OXFORD GENETICS

GENETICS, 2021, 218(1), iyab030
DOI: 10.1093/genetics/iyab030
Advance Access Publication Date: 20 March 2021
Genomic Prediction

Optimal breeding-value prediction using a sparse selection index

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

[1]Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA
[2]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI 48824, USA
[3]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA
[4]Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

*Corresponding author: Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI 48824, USA. gustavoc@msu.edu

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 linkage disequilibrium patterns can lead to heterogeneity in SNP effects. In this context, calibrating genomic predictions using a large, potentially heterogeneous, training data set may not lead to optimal prediction accuracy. Some studies tried to address this sample size/homogeneity trade-off using training set optimization algorithms; 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 is a sparse selection index (SSI) that integrates selection index methodology with sparsity-inducing techniques commonly used for high-dimensional regression. The sparsity of the resulting index is controlled by a regularization parameter (λ); the G-Best Linear Unbiased Predictor (G-BLUP) (the prediction method most commonly used in plant and animal breeding) appears as a special case which happens when $\lambda = 0$. In this study, we present the methodology and demonstrate (using two wheat data sets with phenotypes collected in 10 different environments) that the SSI can achieve significant (anywhere between 5 and 10%) gains in prediction accuracy relative to the G-BLUP.

Keywords: genomic prediction; selection index; prediction accuracy; penalized regression; GenPred; shared data resources

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 candidates of selection or on progeny testing (*e.g.*, Lush 1935). These methods were later extended into selection indices (Smith 1936; Hazel 1943) 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 and often define "families." In this context, borrowing of information spans 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 models (aka, genomic selection, GS; Meuwissen *et al.* 2001) have 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 nonadditive effects coupled with population structure and admixture, can make marker effects vary across subgroups in the sample (Pritchard and Donnelly 2001; de los Campos *et al.* 2015). All these factors can make the genomic relationships derived from markers inaccurate predictors of the genomic relationships realized at causal loci (de los Campos *et al.* 2013b). Therefore, the accuracy of G-BLUP may be suboptimal 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.

Received: September 30, 2020. **Accepted:** February 13, 2021

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; Lehermeier *et al.* 2015) or interaction models (*e.g.*, de los Campos *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 and Calus 2016; Akdemir and 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 identifies, 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).

Materials and Methods

A standard selection index ($\mathfrak{T}_i$) predicts the breeding value of an individual (u_i) using a linear combination of the training phenotypes ($\mathbf{y} = (y_1, \ldots, y_n)'$): $\mathfrak{T}_i = \boldsymbol{\beta}_i'\mathbf{y} = \sum_{j=1}^{n} \beta_{ij} y_j$. Here, phenotypes are assumed to be centered and corrected by nongenetic effects (*e.g.*, experiment and block effects), and $\boldsymbol{\beta}_i = \{\beta_{ij}\}$ is a vector of weights that are obtained as the solution to the following optimization problem:

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

The right-hand side of the above problem expands to $\mathbb{E}(u_i^2) + \boldsymbol{\beta}_i'\mathbb{E}(\mathbf{y}\mathbf{y}')\boldsymbol{\beta}_i - 2\mathbb{E}(\mathbf{y} \times u_i)'\boldsymbol{\beta}_i$. Assuming that genetic (u_i) and nongenetic 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 $\mathbb{E}(u_i - \boldsymbol{\beta}_i'\mathbf{y})^2 = \sigma_u^2 + \boldsymbol{\beta}_i'\mathbf{P}\boldsymbol{\beta}_i - 2\sigma_u^2\mathbf{G}_i'\boldsymbol{\beta}_i$, 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. Because σ_u^2 does not depend on $\boldsymbol{\beta}_i$, the aforementioned optimization problem can be reduced to

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

where $\lambda_0 = \frac{\sigma_\varepsilon^2}{\sigma_u^2} = \frac{1-h^2}{h^2}$ 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

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{G} + \lambda_0\mathbf{I})^{-1}\mathbf{G}_i \quad (2)$$

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 Supplementary File S1 in the Supplementary Material for a proof). Therefore, 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 ($\hat{\mathfrak{T}}_i = \hat{\boldsymbol{\beta}}_i'\mathbf{y}$) 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.

Sparse selection index (SSI) 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 (1); therefore,

$$\tilde{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}_i}\left\{\frac{1}{2}\boldsymbol{\beta}_i'(\mathbf{G} + \lambda_0\mathbf{I})\boldsymbol{\beta}_i - \mathbf{G}_i'\boldsymbol{\beta}_i + \lambda\sum_{j=1}^{n}|\beta_{ij}|\right\}. \quad (3)$$

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). Specifically, in Equation (3), the relationships between the prediction point and the training genotypes ($\mathbf{G}_i$) enters in the optimization problem in the same way the right-hand-side term $\mathbf{X}'\mathbf{y}$ enters in least-square and LASSO problems for linear models of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. On the other hand, the term $(\mathbf{G} + \lambda_0\mathbf{I})$, which accounts for relationships among training genotypes, enters in Equation (3) in the same way that $\mathbf{X}'\mathbf{X}$ enters in least-square and LASSO problems.

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.

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-small) is a small, structured data (see Supplementary Figure S1).

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 5 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, and 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 S2B for boxplots of grain yield) and the heritability of single-plot records varied between 0.23 and 0.57 (see Supplementary Table S1 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 remaining genotypes that were missing were imputed with the sample mean of genotypes at the corresponding loci. The data set has been previously described and analyzed by Pérez-Rodríguez *et al.* (2017).

The Wheat-small 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 (ME1, ..., ME4). 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 S2A for boxplots of grain yield data) with heritability estimates for the least-square means ranging between 0.43 and 0.50 (see Supplementary Table S2). 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 and de los Campos 2014) and has been described and analyzed by previous authors (*e.g.*, de los Campos *et al.* 2009; Crossa *et al.* 2010).

Analyses

For each data set, we computed a *genomic relationship* matrix $\boldsymbol{G}$ using (centered and standardized) marker information, $\boldsymbol{X} = \{x_{im}\}$, as $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Z}'/p$, where p is the number of markers and $\boldsymbol{Z} = \{(x_{im} - \bar{x}_m)/sd_{x_m}\}$ is the matrix of centered and standardized markers obtained by subtracting from each marker entry, x_{im}, the mean of each column, $\bar{x}_m$, 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 (2),

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

(for the standard SI) and Equation (3),

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

(for the SSI), with $\boldsymbol{G} = \boldsymbol{G}_{trn}$ representing the genomic matrix of the training data points (*i.e.*, with dimensions $n_{trn} \times n_{trn}$, where $n_{trn} = 0.7n$), and $\boldsymbol{G}_i = \boldsymbol{G}_{trn,\ tst(i)}$ 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 $\boldsymbol{G}_i$ 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 $\hat{\mathfrak{T}}_i = \hat{\boldsymbol{\beta}}_i'\boldsymbol{y}_{trn}$ (for the standard SI) and $\hat{\mathfrak{T}}_i(\lambda) = \tilde{\boldsymbol{\beta}}_i'(\lambda)\boldsymbol{y}_{trn}$ (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 $\varepsilon_i \sim^{iid} N(0, \sigma^2_\varepsilon)$ and $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2_u\boldsymbol{G})$. 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 $h^2 = \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_\varepsilon}$ for grain yield.

Prediction accuracy (ρ) was measured as the correlation between the phenotype and the index, divided by the square-root of the heritability of the trait (grain yield), $\rho = Acc(\hat{\mathfrak{T}}) = \frac{Cor(\hat{\mathfrak{T}}_i, y_i)}{h}$ (Dekkers 2007). We applied all methods to the same training-testing partitions (trn–tst partitions) and report, from these analyses, the average prediction accuracy and the proportion of times that one method was better than the other.

For the SSI, we estimated the accuracy over a grid of values of the regularization parameter ($\lambda = 0 < \lambda^{(1)} < \lambda^{(2)} < \ldots < \lambda_{max}$) where $\lambda_{max} = \max_i\left\{\frac{G_i}{diag(\boldsymbol{G}) + \lambda_0}\right\}$. Here, λ_{max} is the minimum value of λ that yields an SSI with no active predictors (*i.e.*, with all coefficients β_{ij} equal to zero), 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 the 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 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 $\mathfrak{T}_i(\hat{\lambda}_{cv})$ and evaluated the accuracy of the resulting index in the left-out data from the testing set.

Software

All the analyses were performed in the R environment-language (R Core Team 2019) version 3.5. Genomic relationships were derived using the getG() function of the BGData R-package (Grueneberg and de los Campos 2019). The heritability of the trait 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-small data set are presented in the Supplementary Material, Supplementary File S2. Training set optimization via subset selection (presented in the Section *Discussion*) was implemented in the STPGA R-package (Akdemir *et al.* 2015).

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 March 6th, 2021). The Wheat-small data set can be downloaded from the BGLR R-package by calling "data(wheat)." Supplementary File S1 contains a

proof on the equivalence between the standard SI and the BLUP. Code showing how to perform all analyses is provided in Supplementary File S2. All supplementary figures and tables are contained in Supplementary File S3. All supplementary files are available at figshare: https://doi.org/10.25386/genetics.14098952.

Results

Sparsity improves prediction accuracy

We assessed the effect of sparsity on the accuracy, by fitting the SSI for 100 values of λ ($0 < \lambda^{(1)} < \lambda^{(2)} < \ldots < \lambda_{max}$; 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 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 1 is displayed as the average (across genotypes in the testing set) 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 1). The maximum accuracy in the environment EHT (see Figure 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 1B), the same "belly" pattern can be observed in all environments, except for ME2. This case shows

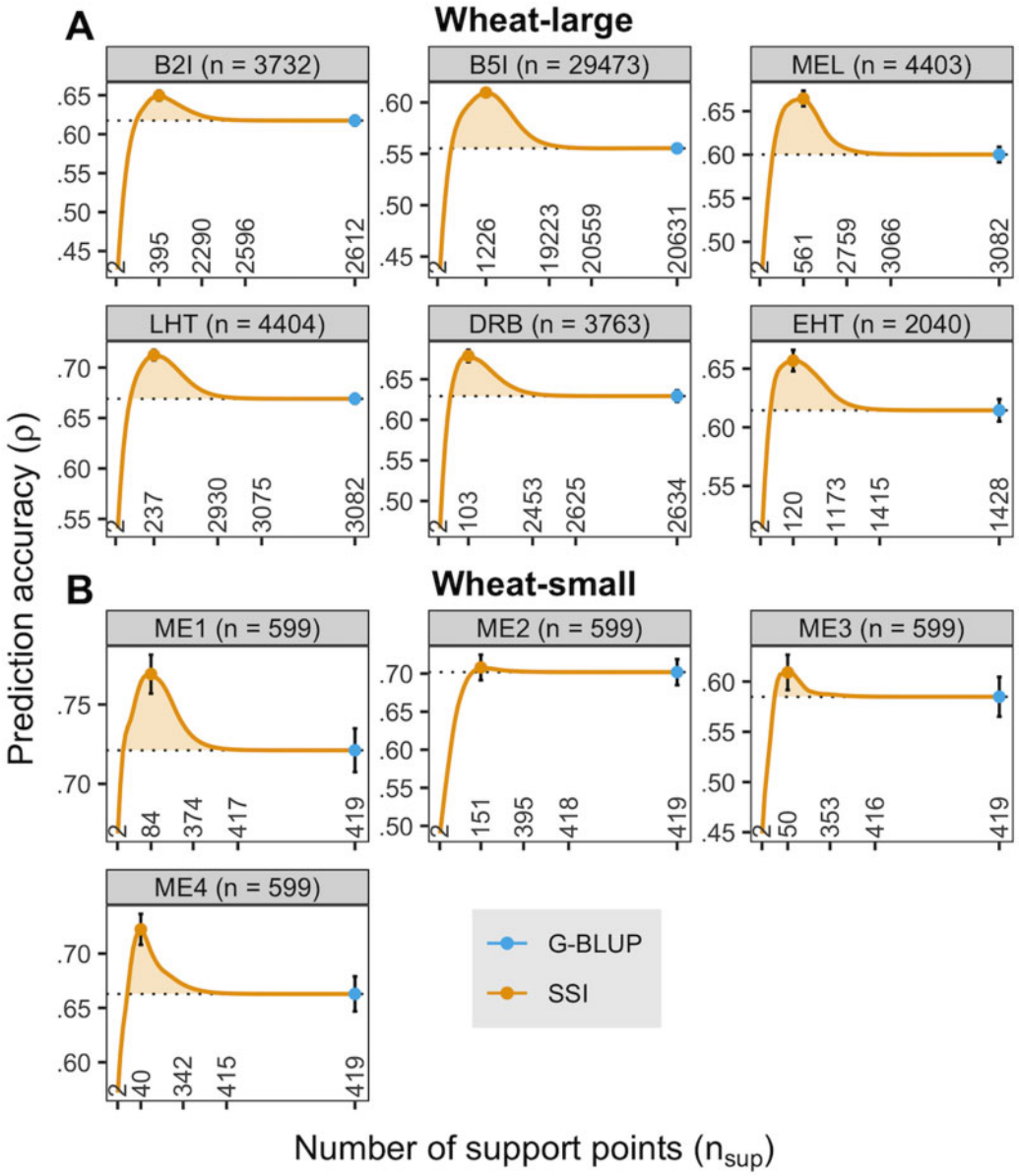

Figure 1 Prediction accuracy for grain yield (average across 100 trn–tst partitions) of the SSI versus the (average) number of support points of the SSIs. The G-BLUP (blue rightmost point) is a special case of an SSI when $\lambda = 0$. Each panel represents one environment within data set. (A) Wheat-large data set. B2I, bed planting + 2 irrigations; B5I, bed planting + 5 irrigations; MEL, flat planting + 5 irrigations; LHT, late planting date; EHT, early planting date; DRB, bed planting + drip irrigation. (B) Wheat-small data set. ME, mega-environment. Vertical bars represent a 95% confidence interval for the average.

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).

Using an internal CV to achieve optimal sparsity

The results in Figure 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 the SSI can achieve, 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 fivefold 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 $\mathfrak{T}_i(\hat{\lambda}_{cv})$ in a testing set that was not used to choose $\hat{\lambda}_{cv}$. This was repeated for 100 trn–tst partitions. Figure 2 shows the accuracy of $\mathfrak{T}_i(\hat{\lambda}_{cv})$ versus that of the G-BLUP, each point in the plot represents a trn–tst partition. In the Wheat-large data set, the optimal SSI outperformed the G-BLUP in 94% of the cases (Table 1). For this data set, the SSI offered accuracy gains ranging from 5% (in the environment B2I) to 10% (in the environments B5I and MEL) in the correlation metric. Similar patterns were observed with the Wheat-small data set. In environments ME1 and ME4 the SSI outperformed the G-BLUP in more than 80% of the trn–tst partitions (Table 1), with gains in accuracy ranging from 5% to 8%. However, in ME2 and ME3, there were no significant gains in accuracy (see Table 1).

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 subjects with a nonzero entry in $\tilde{\boldsymbol{\beta}}_i(\lambda)$. Figure 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 ≈5,000. In 3 of the environments (B2l, MEL, and LHT), the average number of support points was $n_{sup} \approx 450$, that is ∼15-20% of the size of the training set. In environment B5l, the proportion of active training support points was ∼5–10%. On the other hand, in environment EHT predictions relied on an average of $n_{sup} \approx 178$ (out of 1,428) individuals from training (Figure 3). Similar patterns were also observed in the Wheat-small data (Supplementary Figure S3). For instance, testing phenotypes from environment ME1 were optimally predicted using, on average, $n_{sup} \approx 78$ (out of 419); however, the relative sparsity (n_{sup}/n_{trn}) was smaller in the Wheat-large data set (5-17%) than in the Wheat-small data set (12-60%).

Figure 4 shows (for selected testing genotypes) the coordinates on the 1st and 2nd PC of both the prediction point (yellow circle) and the training genotypes. Active training genotypes are represented in a green circle, and those nonactive (*i.e.*, with zero weight in the index) are represented in gray. In some cases, the support set includes training genotypes that are nearby (according to the coordinates on the top 2 PCs) the prediction point. However, in other cases, the support set spanned outside clusters that could be defined by top PCs (a similar plot for the Wheat-small data set is presented in Supplementary Figure S4). We note that the plots in Figure 4 (and Supplementary Figure S4) use coordinates that are based on two PCs that together explain 10.3% (and 16.3%) of the variance in genotypes. Thus, it is still possible that some points that appear distant in the top 2 PCs coordinates may still have a sizable genomic relationship. This could happen if, for instance, two lines share one ancestor but have the other ancestors originating from divergent populations. Thus, in the next section, we study in more detail the link between genomic relationships and the weights on the SSI.

Genomic relationships and weights in standard and sparse selection indices

Figure 5A shows the coefficients of the G-BLUP and the SSI (*i.e.*, the β_{ij}'s derived from Equations (2) and (3), respectively) versus the genomic relationship (g_{ij}, the ij entry of $\boldsymbol{G}$). In Figure 5A, the β_{ij}'s were derived for one trn–tst partition with fixed heritability

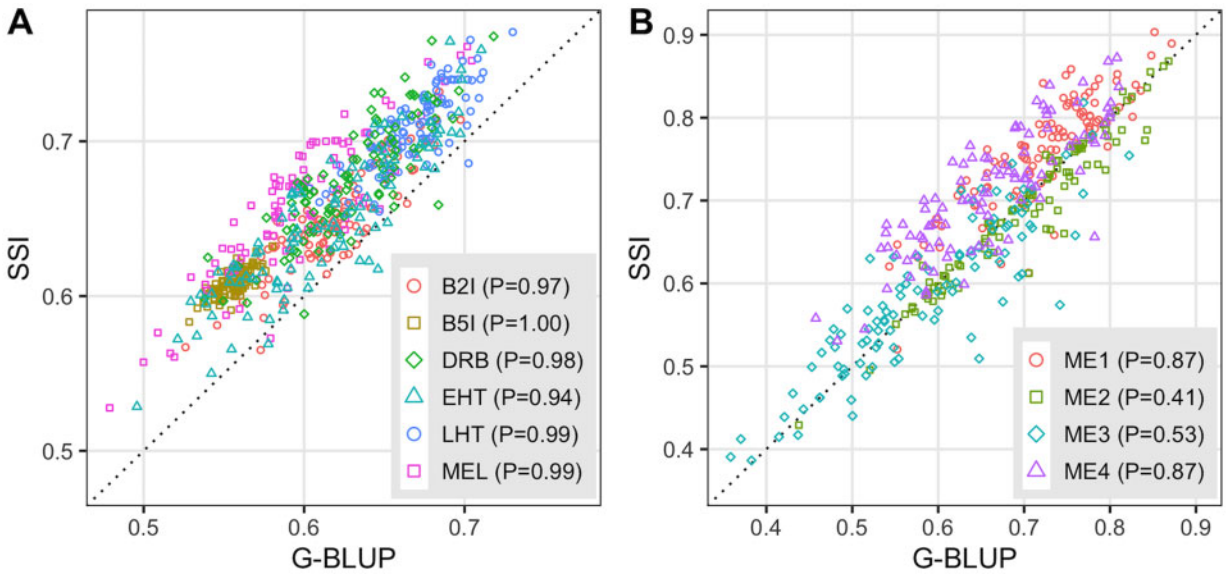

Figure 2 Prediction accuracy for grain yield of the optimal SSI versus that of the G-BLUP. Each point represents a trn–tst partitions (a total of 100 partitions were implemented), the point shape and color represent environments. (A) Wheat-large data set. B2I, bed planting + 2 irrigations; B5I, bed planting + 5 irrigations; MEL, flat planting + 5 irrigations; LHT, late planting date; EHT, early planting date; DRB, bed planting + drip irrigation. (B) Wheat-small data set. ME, mega-environment. The value of λ in the SSI was estimated using 10 fivefold CVs conducted within the training data. In parenthesis, by the legend, P is the proportion of times the SSI was better than the G-BLUP.

Table 1 Prediction accuracy for grain yield (average across 100 partitions) achieved by sparse selection indices (SSIs) and by the G-BLUP (standard SI), by data set and environmental condition

Environment	n_{tst}	n_{trn}	Method	λ_{cv}[a]	n_{sup}[b]	Accuracy (SD)	Pc
Wheat-large							
B2I	1,120	2,612	G-BLUP	0.0000	2,612	0.617 (0.031)	0.97
			SSI	0.0135	434	0.648 (0.031)	
B5I	8,842	20,631	G-BLUP	0.0000	20,631	0.555 (0.010)	1.00
			SSI	0.0107	1,470	0.609 (0.009)	
MEL	1,321	3,082	G-BLUP	0.0000	3,082	0.600 (0.045)	0.99
			SSI	0.0131	524	0.661 (0.046)	
LHT	1,322	3,082	G-BLUP	0.0000	3,082	0.669 (0.024)	0.99
			SSI	0.0168	380	0.709 (0.025)	
DRB	1,129	2,634	G-BLUP	0.0000	2,634	0.629 (0.035)	0.98
			SSI	0.0322	136	0.675 (0.037)	
EHT	612	1,428	G-BLUP	0.0000	1,428	0.614 (0.049)	0.94
			SSI	0.0301	178	0.649 (0.047)	
Wheat-small							
ME1	180	419	G-BLUP	0.0000	419	0.721 (0.070)	0.87
			SSI	0.0413	78	0.760 (0.067)	
ME2	180	419	G-BLUP	0.0000	419	0.702 (0.087)	0.41
			SSI	0.0123	254	0.692 (0.085)	
ME3	180	419	G-BLUP	0.0000	419	0.585 (0.101)	0.53
			SSI	0.0613	84	0.586 (0.093)	
ME4	180	419	G-BLUP	0.0000	419	0.663 (0.082)	0.87
			SSI	0.0617	54	0.714 (0.075)	

B2I, bed planting + 2 irrigations; B5I, bed planting + 5 irrigations; MEL, flat planting + 5 irrigations; LHT, late planting date; EHT, early planting date; DRB, bed planting + drip irrigation; ME, mega-environment; n_{tst} and n_{trn}, size of the testing and training data sets, respectively; SD, standard deviation across trn–tst partitions.

[a] Optimal value of λ (average across partitions) estimated by cross-validating the training set.

[b] Average number of support points in the SSIs. G-BLUP model corresponds to an SSI with $\lambda = 0$ and $n_{sup} = n_{trn}$.

[c] P: proportion of times (out of the 100 partitions) that the SSI outperformed the G-BLUP in prediction accuracy.

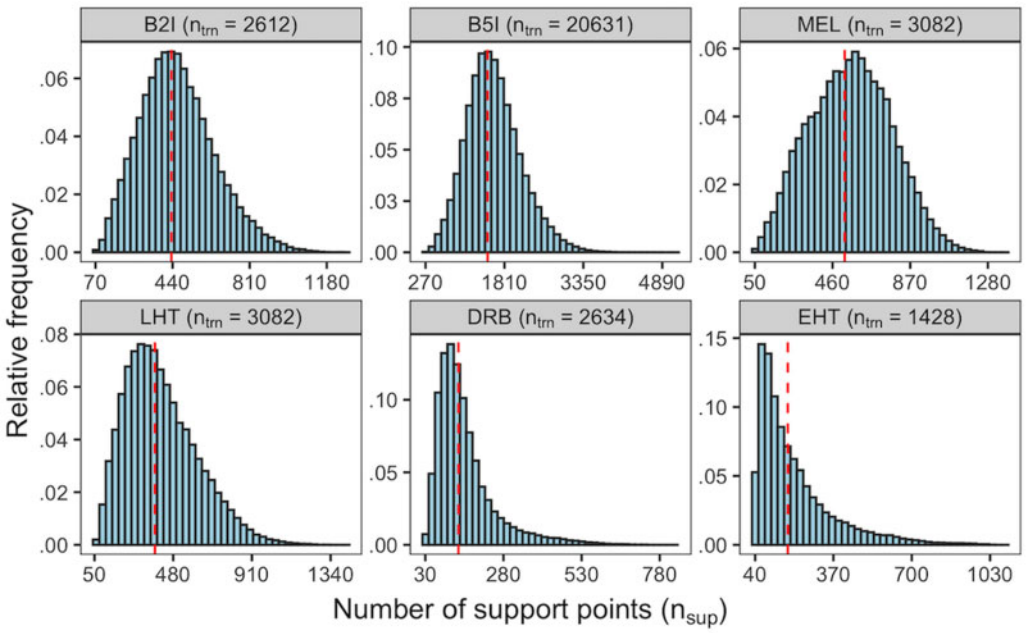

Figure 3 Distribution of the number of training support points (n_{sup}) in the optimal SSI for grain yield (results obtained over 100 trn–tst partitions; n_{trn}, size of the training data set), by environmental condition. B2I, bed planting + 2 irrigations; B5I, bed planting + 5 irrigations; MEL, flat planting + 5 irrigations; LHT, late planting date; EHT, early planting date; DRB, bed planting + drip irrigation. Wheat-large data set.

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 genotypes. On the other hand, as expected, the SSI zero-outs most of the weights. 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 5A).

Figure 5B 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- small (Supplementary Figure S6). Interestingly, the proportion of coefficients zeroed-out also decreases for "negative" genomic relationships, suggesting that the SSI does not only use a "local" support set; instead, the SSI seems to use informative support points.

The linear kernel used here can have negative off-diagonal values (*i.e.*, $g_{ij} < 0$, mostly between pairs of genotypes from different clusters); these negative covariances are, in the context of the additive model fitted here, informative and thus some of the training points with negative covariance with the prediction genotypes can become active in the SSI. However, note that the size of the coefficients corresponding to points with negative genomic relationships is considerably smaller than the size of the coefficients for points with positive relationships with the prediction genotypes. The patterns observed in other environments of the Wheat-large data set and the four environments of the Wheat-small data set were conceptually similar to the ones presented in Figure 5 (see Supplementary Figures S5 and S6).

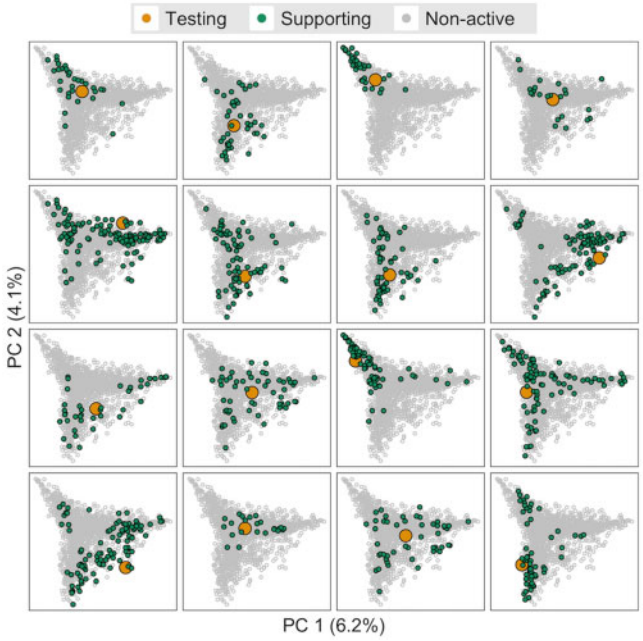

Figure 4 First two principal components coordinates for prediction points (yellow) and the corresponding support points (green). Gray points represent genotypes that did not contribute to the prediction of the genetic value of grain yield of the genotype in yellow. All panels represent solutions for the environment. EHT, early planting date, wheat-large data set.

Discussion

Sample size has been recognized as one of the main factors limiting prediction accuracy in genomic prediction (Lorenzana and Bernardo 2009; de los Campos *et al.* 2013a; Habier *et al.* 2013). In unstructured populations, SNP effects can be assumed to be homogeneous, in this context, genomic prediction accuracy is expected to increase 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 LD-patterns may make SNP effects heterogenous across subgroups in the sample. In that context, a larger training data set may not translate into a higher prediction accuracy. This phenomenon has been recognized in both plant and animal breeding, as well as in complex traits 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 previous 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

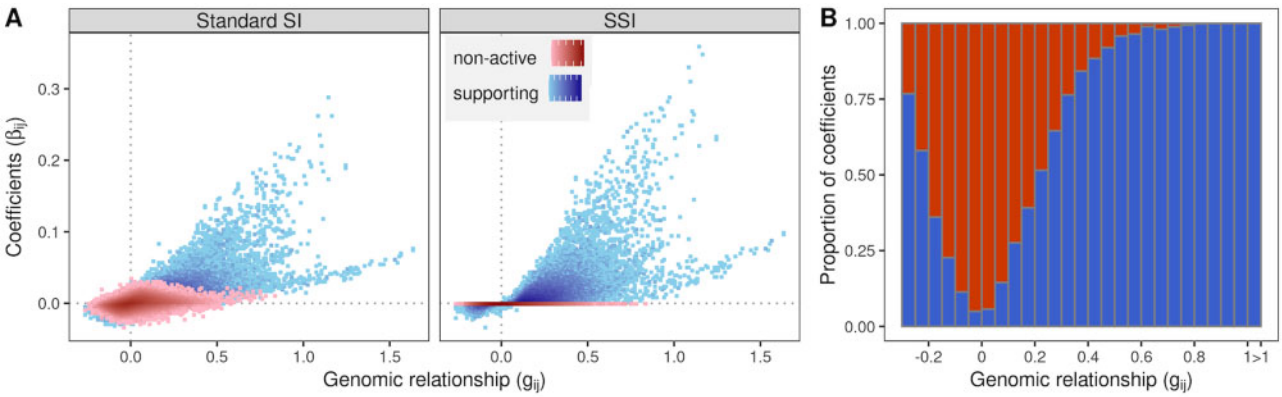

Figure 5 (A) Weights (β_{ij}) of a standard SI (G-BLUP) and the optimal SSI for grain yield versus the genomic relationship (g_{ij}). (B) Proportion of weights in the SSI that were zero (nonactive) and nonzero (support points); environment. EHT, early planting date, wheat-large data set.

individuals may not improve (or may even reduce) prediction accuracy relative to models trained with family data only (Makowsky *et al.* 2011).

Trade-offs between sample size and effect homogeneity

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 effects 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 models effect heterogeneity using group-specific effects (*e.g.*, Veturi *et al.* 2019; Rio *et al.* 2020). This approach is useful when individuals cluster in a few (*e.g.*, 2 or 3) well-defined clusters; however, the approach becomes less useful and more difficult to apply when data are 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).

Training-set optimization techniques: one-size may not fit all

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.*, Jacobson *et al.* 2014; Lorenz and Smith 2015; Wolc *et al.* 2016) or by using more sophisticated optimization algorithms (*e.g.*, Rincent *et al.* 2012; Akdemir *et al.* 2015). Given an available training set, optimization procedures aim at identifying a subset that is optimal for all the genotypes in the prediction set.

We applied the training set optimization methodology described in Rincent *et al.* (2012) to the same trn–tst partitions that we used to evaluate the predictive performance of the SSI (see Section *Materials and Methods* for details of the trn–tst partitions, and Figure 1 and Table 1 for results obtained using an SSI). For each data set and environment, we evaluated the prediction accuracy achieved by the G-BLUP using the entire data set, and using smaller sets chosen either at random or by optimizing the training set using the CDmean criteria (Rincent *et al.* 2012) as implemented in the STPGA R-package (Akdemir *et al.* 2015). Across data sets and environments, the optimized training sets did not produce higher prediction accuracy than the one achieved when using the entire training data (Supplementary Figure S7). Furthermore, in all cases, the SSI outperformed the G-BLUP calibrated with the entire data set and all the G-BLUP models calibrated using smaller (optimized) training sets (Supplementary Figure S7).

The above-results highlight the challenge of selecting a training set that is optimal for all the genotypes in a prediction set. Our approach tackles this problem by identifying an optimal training set for each testing genotype. Because different sparse indices are obtained for each individual in the prediction set, almost all individuals in the training set end up contributing to the index of one or more testing genotypes.

Sparsity of the SSI

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 CV 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) trn–tst partitions are enough to obtain an accurate estimate of the value of the regularization parameter that maximizes accuracy.

SSI and k-nearest neighbor

As noted, an SSI identifies, for each individual in the prediction set, a network of genotypes in the training data set (see Figure 4 and Supplementary Figure S4) that contribute to the prediction. At first glance, this appears similar to the approach used in a *k*-nearest neighbor (KNN) regression (Cover and 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 identify a "neighborhood." However, the SSI also considers the correlations (*i.e.*, redundancies between training genotypes which are described by off-diagonal matrices of the $\boldsymbol{G}$ matrix). As a consequence, the optimal support set of the SSI may zero-out coefficients for close relatives of prediction genotypes, and may include active coefficients for some distantly related individuals (see Figure 4 and Supplementary Figure S4). 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.

SSI and sparse genomic relationship matrices

At first glance, it may seem that an SSI could be obtained using G-BLUP equations by simply zeroing-out small off-diagonal coefficients of the $\boldsymbol{G}$ matrix. However, this approach would be different and will not necessarily yield a sparse index. First, as noted before, simply zeroing-out small coefficients of a $\boldsymbol{G}$ matrix does not consider the fact that training genotypes are also related. (The results in Figure 5 show that the SSI also zero-out weights for training genotypes with sizable genomic relationships with testing genotypes.) Second, zeroing-out coefficients of the $\boldsymbol{G}$ matrix does not guarantee that the inverse of $\boldsymbol{G}$ (and therefore, the G-BLUP equations) will be sparse (making sparse the inverse of $\boldsymbol{G}$, as in Graphical-LASSO, Friedman *et al.* 2008, may be more effective). Third, zeroing-out off-diagonals coefficients of $\boldsymbol{G}$ that are close or below zero ignores the fact that genotypes with negative genomic relationships with testing genotypes may be informative, simply because negative (co)variances are informative (this can also be seen in Figure 5, where some points with the negative genomic relationship are active in some SSIs).

Borrowing of information in the SSI

The fact that some points with negative genomic relationships contribute to the prediction equations of SSIs (see Figure 5) may be counter-intuitive, and may be considered undesirable. This happens simply because negative (co)variances are informative. However, the weights for training genotypes with negative genomic relationships with testing genotypes, if not zero, are small in absolute value. In other words, the SSI draws information primarily from closely related individuals. If one would like to obtain an SSI that is "strictly local" (*i.e.*, that only training genotypes are closely related to testing genotypes) one would need to use a

kernel that specify nonnegative prior (co)variances (*e.g.*, a Gaussian kernel or additive-by-additive structure **G#G**).

SSI and elastic-net penalty

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 (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 and Hastie 2005) regression. However, in the SSI the weight on the L2-penalty is only 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 CV, of the relative weights of the penalization parameters of the Elastic-Net (α and λ) optimization problem:

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

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 S3 for the Wheat-large data set) relative to the original SSI method (Equation 3). However, this practice did not provide any additional advantage over the original SSI in the Wheat-small data (see Supplementary Table S4).

The effect of sample size on the relative performance of the SSI

Sample size, SNP density, and genetic structure are features that may affect the ability of the SSI to achieve a higher prediction accuracy than the G-BLUP. To shed light on the effect of sample size we repeated the analysis presented before with varying sample size within each data set and environment (*i.e.*, holding the structure and the number of SNPs constant). The results (Supplementary Figure S8) showed that the difference in prediction performance of the SSI and that of the G-BLUP increased with sample size. Clearly, the use of an SSI is more appealing when either there is a strong structure and the sample size is very large. Such conditions offer opportunities for the SSI to identify optimal support points for each genotype in the prediction set.

Finally, we note that the derivation of the weights of the SSI depends on the trait only through the heritability (see Equation 3). Therefore, one could imagine deriving the weights of the SSI for each candidate of selection and then using these weights to predict breeding values for a range of traits with comparable heritability.

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 genotype in a 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 multi-generation data sets. The superiority of the SSI relative to a standard G-BLUP is clearer with large sample size.

Acknowledgments

We are grateful to CIMMYT's Global Wheat Program that provided both the experimental field and marker data used in this study.

Funding

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. G.D.L.C. received support from the National Institute for Food and Agriculture (NIFA) of the USDA (award #2021-67015-33413).

Conflicts of interest

None declared.

Literature cited

Akdemir D, Sanchez JI, Jannink J-L. 2015. Optimization of genomic selection training populations with a genetic algorithm. Genet Sel Evol. **47**:1–10.

Akdemir D, Isidro-Sanchez J. 2019. Design of training populations for selective phenotyping in genomic prediction. Sci Rep. **9**:1–15.

Cover T, Hart P. 1967. Nearest neighbor pattern classification. IEEE Trans Inform Theory. **13**:21–27.

Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, *et al.* 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics **186**: 713–724.

Daetwyler HD, Villanueva B, Woolliams JA. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One **3**:e3395.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. 2013a. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193**: 327–345.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, *et al.* 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics **182**:375–385.

de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. 2013b. Prediction of complex human traits using the Genomic Best Linear Unbiased Predictor. PLoS Genet. **9**:e1003608.

de los Campos G, Veturi Y, Vazquez AI, Lehermeier C, Pérez-Rodríguez P. 2015. Incorporating genetic heterogeneity in whole-genome regressions using interactions. J Agric Biol Environ Stat. **20**:467–490.

Dekkers JCM. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet. **124**:331–341.

Endelman JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome J. **4**:250–255.

Friedman J, Hastie T, Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**:432–441.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. **33**:1–22.

Grueneberg A, de los Campos G. 2019. BGData - A suite of R packages for genomic analysis with big data. G3 (Bethesda) **9**:1377–1383.

Habier D, Fernando RL, Dekkers JCM. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**:2389–2397.

Habier D, Fernando RL, Garrick DJ. 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics **194**: 597–607.

Hadley W. 2016. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag.

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol. **41**:51.

Hazel LN. 1943. The genetic basis for constructing selection indexes. Genetics **28**:476–490.

Henderson CR. 1950. Estimation of genetic parameters. Ann Math Stat. **21**:309–310.

Henderson CR. 1963. Selection index and expected genetic advance. In: Statistical Genetics and Plant Breeding: A Symposium and Workshop. Washington, D.C: National Academy of Sciences-National Research Council. p. 141–163.

Isidro J, Jean-Luc J, Akdemir D, Poland J, Heslot N, *et al.* 2015. Training set optimization under population structure in genomic selection. Theor Appl Genet. **128**:145–158.

Jacobson A, Lian L, Zhong S, Bernardo R. 2014. General combining ability model for genomewide selection in a biparental cross. Crop Sci. **54**:895–905.

Lehermeier C, Schön CC, de los Campos G. 2015. Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. Genetics **201**:323–337.

Lopez-Cruz M, Olson E, Rovere G, Crossa J, Dreisigacker S, *et al.* 2020. Regularized selection indices for breeding value prediction using hyper-spectral image data. Sci Rep. **10**:8195.

Lorenz AJ, Smith KP. 2015. Adding genetically distant individuals to training populations reduces genomic prediction accuracy in Barley. Crop Sci. **55**:2657–2667.

Lorenzana RE, Bernardo R. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet. **120**:151–161.

Lush JL. 1935. Progeny test and individual performance as indicators of an animal's breeding value. J Dairy Sci. **18**:1–19.

Lush JL. 1948. The Genetics of Populations. Ames, IA: Iowa State College.

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, *et al.* 2011. Beyond missing heritability: prediction of complex traits. PLoS Genet. **7**:e1002051.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**:1819–1829.

Olson KM, VanRaden PM, Tooker ME. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. **95**:5378–5383.

Pérez-Rodríguez P, Crossa J, Rutkoski J, Poland J, Singh R, *et al.* 2017. Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. Plant Genome **10**:1–15.

Perez P, de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics **198**:483–495.

Pritchard JK, Donnelly P. 2001. Case–control studies of association in structured or admixed populations. Theor Popul Biol. **60**:227–237.

Pszczola M, Calus MPL. 2016. Updating the reference population to achieve constant genomic prediction reliability across generations. Animal **10**:1018–1024.

R Core Team 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rincent R, Nicolas S, Altmann T, Brunel D, Revilla P, *et al.* 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics **192**:715–728.

Rio S, Moreau L, Charcosset A, Mary-Huard T. 2020. Accounting for group-specific allele effects and admixture in genomic predictions: theory and experimental evaluation in maize. Genetics **216**:27–41.

Schulz-Streeck T, Ogutu JO, Karaman Z, Knaak C, Piepho HP. 2012. Genomic selection using multiple populations. Crop Sci. **52**: 2453–2461.

Smith HF. 1936. A discrimant function for plant selection. Ann Eugen. **7**:240–250.

VanRaden PM. 2007. Genomic measures of relationship and inbreeding. Interbull Bull. **37**:33–36.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. **91**:4414–4423.

Veturi Y, de Los Campos G, Yi N, Huang W, Vazquez AI, *et al.* 2019. Modeling heterogeneity in the genetic architecture of ethnically diverse groups using random effect interaction models. Genetics **211**:1395–1407.

Wolc A, Kranis A, Arango J, Settar P, Fulton JE, *et al.* 2016. Implementation of genomic selection in the poultry industry. Anim Front. **6**:23–31.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. J R Stat Soc B. **67**:301–320.

Communicating editor: J. B. Endelman