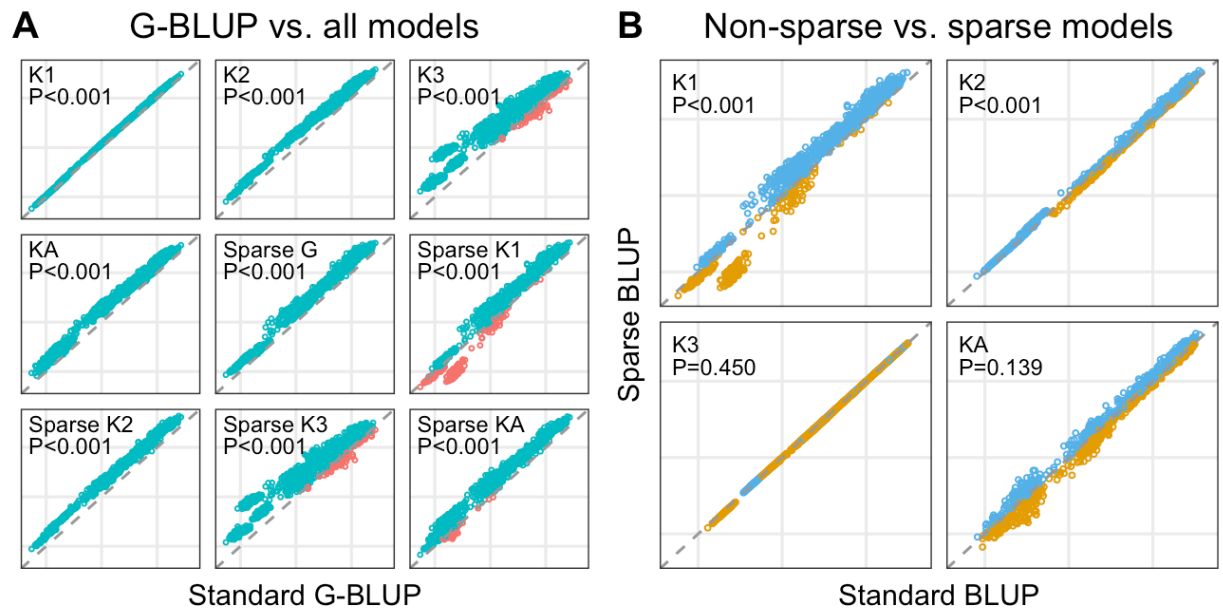

Figure C.4: (A) Prediction accuracy of the standard (non-sparse) G-BLUP model (horizontal axis) versus the prediction accuracy of all other models (vertical axis of each panel). (B) Prediction accuracy of the standard *-BLUP model (horizontal axis) versus the prediction accuracy of its sparse version (vertical axis), by type of kernel used in panels. Each point represent a training-testing partition within each training set composition. Colored points above (below) the 45 degree line represent cases for which one model outperformed the other model. P: p-value for the test (from ANOVA) for differences in accuracy between the two models. GY-DRT trait-environment combination.

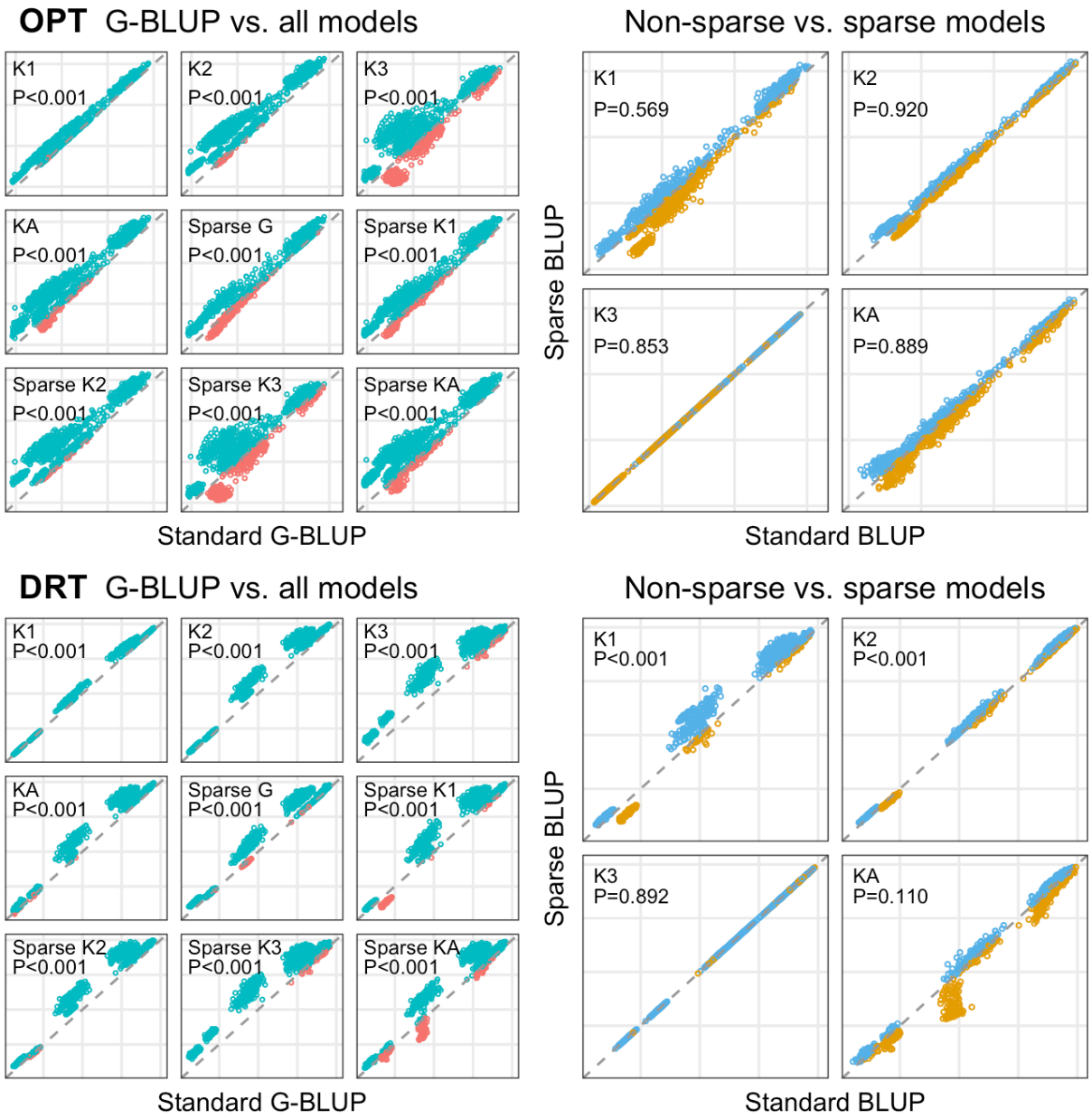

Figure C.5: (left) Prediction accuracy of the standard (non-sparse) G-BLUP model (horizontal axis) versus the prediction accuracy of all other models (vertical axis of each panel). (right) Prediction accuracy of the standard *-BLUP model (horizontal axis) versus the prediction accuracy of its sparse version (vertical axis), by type of kernel used in panels. Each point represent a training-testing partition within each training set composition. Colored points above (below) the 45 degree line represent cases for which one model outperformed the other model. P: p-value for the test (from ANOVA) for differences in accuracy between the two models. Trait PH (OPT and DRT environments).

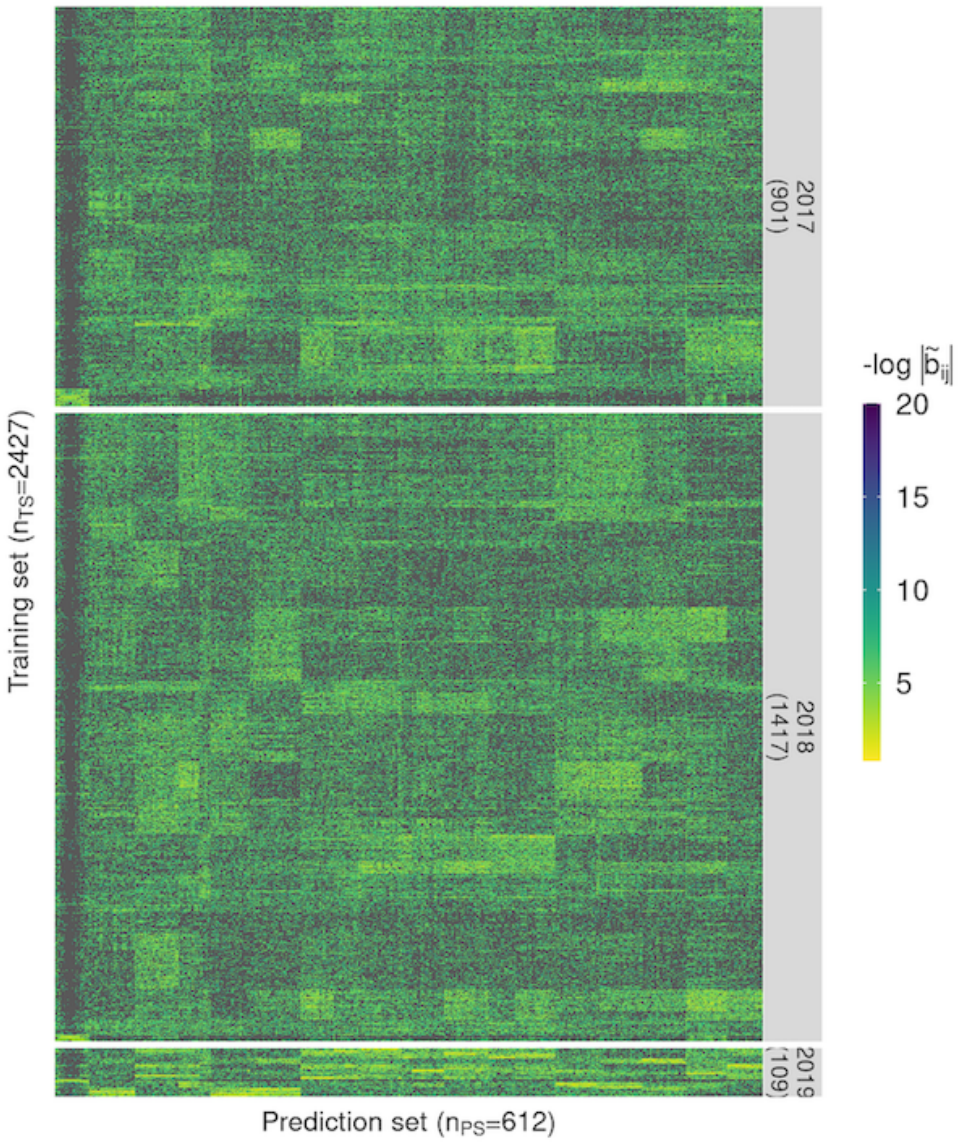

Figure C.6: Heatmap of the coefficients in the Hat matrix ($\tilde{\mathbf{B}}(\lambda)_K$) of the sparse KA-BLUP model for one training-prediction (TS-PS) partition in the prediction of $n_{PS} = 612$ individuals from 2019 using $n_{TS} = 2427$ individuals (2017+2018 plus 15% of the 2019 set). Predicted individuals are presented in columns and training individuals are presented in rows separated by cycle and number of individuals in parentheses. The value of λ was obtained by cross-validation. Each column represents values of the vector $\tilde{\mathbf{b}}(\lambda)_{i_K} = \{\tilde{b}_{ij}\}$, $j = 1, ..., 2427$ (Equation (4.6)). Individuals no contributing to the prediction have a coefficient $\tilde{b}_{ij} = 0$ represented in grey color. Individuals with a non-zero coefficient are shown in a yellow-blue logarithm scale (in the original scale, yellow indicates large values and blue indicates small value). GY-OPT trait-environment combination.

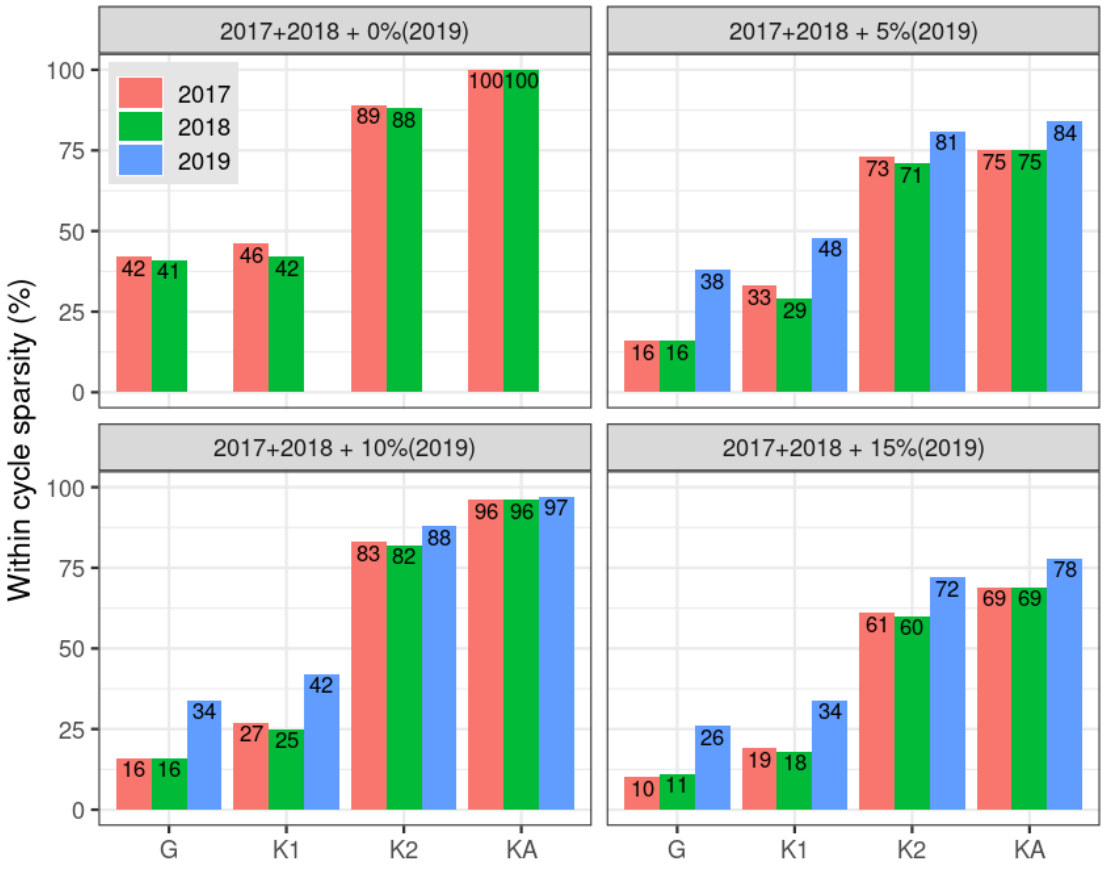

Figure C.7: Proportion of the training individuals from each cycle that contributed to the prediction of the 612 testing genotypes from 2019, using sparse models with different relationship matrices (horizontal axis): $\mathbf{G}$, $\mathbf{K}_1$, $\mathbf{K}_2$, or $\mathbf{K}_A$. The training set was composed by individuals from 2017 (n = 901) and 2018 (n = 1417) alone (top-left panel) or in combination with a proportion (5%, 10%, 15%) of the data from the 2019 cycle. GY-DRT trait-environment combination.

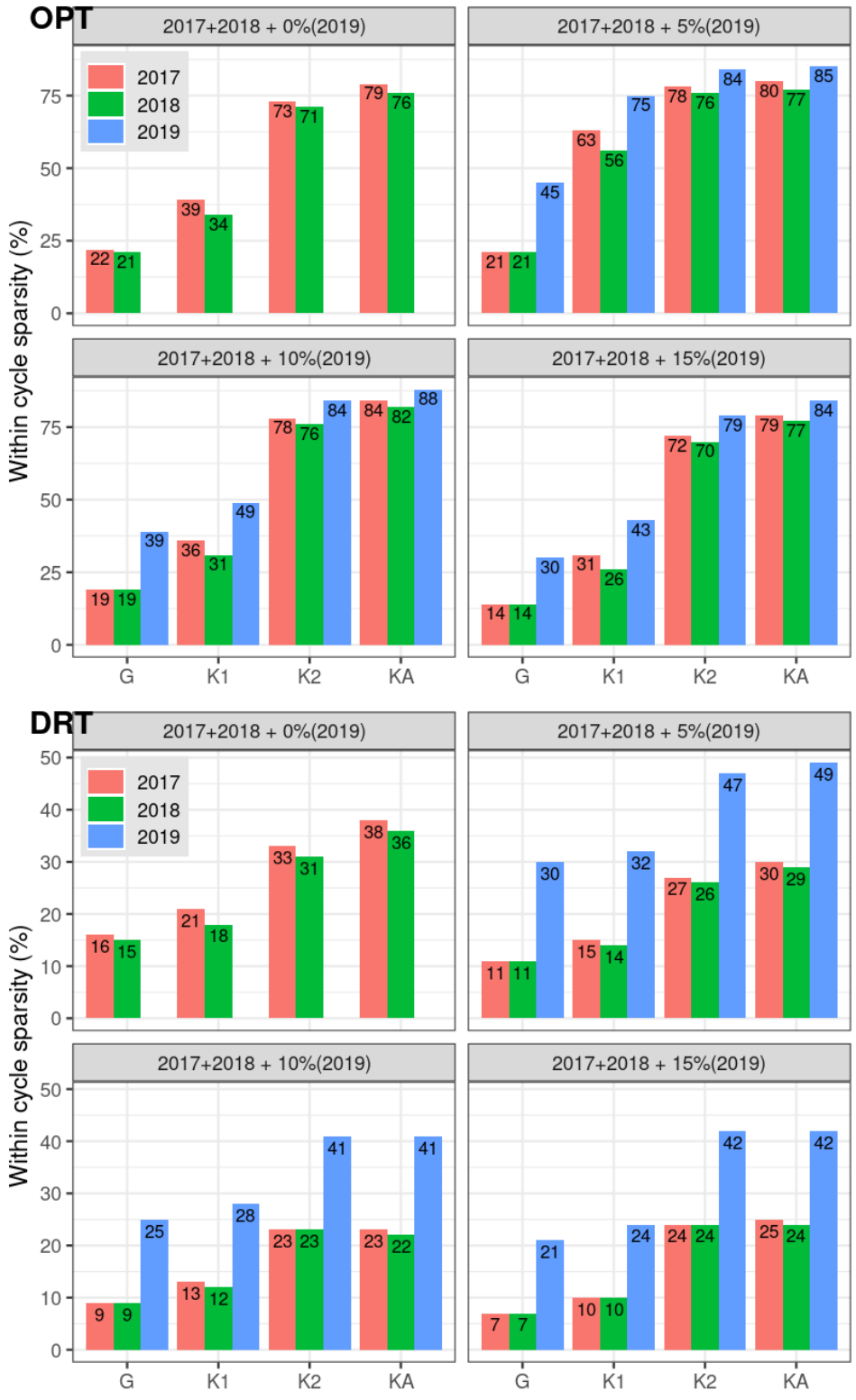

Figure C.8: Proportion of the training individuals from each cycle that contributed to the prediction of the 612 testing genotypes from 2019, using sparse models with different relationship matrices (horizontal axis): $\mathbf{G}$, $\mathbf{K}_1$, $\mathbf{K}_2$, or $\mathbf{K}_A$. The training set was composed by individuals from 2017 (n = 901) and 2018 (n = 1417) alone (top-left panels) or in combination with a proportion (5%, 10%, 15%) of the data from the 2019 cycle. Trait PH (OPT and DRT environments).

Table C.1: Heritability and accuracy of prediction for each training set (TS) composition (including 0%, 5%, and 10% of subjects from the 2019 cycle), trait GY, environment OPT.

						Accuracy (SD)		% Gain	
	TS (n_{TS})	**GRM**	λ_{CV}[a]	n_{sup} **(RS)**[b]	h^2	**Standard**	**Sparse**	**I**[c]	**II**[d]
TS+0%(2019)	2017 (901)	G	0.0139	271 (30)	0.51	0.20 (0.017)	0.25 (0.018)	0	24
		K1	0.0028	307 (34)	0.86	0.21 (0.017)	0.20 (0.017)	5	-9
		K2	0.0037	388 (43)	0.73	0.24 (0.017)	0.26 (0.017)	17	7
		K3	0.0000	901 (100)	0.96	0.26 (0.016)	0.26 (0.016)	28	0
		KA	0.0023	435 (48)	0.86	0.24 (0.019)	0.26 (0.029)	19	6
	2018 (1417)	G	0.0086	469 (33)	0.61	0.25 (0.015)	0.22 (0.015)	0	-11
		K1	0.0011	664 (47)	0.91	0.25 (0.015)	0.22 (0.015)	2	-13
		K2	0.0019	743 (52)	0.79	0.26 (0.015)	0.24 (0.016)	5	-8
		K3	0.0000	1417 (100)	0.92	0.23 (0.015)	0.23 (0.015)	-7	0
		KA	0.0015	680 (48)	0.88	0.26 (0.016)	0.24 (0.017)	4	-8
	2017+18 (2318)	G	0.0120	569 (25)	0.52	0.33 (0.015)	0.33 (0.015)	0	2
		K1	0.0028	484 (21)	0.88	0.33 (0.015)	0.31 (0.015)	2	-6
		K2	0.0013	1374 (59)	0.76	0.34 (0.015)	0.33 (0.015)	4	-2
		K3	0.0000	2316 (100)	0.88	0.31 (0.015)	0.31 (0.015)	-4	0
		KA	0.0010	1446 (62)	0.83	0.33 (0.016)	0.33 (0.016)	3	-3
TS+5%(2019)	2017 (938)	G	0.0149	241 (26)	0.50	0.32 (0.036)	0.36 (0.035)	0	12
		K1	0.0027	292 (31)	0.86	0.34 (0.036)	0.36 (0.036)	5	8
		K2	0.0028	469 (50)	0.73	0.37 (0.037)	0.38 (0.038)	16	2
		K3	0.0000	938 (100)	0.93	0.34 (0.038)	0.34 (0.038)	7	0
		KA	0.0015	550 (59)	0.84	0.36 (0.038)	0.37 (0.038)	14	1
	2018 (1454)	G	0.0097	412 (28)	0.61	0.26 (0.021)	0.24 (0.020)	0	-7
		K1	0.0014	563 (39)	0.91	0.27 (0.021)	0.26 (0.024)	2	-3
		K2	0.0015	835 (57)	0.79	0.28 (0.021)	0.27 (0.021)	7	-3
		K3	0.0000	1454 (100)	0.92	0.28 (0.021)	0.28 (0.021)	6	0
		KA	0.0012	749 (52)	0.89	0.28 (0.021)	0.26 (0.024)	6	-6
	2017+18 (2355)	G	0.0142	441 (19)	0.53	0.34 (0.019)	0.36 (0.019)	0	4
		K1	0.0025	534 (23)	0.88	0.35 (0.019)	0.34 (0.024)	3	-4
		K2	0.0018	1117 (47)	0.76	0.37 (0.021)	0.37 (0.021)	9	-1
		K3	0.0000	2354 (100)	0.88	0.36 (0.022)	0.36 (0.022)	4	0
		KA	0.0013	1235 (52)	0.83	0.37 (0.021)	0.37 (0.020)	8	-2
TS+10%(2019)	2017 (974)	G	0.0186	183 (19)	0.52	0.42 (0.029)	0.45 (0.030)	0	7
		K1	0.0041	180 (18)	0.87	0.44 (0.029)	0.46 (0.030)	4	4
		K2	0.0067	218 (22)	0.75	0.47 (0.030)	0.47 (0.030)	12	1
		K3	0.0000	974 (100)	0.92	0.48 (0.032)	0.48 (0.032)	14	0
		KA	0.0045	244 (25)	0.85	0.46 (0.030)	0.47 (0.032)	10	1
	2018 (1417)	G	0.0111	349 (23)	0.61	0.40 (0.029)	0.41 (0.030)	0	2
		K1	0.0017	480 (32)	0.91	0.41 (0.029)	0.42 (0.030)	3	2

Table C.1 (cont'd)

2018	K2	0.0017	781 (52)	0.79	0.44 (0.029)	0.43 (0.029)	10	-1
(1490)	K3	0.0000	1490 (100)	0.90	0.42 (0.028)	0.42 (0.028)	5	0
	KA	0.0012	781 (52)	0.88	0.43 (0.029)	0.42 (0.029)	8	-1
	G	0.0126	455 (19)	0.53	0.42 (0.021)	0.44 (0.021)	0	4
	K1	0.0027	459 (19)	0.88	0.44 (0.021)	0.44 (0.026)	4	1
2017+18	K2	0.0021	1032 (43)	0.76	0.46 (0.023)	0.46 (0.023)	10	-1
(2391)	K3	0.0000	2390 (100)	0.87	0.46 (0.025)	0.46 (0.025)	9	0
	KA	0.0015	1153 (48)	0.83	0.46 (0.023)	0.46 (0.023)	9	-1

GRM: Genetic relationship matrix. SD: standard deviation. [a]Penalization parameter in Equation (4.6) found by cross-validating the TS. [b]n_{sup}=average number of individuals from the TS with a non-zero coefficient in the sparse Hat matrix (support set). RS: relative sparsity ($100n_{TS}/n_{sup}$). In the standard models λ_{CV} is equal to zero and n_{sup} is equal to the total TS size. Within each TS cycle, percentage of gain in accuracy of the [c]standard K-BLUP relative to the standard G-BLUP, and [d]sparse *-BLUP relative to the standard *-BLUP (*=G- or K-).

Table C.2: Heritability and accuracy of prediction for each training set (TS) composition (including 0%, 5%, 10%, and 15% of subjects from the 2019 cycle), trait GY, environment DRT.

						Accuracy (SD)		% Gain	
	TS (n_{TS})	**GRM**	λ_{CV}[a]	n_{sup} **(RS)**[b]	h^2	**Standard**	**Sparse**	**I**[c]	**II**[d]
TS+0%(2019)	2017 (901)	G	0.0071	418 (46)	0.57	0.27 (0.014)	0.29 (0.014)	0	6
		K1	0.0012	522 (58)	0.89	0.28 (0.014)	0.20 (0.020)	4	-30
		K2	0.0015	597 (66)	0.77	0.31 (0.014)	0.33 (0.014)	15	4
		K3	0.0000	901 (100)	0.94	0.34 (0.013)	0.34 (0.013)	23	0
		KA	0.0008	680 (75)	0.88	0.31 (0.018)	0.30 (0.033)	13	-2
	2018 (1417)	G	0.0075	561 (40)	0.41	0.20 (0.018)	0.23 (0.018)	0	17
		K1	0.0011	768 (54)	0.81	0.20 (0.018)	0.19 (0.017)	3	-8
		K2	0.0001	1387 (98)	0.60	0.23 (0.018)	0.24 (0.018)	18	2
		K3	0.0000	1417 (100)	0.66	0.26 (0.013)	0.26 (0.014)	34	-2
		KA	0.0000	1417 (100)	0.69	0.23 (0.020)	0.23 (0.021)	19	-3
	2017+18 (2318)	G	0.0066	954 (41)	0.44	0.23 (0.018)	0.27 (0.018)	0	14
		K1	0.0013	1013 (44)	0.84	0.24 (0.018)	0.25 (0.018)	2	5
		K2	0.0003	2049 (88)	0.65	0.28 (0.017)	0.29 (0.017)	21	1
		K3	0.0000	2318 (100)	0.75	0.38 (0.013)	0.38 (0.013)	63	0
		KA	0.0000	2318 (100)	0.73	0.29 (0.019)	0.29 (0.029)	22	2
TS+5%(2019)	2017 (938)	G	0.0151	243 (26)	0.57	0.41 (0.028)	0.46 (0.027)	0	11
		K1	0.0031	331 (35)	0.89	0.43 (0.028)	0.44 (0.053)	4	2
		K2	0.0035	484 (52)	0.77	0.46 (0.026)	0.46 (0.026)	12	-1
		K3	0.0000	938 (100)	0.92	0.45 (0.021)	0.45 (0.021)	9	0
		KA	0.0010	659 (70)	0.87	0.46 (0.027)	0.45 (0.033)	10	-1
	2018 (1454)	G	0.0126	373 (26)	0.41	0.39 (0.034)	0.43 (0.031)	0	10
		K1	0.0013	678 (47)	0.82	0.41 (0.033)	0.42 (0.035)	4	4
		K2	0.0000	1453 (100)	0.61	0.45 (0.030)	0.45 (0.032)	15	0
		K3	0.0000	1454 (100)	0.67	0.48 (0.026)	0.48 (0.026)	22	-1
		KA	0.0000	1454 (100)	0.70	0.46 (0.029)	0.45 (0.031)	16	0
	2017+18 (2355)	G	0.0163	387 (16)	0.44	0.40 (0.035)	0.46 (0.024)	0	16
		K1	0.0020	728 (31)	0.84	0.41 (0.034)	0.46 (0.030)	3	11
		K2	0.0008	1696 (72)	0.65	0.46 (0.031)	0.45 (0.034)	15	-1
		K3	0.0000	2355 (100)	0.76	0.49 (0.020)	0.49 (0.020)	24	0
		KA	0.0006	1773 (75)	0.73	0.46 (0.031)	0.46 (0.038)	15	0
TS+10%(2019)	2017 (974)	G	0.0174	190 (19)	0.58	0.54 (0.033)	0.58 (0.030)	0	7
		K1	0.0038	184 (19)	0.89	0.55 (0.032)	0.58 (0.028)	2	4
		K2	0.0064	229 (23)	0.78	0.57 (0.029)	0.59 (0.029)	6	2
		K3	0.0000	974 (100)	0.92	0.54 (0.031)	0.54 (0.031)	0	0
		KA	0.0040	284 (29)	0.88	0.57 (0.029)	0.58 (0.031)	5	2
		G	0.0124	354 (24)	0.41	0.49 (0.026)	0.53 (0.025)	0	9
		K1	0.0020	488 (33)	0.82	0.50 (0.026)	0.53 (0.027)	3	6

Table C.2 (cont'd)

TS+15%(2019)	2018 (1490)	K2	0.0002	1416 (95)	0.61	0.54 (0.024)	0.54 (0.025)	11	0
		K3	0.0000	1490 (100)	0.68	0.56 (0.021)	0.56 (0.022)	15	0
		KA	0.0000	1482 (99)	0.70	0.54 (0.024)	0.54 (0.026)	11	0
	2017+18 (2391)	G	0.0152	389 (16)	0.45	0.51 (0.030)	0.56 (0.028)	0	10
		K1	0.0022	622 (26)	0.84	0.52 (0.029)	0.55 (0.028)	3	6
		K2	0.0005	1981 (83)	0.66	0.56 (0.026)	0.56 (0.027)	10	0
		K3	0.0000	2391 (100)	0.76	0.58 (0.021)	0.58 (0.021)	14	0
		KA	0.0001	2292 (96)	0.73	0.56 (0.026)	0.56 (0.027)	10	0
	2017 (1010)	G	0.0236	126 (12)	0.58	0.64 (0.029)	0.68 (0.025)	0	7
		K1	0.0058	125 (12)	0.89	0.65 (0.028)	0.68 (0.025)	2	4
		K2	0.0097	125 (12)	0.79	0.67 (0.026)	0.68 (0.023)	5	2
		K3	0.0000	1010 (100)	0.92	0.64 (0.025)	0.64 (0.025)	1	0
		KA	0.0061	145 (14)	0.87	0.66 (0.026)	0.68 (0.024)	4	3
	2018 (1526)	G	0.0126	343 (22)	0.42	0.60 (0.026)	0.65 (0.030)	0	8
		K1	0.0013	705 (46)	0.83	0.62 (0.026)	0.64 (0.028)	3	3
		K2	0.0001	1491 (98)	0.62	0.65 (0.025)	0.65 (0.025)	8	0
		K3	0.0000	1526 (100)	0.69	0.65 (0.024)	0.65 (0.024)	8	0
		KA	0.0000	1518 (99)	0.71	0.65 (0.026)	0.65 (0.026)	8	0
	2017+18 (2464)	G	0.0175	273 (11)	0.45	0.60 (0.030)	0.67 (0.026)	0	11
		K1	0.0028	458 (19)	0.84	0.62 (0.029)	0.66 (0.026)	3	7
		K2	0.0011	1488 (61)	0.66	0.65 (0.027)	0.66 (0.026)	9	1
		K3	0.0000	2427 (100)	0.76	0.65 (0.024)	0.65 (0.024)	8	0
		KA	0.0007	1682 (69)	0.74	0.65 (0.027)	0.66 (0.026)	9	0

GRM: Genetic relationship matrix. SD: standard deviation. [a]Penalization parameter in Equation (4.6) found by cross-validating the TS. [b]n_{sup}=average number of individuals from the TS with a non-zero coefficient in the sparse Hat matrix (support set). RS: relative sparsity ($100n_{TS}/n_{sup}$). In the standard models λ_{CV} is equal to zero and n_{sup} is equal to the total TS size. Within each TS cycle, percentage of gain in accuracy of the [c]standard K-BLUP relative to the standard G-BLUP, and [d]sparse *-BLUP relative to the standard *-BLUP (*=G- or K-).

Table C.3: Heritability and accuracy of prediction for each training set (TS) composition (including 0%, 5%, and 10% of subjects from the 2019 cycle), trait PH, environment OPT.

						Accuracy (SD)		% Gain	
	TS (n_{TS})	**GRM**	λ_{CV}[a]	n_{sup} **(RS)**[b]	h^2	**Standard**	**Sparse**	**I**[c]	**II**[d]
TS+0%(2019)	2017 (901)	G	0.0024	670 (74)	0.55	0.14 (0.014)	0.12 (0.014)	0	-9
		K1	0.0005	735 (82)	0.88	0.14 (0.014)	0.13 (0.015)	1	-8
		K2	0.0006	763 (85)	0.76	0.13 (0.014)	0.14 (0.014)	-2	3
		K3	0.0000	901 (100)	0.96	0.04 (0.012)	0.04 (0.012)	-73	-3
		KA	0.0002	857 (95)	0.87	0.12 (0.019)	0.11 (0.028)	-14	-3
	2018 (1417)	G	0.0131	317 (22)	0.71	0.02 (0.015)	0.08 (0.015)	0	430
		K1	0.0020	387 (27)	0.95	0.06 (0.015)	0.08 (0.015)	265	43
		K2	0.0024	615 (43)	0.87	0.10 (0.015)	0.12 (0.014)	564	22
		K3	0.0000	1417 (100)	0.93	0.06 (0.013)	0.06 (0.013)	322	-1
		KA	0.0019	545 (38)	0.93	0.09 (0.021)	0.13 (0.017)	517	33
	2017+18 (2318)	G	0.0129	499 (22)	0.64	0.11 (0.013)	0.07 (0.014)	0	-38
		K1	0.0014	835 (36)	0.93	0.15 (0.013)	0.07 (0.013)	39	-51
		K2	0.0008	1660 (72)	0.84	0.16 (0.013)	0.13 (0.014)	49	-19
		K3	0.0000	2318 (100)	0.94	0.06 (0.013)	0.06 (0.013)	-46	-1
		KA	0.0005	1786 (77)	0.89	0.15 (0.018)	0.14 (0.024)	40	-8
TS+5%(2019)	2017 (938)	G	0.0010	819 (87)	0.55	0.23 (0.031)	0.22 (0.031)	0	-1
		K1	0.0002	859 (92)	0.88	0.23 (0.031)	0.23 (0.032)	3	-3
		K2	0.0003	860 (92)	0.75	0.25 (0.031)	0.24 (0.031)	9	-2
		K3	0.0000	938 (100)	0.92	0.20 (0.032)	0.20 (0.032)	-10	0
		KA	0.0002	882 (94)	0.87	0.24 (0.033)	0.22 (0.035)	5	-5
	2018 (1454)	G	0.0103	373 (26)	0.70	0.10 (0.037)	0.18 (0.036)	0	76
		K1	0.0016	458 (31)	0.94	0.16 (0.036)	0.19 (0.036)	59	17
		K2	0.0018	714 (49)	0.87	0.24 (0.032)	0.25 (0.032)	136	5
		K3	0.0000	1454 (100)	0.92	0.24 (0.027)	0.24 (0.027)	138	0
		KA	0.0013	682 (47)	0.92	0.23 (0.034)	0.24 (0.033)	121	5
	2017+18 (2355)	G	0.0121	500 (21)	0.63	0.16 (0.024)	0.14 (0.028)	0	-11
		K1	0.0006	1388 (59)	0.92	0.22 (0.025)	0.17 (0.027)	34	-19
		K2	0.0006	1815 (77)	0.84	0.26 (0.026)	0.25 (0.026)	61	-4
		K3	0.0000	2355 (100)	0.92	0.21 (0.027)	0.21 (0.027)	28	0
		KA	0.0004	1847 (78)	0.89	0.25 (0.028)	0.24 (0.026)	53	-5
TS+10%(2019)	2017 (974)	G	0.0026	706 (72)	0.55	0.28 (0.043)	0.28 (0.044)	0	-1
		K1	0.0004	817 (84)	0.88	0.29 (0.042)	0.28 (0.044)	3	-2
		K2	0.0004	857 (88)	0.76	0.30 (0.043)	0.30 (0.043)	8	-1
		K3	0.0000	974 (100)	0.92	0.28 (0.046)	0.28 (0.046)	-1	0
		KA	0.0002	907 (93)	0.86	0.29 (0.042)	0.28 (0.043)	4	-2
		G	0.0128	291 (20)	0.70	0.17 (0.072)	0.25 (0.065)	0	48
		K1	0.0020	363 (24)	0.94	0.23 (0.066)	0.25 (0.066)	35	11

Table C.3 (cont'd)

2018	K2	0.0022	624 (42)	0.87	0.30 (0.054)	0.31 (0.055)	79	2
(1490)	K3	0.0000	1490 (100)	0.92	0.32 (0.043)	0.32 (0.043)	92	0
	KA	0.0019	540 (36)	0.92	0.29 (0.057)	0.30 (0.055)	70	4
	G	0.0117	468 (20)	0.63	0.22 (0.047)	0.22 (0.055)	0	0
	K1	0.0013	798 (33)	0.92	0.28 (0.044)	0.25 (0.053)	26	-8
2017+18	K2	0.0005	1846 (77)	0.84	0.32 (0.042)	0.31 (0.042)	47	-3
(2391)	K3	0.0000	2391 (100)	0.92	0.29 (0.042)	0.29 (0.042)	32	0
	KA	0.0003	1982 (83)	0.89	0.31 (0.043)	0.30 (0.044)	43	-3

GRM: Genetic relationship matrix. SD: standard deviation. [a]Penalization parameter in Equation (4.6) found by cross-validating the TS. [b]n_{sup}=average number of individuals from the TS with a non-zero coefficient in the sparse Hat matrix (support set). RS: relative sparsity ($100n_{TS}/n_{sup}$). In the standard models λ_{CV} is equal to zero and n_{sup} is equal to the total TS size. Within each TS cycle, percentage of gain in accuracy of the [c]standard K-BLUP relative to the standard G-BLUP, and [d]sparse *-BLUP relative to the standard *-BLUP (*=G- or K-).

Table C.4: Heritability and accuracy of prediction for each training set (TS) composition (including 0%, 5%, 10%, and 15% of subjects from the 2019 cycle), trait PH, environment DRT.

TS (n_{TS})		GRM	λ_{CV}[a]	n_{sup} (RS)[b]	h^2	Accuracy (SD)		% Gain	
						Standard	Sparse	I[c]	II[d]
TS+0%(2019)	2017 (901)	G	0.0025	658 (73)	0.64	0.18 (0.013)	0.17 (0.013)	0	-4
		K1	0.0005	718 (80)	0.92	0.19 (0.013)	0.34 (0.012)	9	78
		K2	0.0003	842 (93)	0.84	0.24 (0.013)	0.25 (0.013)	34	4
		K3	0.0000	901 (100)	1.00	0.33 (0.012)	0.33 (0.012)	84	1
		KA	0.0000	901 (100)	0.92	0.22 (0.024)	0.11 (0.063)	26	-51
	2018 (1417)	G	0.0175	263 (19)	0.59	-0.15 (0.017)	-0.11 (0.017)	0	-29
		K1	0.0028	319 (23)	0.91	-0.15 (0.017)	-0.13 (0.017)	-3	-13
		K2	0.0027	589 (42)	0.81	-0.13 (0.018)	-0.12 (0.018)	-14	-9
		K3	0.0000	1417 (100)	0.96	-0.06 (0.019)	-0.06 (0.019)	-63	0
		KA	0.0017	675 (48)	0.87	-0.14 (0.021)	-0.13 (0.018)	-11	-8
	2017+18 (2318)	G	0.0178	363 (16)	0.58	-0.05 (0.017)	-0.04 (0.017)	0	-18
		K1	0.0028	447 (19)	0.91	-0.05 (0.017)	-0.11 (0.018)	-3	115
		K2	0.0033	741 (32)	0.82	-0.04 (0.017)	-0.05 (0.017)	-28	36
		K3	0.0000	2316 (100)	0.97	0.08 (0.019)	0.08 (0.019)	-243	0
		KA	0.0022	853 (37)	0.87	-0.04 (0.020)	-0.05 (0.026)	-30	39
TS+5%(2019)	2017 (938)	G	0.0025	664 (71)	0.66	0.55 (0.024)	0.56 (0.025)	0	2
		K1	0.0003	781 (83)	0.92	0.57 (0.023)	0.60 (0.021)	3	5
		K2	0.0001	907 (97)	0.86	0.61 (0.020)	0.60 (0.021)	9	0
		K3	0.0000	938 (100)	0.98	0.58 (0.030)	0.58 (0.029)	5	0
		KA	0.0000	919 (98)	0.93	0.60 (0.022)	0.57 (0.039)	7	-5
	2018 (1454)	G	0.0157	271 (19)	0.59	0.17 (0.040)	0.26 (0.044)	0	60
		K1	0.0030	286 (20)	0.91	0.22 (0.043)	0.29 (0.044)	32	34
		K2	0.0024	625 (43)	0.82	0.31 (0.045)	0.33 (0.045)	87	7
		K3	0.0000	1454 (100)	0.95	0.34 (0.037)	0.34 (0.037)	104	0
		KA	0.0018	658 (45)	0.87	0.30 (0.046)	0.32 (0.047)	83	5
	2017+18 (2355)	G	0.0214	261 (11)	0.58	0.21 (0.030)	0.34 (0.033)	0	59
		K1	0.0034	344 (15)	0.91	0.27 (0.032)	0.32 (0.080)	24	20
		K2	0.0037	636 (27)	0.82	0.36 (0.034)	0.36 (0.035)	68	1
		K3	0.0000	2355 (100)	0.96	0.43 (0.027)	0.43 (0.027)	102	0
		KA	0.0028	699 (30)	0.87	0.35 (0.037)	0.36 (0.038)	65	2
TS+10%(2019)	2017 (974)	G	0.0050	625 (64)	0.67	0.63 (0.025)	0.64 (0.028)	0	1
		K1	0.0006	769 (79)	0.93	0.65 (0.024)	0.65 (0.027)	2	0
		K2	0.0000	970 (99)	0.86	0.68 (0.023)	0.67 (0.025)	6	0
		K3	0.0000	974 (100)	0.96	0.67 (0.029)	0.67 (0.029)	5	0
		KA	0.0001	947 (97)	0.93	0.67 (0.024)	0.66 (0.028)	5	-1
		G	0.0184	211 (14)	0.59	0.51 (0.037)	0.62 (0.037)	0	23
		K1	0.0030	260 (17)	0.91	0.58 (0.033)	0.64 (0.035)	13	10

Table C.4 (cont'd)

TS+15%(2019)	2018 (1490)	K2	0.0023	628 (42)	0.83	0.65 (0.028)	0.66 (0.028)	28	2
		K3	0.0000	1490 (100)	0.94	0.60 (0.031)	0.60 (0.031)	17	0
		KA	0.0018	653 (44)	0.87	0.64 (0.029)	0.66 (0.030)	27	2
	2017+18 (2391)	G	0.0212	231 (10)	0.58	0.49 (0.030)	0.63 (0.027)	0	28
		K1	0.0034	309 (13)	0.91	0.57 (0.028)	0.65 (0.027)	15	13
		K2	0.0040	559 (23)	0.82	0.65 (0.023)	0.67 (0.023)	33	2
		K3	0.0000	2391 (100)	0.95	0.64 (0.022)	0.64 (0.022)	31	0
		KA	0.0033	556 (23)	0.87	0.65 (0.024)	0.67 (0.025)	32	3
	2017 (1010)	G	0.0055	487 (48)	0.66	0.67 (0.023)	0.69 (0.025)	0	3
		K1	0.0039	613 (61)	0.93	0.69 (0.022)	0.70 (0.020)	2	2
		K2	0.0006	804 (80)	0.86	0.71 (0.021)	0.72 (0.021)	6	0
		K3	0.0000	1010 (100)	0.97	0.70 (0.027)	0.69 (0.027)	3	0
		KA	0.0003	857 (85)	0.92	0.71 (0.022)	0.70 (0.025)	5	-1
	2018 (1526)	G	0.0198	173 (11)	0.59	0.54 (0.032)	0.65 (0.031)	0	20
		K1	0.0034	204 (13)	0.91	0.60 (0.029)	0.66 (0.027)	11	10
		K2	0.0025	567 (37)	0.83	0.67 (0.025)	0.68 (0.026)	23	2
		K3	0.0000	1526 (100)	0.94	0.67 (0.027)	0.67 (0.027)	22	0
		KA	0.0018	626 (41)	0.87	0.67 (0.025)	0.68 (0.026)	23	1
	2017+18 (2464)	G	0.0213	193 (8)	0.58	0.51 (0.028)	0.65 (0.025)	0	27
		K1	0.0038	253 (10)	0.91	0.59 (0.026)	0.66 (0.025)	14	13
		K2	0.0034	600 (25)	0.82	0.67 (0.023)	0.68 (0.023)	30	2
		K3	0.0000	2427 (100)	0.95	0.67 (0.022)	0.67 (0.022)	32	0
		KA	0.0028	618 (25)	0.87	0.66 (0.023)	0.67 (0.025)	29	2

GRM: Genetic relationship matrix. SD: standard deviation. [a]Penalization parameter in Equation (4.6) found by cross-validating the TS. [b]n_{sup}=average number of individuals from the TS with a non-zero coefficient in the sparse Hat matrix (support set). RS: relative sparsity ($100n_{TS}/n_{sup}$). In the standard models λ_{CV} is equal to zero and n_{sup} is equal to the total TS size. Within each TS cycle, percentage of gain in accuracy of the [c]standard K-BLUP relative to the standard G-BLUP, and [d]sparse *-BLUP relative to the standard *-BLUP (*=G- or K-).

BIBLIOGRAPHY

BIBLIOGRAPHY

Aguate, Fernando M., Samuel Trachsel, Lorena González-Pérez, Juan Burgueño, José Crossa, Mónica Balzarini, David Gouache, Matthieu Bogard & Gustavo de los Campos. 2017. Use of hyperspectral image data outperforms vegetation indices in prediction of maize yield. *Crop Science* 57. 2517–2524.

Akdemir, Deniz & Julio Isidro-Sanchez. 2019. Design of training populations for selective phenotyping in genomic prediction. *Scientific Reports* 9. 1–15.

Akdemir, Deniz, Julio I Sanchez & Jean-Luc Jannink. 2015. Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution* 47(38). 1–10.

Alvarado, Gregorio, Francisco M. Rodríguez, Angela Pacheco, Juan Burgueño, José Crossa, Mateo Vargas, Paulino Pérez-Rodríguez & Marco A. Lopez-Cruz. 2020. META-R: A software to analyze data from multi-environment plant breeding trials. *The Crop Journal* 8(5). 745–756.

Araus, Luis & Jill E Cairns. 2014. Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science* 19(1). 52–61.

Atanda, Sikiru Adeniyi, Michael Olsen, Juan Burgueño, Jose Crossa, Daniel Dzidzienyo, Yoseph Beyene, Manje Gowda, Kate Dreher, Xuecai Zhang, Boddupalli M. Prasanna, Pangirayi Tongoona, Eric Yirenkyi Danquah, Gbadebo Olaoye & Kelly R. Robbins. 2020. Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theoretical and Applied Genetics* .

Babar, M A, M P Reynolds, M Van Ginkel, A R Klatt, W R Raun & M L Stone. 2006. Spectral reflectance to estimate genetic variation for in-season biomass, leaf chlorophyll, and canopy temperature in wheat. *Crop Science* 46. 1046–1057.

Bates, Douglas, Martin Mächler, Benjamin M Bolker & Steven C Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Bernardo, Rex & Jianming Yu. 2007. Prospects for Genomewide Selection for Quantitative Traits in Maize. *Crop Science* 47. 1082–1090.

Beyene, Yoseph, Manje Gowda, Michael Olsen, Kelly R. Robbins, Paulino Pérez-Rodríguez, Gregorio Alvarado, Kate Dreher, Star Yanxin Gao, Stephen Mugo, Boddupalli M. Prasanna & Jose Crossa. 2019. Empirical Comparison of Tropical Maize Hybrids Selected Through Genomic and Phenotypic Selections. *Frontiers in Plant Science* 10(1502). 1–11.

Bulmer, M G. 1985. *The mathematical theory of quantitative genetics*. New York, USA: Oxford University Press.

de los Campos, G., D. Gianola & G. J. Rosa. 2009a. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of animal science* 87. 1883–1887.

de los Campos, Gustavo, Daniel Gianola, Guilherme J.M. Rosa, Kent A. Weigel & Jose Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research* 92(4). 295–308.

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

de los Campos, Gustavo, Hugo Naya, Daniel Gianola, José Crossa, Andrés Legarra, Eduardo Manfredi, Kent Weigel & José Miguel Cotes. 2009b. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182. 375–385.

de los Campos, Gustavo, Ana I Vazquez, Rohan Fernando, Yann C Klimentidis & Daniel Sorensen. 2013b. Prediction of complex human traits using the Genomic Best Linear Unbiased Predictor. *PLoS Genetics* 9(7). 1–15.

de los Campos, Gustavo, Yogasudha Veturi, Ana I. Vazquez, Christina Lehermeier & Paulino Pérez-Rodríguez. 2015. Incorporating genetic heterogeneity in whole-genome regressions using interactions. *Journal of Agricultural, Biological, and Environmental Statistics* 20(4). 467–490.

Clark, Samuel A., John M. Hickey, Hans D. Daetwyler & Julius H.J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics Selection Evolution* 44(4). 1–9.

Combs, Emily & Rex Bernardo. 2013. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *The Plant Genome* 6(1). 1–7.

Cover, T & P Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1). 21–27.

Crossa, José, Gustavo de los Campos, Paulino Pérez, Daniel Gianola, Juan Burgueño, José Luis Araus, Dan Makumbi, Ravi P. Singh, Susanne Dreisigacker, Jianbing Yan, Vivi Arief, Marianne Banziger & Hans Joachim Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186(2). 713–724.

Daetwyler, Hans D., Beatriz Villanueva & John A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* 3(10). 1–8.

Dagnachew, B.S., T.H.E. Meuwissen & T. Ådnøy. 2013. Genetic components of milk Fourier-transform infrared spectra used to predict breeding values for milk composition and quality traits in dairy goats. *Journal of Dairy Science* 96(9). 5933–5942.

Dawson, Julie C., Jeffrey B. Endelman, Nicolas Heslot, Jose Crossa, Jesse Poland, Susanne Dreisigacker, Yann Manès, Mark E. Sorrells & Jean Luc Jannink. 2013. The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research* 154. 12–22.

Dekkers, J C M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. *Journal of Animal Breeding and Genetics* 124. 331–341.

Efron, B, T Hastie, I Johnstone & R Tibshirani. 2004. Least angle regression. *The Annals of Statistics* 32(2). 407–499.

Endelman, Jeffrey B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal* 4(3). 250–255.

Falconer, Douglas S & Trudy F C Mackay. 1996. *Introduction to quantitative genetics*. Essex, UK: Prentice Hall 4th edn.

Ferragina, A, G de los Campos, A I Vazquez, A Cecchinato & G Bittante. 2015. Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. *Journal of Dairy Science* 98(11). 8133–8151.

Ferrio, J P, D Villegas, J Zarco, N Aparicio, J L Araus & C Royo. 2005. Assessment of durum wheat yield using visible and near-infrared reflectance spectra of canopies. *Field Crops Research* 94. 126–148.

Friedman, Jerome, Trevor Hastie, Holger Höfling & Robert Tibshirani. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1(2). 302–332.

Friedman, Jerome, Trevor Hastie & Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1). 1–22.

Fu, Wenjiang J. 1998. Penalized regressions: the Bridge versus the LASSO. *Journal of Computational and Graphical Statistics* 7(3). 397–416.

Garnsworthy, P C, J Wiseman & K Fegeros. 2000. Prediction of chemical, nutritive and agronomic characteristics of wheat by near infrared spectroscopy. *Journal of Agricultural Science* 135. 409–417.

Garrick, Dorian J. 2011. The nature, scope and impact of genomic prediction in beef cattle in the United States. *Genetics Selection Evolution* 43(1). 1–11.

Garriga, Miguel, Sebastián Romero-Bravo, Félix Estrada, Alejandro Escobar, Ivan A Matus, Alejandro del Pozo, Cesar A Astudillo & Gustavo A Lobos. 2017. Assessing wheat traits by spectral reflectance: do we really need to focus on predicted trait-values or directly identify the elite genotypes group? *Frontiers in Plant Science* 8(280). 1–12.

Gianola, Daniel, Rohan L. Fernando & Alessandra Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173. 1761–1776.

Gitelson, Anatoly A, Yoram J Kaufman & Mark N Merzlyak. 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment* 58(3). 289–298.

Goddard, Mike. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* 136. 245–257.

González-Camacho, J. M., G. de los Campos, P. Pérez, D. Gianola, J. E. Cairns, G. Mahuku, R. Babu & J. Crossa. 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics* 125(4). 759–771.

Habier, D, R L Fernando & J C M Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177. 2389–2397.

Habier, D, Rohan L Fernando & Dorian J Garrick. 2013. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. *Genetics* 194. 597–607.

Habier, David, Jens Tetens, Franz Reinhold Seefried, Peter Lichtner & Georg Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution* 42(5). 1–12.

Haboudane, Driss, John R Miller, Nicolas Tremblay, Pablo J Zarco-Tejada & Louise Dextraze. 2002. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sensing of Environment* 81. 416–426.

Hadley, Wickman. 2016. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Hansen, P M & J K Schjoerring. 2003. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote Sensing of Environment* 86. 542–553.

Hastie, Trevor, Robert Tibshirani & J. H. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. New York, USA: Springer 2nd edn.

Hayes, Ben J, Phillip J Bowman, Amanda C Chamberlain, Klara Verbyla & Mike E Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution* 41. 51.

Hazel, L N. 1943. The genetic basis for constructing selection indexes. *Genetics* 28(6). 476–490.

Heffner, Elliot L, Mark E Sorrells & Jean-Luc Jannink. 2009. Genomic Selection for Crop Improvement. *Crop Science* 49. 1–12.

Henderson, C R. 1950. Estimation of genetic parameters. *The Annals of Mathematical Statistics* 21. 309–310.

Henderson, C R. 1963. Selection index and expected genetic advance. In *Statistical genetics and plant breeding: A symposium and workshop*, 141–163. Washington, D.C.: National Academy of Sciences-National Research Council.

Henderson, C R & R L Quaas. 1976. Multiple trait evaluation using relatives' records. *Journal of Animal Science* 43(6). 1188–1197.

Hernandez, Javier, Gustavo A Lobos, Iván Matus, Alejandro del Pozo, Paola Silva & Mauricio Galleguillos. 2015. Using ridge regression models to estimate grain yield from field spectral data in bread wheat (Triticum aestivum L.) grown under three water regimes. *Remote Sensing* 7. 2109–2126.

Isidro, Julio, Jannink Jean-Luc, Deniz Akdemir, Jesse Poland, Nicolas Heslot & Mark E Sorrells. 2015. Training set optimization under population structure in genomic selection. *Theor Appl Genet* 128. 145–158.

Jacobson, Amy, Lian Lian, Shengqiang Zhong & Rex Bernardo. 2014. General combining ability model for genomewide selection in a biparental cross. *Crop Science* 54(3). 895–905.

Jiang, Guo-Liang. 2013. Molecular Markers and Marker-Assisted Breeding in Plants. In *Plant breeding from laboratories to fields*, InTech.

Lehermeier, Christina, Nicole Krämer, Eva Bauer, Cyril Bauland, Christian Camisan, Laura Campo, Pascal Flament, Albrecht E Melchinger, Monica Menz, Nina Meyer, Laurence Moreau, Jesús Moreno-González, Milena Ouzunova, Hubert Pausch, Nicolas Ranc, Wolfgang Schipprack, Manfred Schönleben, Hildrun Walter, Alain Charcosset & Chris-Carolin Schön. 2014. Usefulness of Multiparental Populations of Maize. *Genetics* 198. 3–16.

Lehermeier, Christina, Chris Carolin Schön & Gustavo de los Campos. 2015. Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics* 201(1). 323–337.

Lopez-Cruz, Marco, Eric Olson, Gabriel Rovere, Jose Crossa, Susanne Dreisigacker, Mondal Suchismita, Ravi Singh & Gustavo de los Campos. 2020. Regularized selection indices for breeding value prediction using hyper-spectral image data. *Scientific Reports* 10(8195).

Lorenz, Aaron & Kevin P Smith. 2015. Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. *Crop Science* 55. 2657–2667.

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

Lush, Jay L. 1935. Progeny test and individual performance as indicators of an animal's breeding value. *Journal of Dairy Science* 18(1). 1–19.

Lush, Jay L. 1937. *Animal breeding plans*. Ames, Iowa: Iowa State College, Ames.

Lush, Jay L. 1948. *The genetics of populations*. Iowa: Iowa State College, Ames special re edn.

Mahlein, A K, E C Oerke, U Steiner & H Wilhelm Dehne. 2012. Recent advances in sensing plant diseases for precision crop protection. *European Journal of Plant Pathology* 133. 197–209.

Melchinger, Albrecht E., H. Friedrich Utz & Chris C. Schör. 1998. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* 149(1). 383–403.

Meuwissen, T H E, B J Hayes & M E Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4). 1819–1829.

Miah, G., M. Y. Rafii, M. R. Ismail, A. B. Puteh, H. A. Rahim, R. Asfaliza & M. A. Latif. 2013. Blast resistance in rice: a review of conventional breeding to molecular approaches. *Molecular Biology Reports* 40(3). 2369–2388.

Montes, J M, H F Utz, W Schipprack, B Kusterer, J Muminovic, C Paul & A E Melchinger. 2006. Near-infrared spectroscopy on combine harvesters to measure maize grain dry matter content and quality parameters. *Plant Breeding* 125. 591–595.

Montes, Juan M, Albrecht E Melchinger & Jochen C Reif. 2007. Novel throughput phenotyping platforms in plant genetic studies. *Trends in Plant Science* 12(10). 10–13.

Montesinos-López, Osval, Abelardo Montesinos-López, Jose Crossa, Gustavo de los Campos, Gregorio Alvarado, Mondal Suchismita, Jessica Rutkoski, Lorena González-Pérez & Juan Burgueño. 2017. Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. *Plant Methods* 13(4). 1–23.

Morota, Gota & Daniel Gianola. 2014. Kernel-based whole-genome prediction of complex traits: A review. *Frontiers in Genetics* 5. 1–13.

Nagel, Kerstin A., Alexander Putz, Frank Gilmer, Kathrin Heinz, Andreas Fischbach, Johannes Pfeifer, Marc Faget, Stephan Blossfeld, Michaela Ernst, Chryssa Dimaki, Bernd Kastenholz, Ann Katrin Kleinert, Anna Galinski, Hanno Scharr, Fabio Fiorani & Ulrich Schurr. 2012. GROWSCREEN-Rhizo is a novel phenotyping robot enabling simultaneous measurements of root and shoot growth for plants grown in soil-filled rhizotrons. *Functional Plant Biology* 39(11). 891–904.

Oblath, Emily A, Terry A Isbell, Mark A Berhow, Brett Allen, David Archer, Jack Brown, Russell W Gesch, Jerry L Hatfield, Jalal D Jabro, James R Kiniry & Daniel S Long. 2016. Development of near-infrared spectroscopy calibrations to measure quality characteristics in intact Brassicaceae germplasm. *Industrial Crops & Products* 89. 52–58.

Olson, K. M., P. M. VanRaden & M. E. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *Journal of Dairy Science* 95(9). 5378–5383.

Perez, Paulino & Gustavo de los Campos. 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198(2). 483–495.

Pérez-Rodríguez, Paulino, José Crossa, Jessica Rutkoski, Jesse Poland, Ravi Singh, Andrés Legarra, Enrique Autrique, Gustavo De Los Campos, Juan Burgueño & Susanne Dreisigacker. 2017. Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. *The Plant Genome Journal* 10(2). 1–15.

Poland, Jesse, Jeffrey Endelman, Julie Dawson, Jessica Rutkoski, Shuangye Wu, Yann Manes, Susanne Dreisigacker, José Crossa, Héctor Sánchez-villeda, Mark Sorrells & Jean-Luc Jannink. 2012. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome Journal* 5(3). 103–113.

Pritchard, Jonathan K & Peter Donnelly. 2001. Case–Control Studies of Association in Structured or Admixed Populations. *Theoretical Population Biology* 60. 227–237.

Pritchard, Jonathan K., Matthew Stephens & Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155. 945–959.

Pszczola, M & M P L Calus. 2016. Updating the reference population to achieve constant genomic prediction reliability across generations. *Animal* 10(6). 1018–1024.

R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Riedelsheimer, Christian, Jeffrey B. Endelman, Michael Stange, Mark E. Sorrells, Jean Luc Jannink & Albrecht E. Melchinger. 2013. Genomic predictability of interconnected biparental maize populations. *Genetics* 194. 493–503.

Rincent, R, S Nicolas, T Altmann, D Brunel, P Revilla, A Melchinger, E Bauer, N Meyer, C Giauffret, C Bauland, P Jamin, J Laborde, H Monod, P Flament, A Charcosset & L Moreau. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). *Genetics* 192. 715–728.

Rio, Simon, Laurence Moreau, Alain Charcosset & Tristan Mary-Huard. 2020. Accounting for group-specific allele effects and admixture in genomic predictions: Theory and experimental evaluation in maize. *Genetics* 216(1). 27–41.

Roemer, C, M Wahabzada, A Ballvora, F Pinto, M Rossini, C Panigada, J Behmann, J Leon, C Thurau, C Bauckhage, K Kersting, U Rascher & L Pluemer. 2012. Early Drought Stress Detection in Cereals: Simplex Volume Maximization for Hyperspectral Image Analysis. *Functional Plant Biology* 39. 878–890.

Roth, Morgane, Hélène Muranty, Mario Di Guardo, Walter Guerra, Andrea Patocchi & Fabrizio Costa. 2020. Genomic prediction of fruit texture and training population optimization towards the application of genomic selection in apple. *Horticulture Research* 7(1).

Rutkoski, Jessica, Jesse Poland, Suchismita Mondal, Enrique Autrique, Lorena Pérez-González, José Crossa, Matthew Reynolds & Ravi Singh. 2016. Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in heat. *G3: Genes, Genomes, Genetics* 6. 2799–2808.

Schulz-Streeck, T, J O Ogutu, Z Karaman, C Knaak & H P Piepho. 2012. Genomic Selection using Multiple Populations. *Crop Science* 52. 2453–2461.

Searle, S R, G Casella & C E McCulloch. 1992. *Variance components*. John Wiley & Sons, Inc.

Smith, H F. 1936. A discrimant function for plant selection. *Annals of Eugenics* 7. 240–250.

Soyeurt, H, I Misztal & N Gengler. 2010. Genetic variability of milk components based on mid-infrared spectral data. *Journal of Dairy Science* 93(4). 1722–8.

Spielbauer, Gertraud, Paul Armstrong, John W Baier, William B. Allen, Katina Richardson, Bo Shen & A Mark Settles. 2009. High-throughput near-infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. *Cereal Chemistry* 86(5). 556–564.

Tang, Haibao, Uzay Sezen & Andrew H. Paterson. 2010. Domestication and plant genomes. *Current Opinion in Plant Biology* 13. 160–166.

Tattaris, Maria, Matthew P. Reynolds & Scott C. Chapman. 2016. A direct comparison of remote sensing approaches for high-throughput phenotyping in plant breeding. *Frontiers in Plant Science* 7. 1–9.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B* 58(1). 267–288.

Tucker, Compton J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment* 8(2). 127–150.

VanRaden, P M. 2007. Genomic measures of relationship and inbreeding. *Interbull Bulletin* 37. 33–36.

VanRaden, P M. 2008. Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91(11). 4414–4423.

Veturi, Yogasudha, Gustavo de los Campos, Nengjun Yi, Wen Huang, Ana I Vazquez & Brigitte Kühnel. 2019. Modeling heterogeneity in the genetic architecture of ethnically diverse groups using random effect interaction models. *Genetics* 211(April). 1395–1407.

Weber, V S, J L Araus, J E Cairns, C Sanchez, A E Melchinger & E Orsini. 2012. Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes. *Field Crops Research* 128. 82–90.

White, Jeffrey W, Pedro Andrade-Sanchez, Michael A Gore, Kevin F Bronson, Terry A Coffelt, Matthew M Conley, Kenneth A Feldmann, Andrew N French, John T Heun, Douglas J Hunsaker, Matthew A Jenks, Bruce A Kimball, Robert L Roth, Robert J Strand, Kelly R Thorp, Gerard W Wall & Guangyao Wang. 2012. Field Crops Research Field-based phenomics for plant genetics research. *Field Crops Research* 133. 101–112.

Wientjes, Yvonne C J, Roel F Veerkamp & Mario P L Calus. 2013. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. *Genetics* 193. 621–631.

William, H. M., R. Trethowan & E. M. Crosby-Galvan. 2007. Wheat breeding assisted by markers: CIMMYT's experience. *Euphytica* 157(3). 307–319.

Wolc, Anna, Jesus Arango, Petek Settar, Janet E. Fulton, Neil P. O'Sullivan, Rudolf Preisinger, David Habier, Rohan Fernando, Dorian J. Garrick & Jack C.M. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genetics Selection Evolution* 43(23). 1–8.

Wolc, Anna, A. Kranis, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, A. Avendano, K. A. Watson, J. M. Hickey, G. de los Campos, R. L. Fernando, D. J. Garrick & J. C.M. Dekkers. 2016. Implementation of genomic selection in the poultry industry. *Animal Frontiers* 6(1). 23–31.

Xu, Yunbi & Jonathan H. Crouch. 2008. Marker-assisted selection in plant breeding: From publications to practice. *Crop Science* 48(2). 391–407.

Xu, Yunbi, Xiaogang Liu, Junjie Fu, Hongwu Wang, Jiankang Wang, Changling Huang, Boddupalli M. Prasanna, Michael S. Olsen, Guoying Wang & Aimin Zhang. 2020. Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Communications* 1(1). 100005.

Xu, Yunbi, Debra J. Skinner, Huixia Wu, Natalia Palacios-Rojas, Jose Luis Araus, Jianbing Yan, Shibin Gao, Marilyn L. Warburton & Jonathan H. Crouch. 2009. Advances in maize genomics and their value for enhancing genetic gains from breeding. *International Journal of Plant Genomics* 1(957602).

Yu, Jianming, James B Holland, Michael D Mcmullen & Edward S Buckler. 2008. Genetic Design and Statistical Power of Nested Association Mapping in Maize. *Genetics Selection Evolution* 178. 539–551.

Zhu, Chengsong, Michael Gore, Edward S Buckler & Jianming Yu. 2008. Status and prospects of association mapping in plants. *The Plant Genome Journal* 1(1). 5–20.

Zou, Hui & Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* 67(2). 301–320.