Genetic image-processing using regularized selection indices

by

Marco Lopez-Cruz[1], Eric Olson[1], Gabriel Rovere[2,3,4], Jose Crossa[6], Susanne Dreisigacker[6], Sushismita 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

* Corresponding author. E-mail: gustavoc@msu.edu. (GDLC)

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 image-based regularized SIs offer consistently higher accuracy for grain yield than those achieved by canonical 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.

Author summary

A more intensive use of High-throughput phenotyping (HTP) in breeding programs can increase selection gains and can enable breeding for traits that are otherwise difficult to measure and to breed for (e.g., drought resistance in plants). Most of the phenotypes generated by HTP platforms are high-dimensional, making the use of these data for breeding decisions challenging. We

propose to address this problem by using regularized selection indices (SIs). The methodology combines ideas from quantitative genetics with methods used in high-dimensional regressions. Using wheat data from CIMMYT's wheat breeding program we show that regularized SIs deliver more accurate selection decisions than that of canonical SIs.

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 [1]) to extensive HTP using sensing devices mounted on aerial (e.g., hyper-spectral cameras mounted on aerial vehicles [2]) and terrestrial equipment such as tractors and combine harvesters [3]. Modern agricultural production systems use HTP data to optimize management practices [4], forecast agricultural outputs [5] and to assess the quality (e.g., protein content) of agricultural commodities [6]. 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. An extensive body of research deals with the use HTP data to predict phenotypes such as grain yield [5,7–9], dry matter [3], oil and protein content [10,11]. 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

60 correlated phenotypes was first addressed by Smith [12] and Hazel [13] who introduced the
61 concept of selection index (SI) in plant and animal breeding, respectively.

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

69 To address this problem we developed regularized selection indices (including penalized
70 and reduced-rank methods) that are tailored to achieve accurate prediction of genetic values
71 using high-dimensional phenotypes. The proposed methodology integrates into the SI framework
72 methods often used to prevent overfitting in high-dimensional phenotypic regressions [14].
73 Using extensive multi-environment crop imaging data from CIMMYT's wheat breeding program
74 we show that regularized SIs offer improved accuracy of indirect selection in both optimal and
75 stress environments.

76

77 ## Results

78 A **selection index** is a linear combination of p measured phenotypes, $\boldsymbol{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)'$, of the
79 form $I_i = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)'$ is a vector of regression coefficients whose entries define
80 the weights of each of the measured phenotypes in the SI. In a canonical SI those weights are

derived by minimizing the squared deviation between the genetic merit for the selection target (g_{y_i}, e.g., the genetic merit for grain yield of the i^{th} genotype) and the SI, that is:

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

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

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{P}_x^{-1}\boldsymbol{G}_{x,y}, \quad (2)$$

where $\boldsymbol{G}_{x,y} = \mathbb{E}\left(\boldsymbol{x}_i g_{y_i}\right) = \left(G_{x_1,y}, \ldots, G_{x_p,y}\right)'$ is a vector containing the genetic covariances between the selection objective (y_i) and each of the measured traits ($\boldsymbol{x}_i$), and $\boldsymbol{P}_x$ is the (population) phenotypic variance-covariance matrix of the measured phenotypes, that is, $\boldsymbol{P}_x = \mathbb{E}(\boldsymbol{x}_i\,\boldsymbol{x}_i') = \mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_i')$. Thus, a canonical SI takes the form $I_i = \boldsymbol{x}_i'\boldsymbol{P}_x^{-1}\boldsymbol{G}_{x,y}$. The theory underlying the derivation of SIs and response to indirect selection is well established [15,16].

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

Regularized selection indices

Reduced-rank (e.g., principal components methods) and penalized regression [14] 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.

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 expression (1). The solution of expression (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 $q = 1,2, \dots, p$ PCs ($q = p$ renders the canonical SI). For the sequence of estimates $(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \dots, \widehat{\boldsymbol{\beta}}^{(p)})$, 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.

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

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \left\{ \frac{1}{2} \mathbb{E}\left(g_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 + \lambda J(\boldsymbol{\beta}) \right\}, \quad (3)$$

where λ is a penalty parameter ($\lambda = 0$ yields the coefficients for the canonical 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 [17], or a weighted sum of the two [18].

Using $J(\boldsymbol{\beta}) = \frac{1}{2}\sum_{j=1}^{p} \beta_j^2$ in expression (3) renders a **Ridge-regression-type PSI** (RR-PSI, see *Methods*):

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

where $\boldsymbol{I}$ 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 [19]), or in combination with the L2-norm, $J(\boldsymbol{\beta}) = \frac{1}{2}(1-\alpha)\sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|$ (elastic-net [18]). 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. However, solutions for PSIs involving an L1-penalty can be obtained using iterative procedures such as the coordinate descent [20] and the least angle regression [21] (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.

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, $Acc(I) = cor(I_i, g_{y_i})$. This parameter is equal to the product of the square root of the heritability of the SI (h_I) times the genetic correlation between the SI and the selection target, $cor(g_{I_i}, g_{y_i})$ [16]. To avoid estimation bias $Acc(I)$ must be estimated using data that was not used to derive the coefficients of the index (Fig 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 ($I_i = \boldsymbol{x}_i'\boldsymbol{\beta}$), and (*iv*) estimated h_I, $cor(g_{I_i}, g_{y_i})$, and $Acc(I)$ in the testing set. Furthermore, we quantified the efficiency of indirect selection relative to mass phenotypic selection (RE) using $RE = \frac{h_I}{h_y} cor(g_{I_i}, g_{y_i})$ [16].

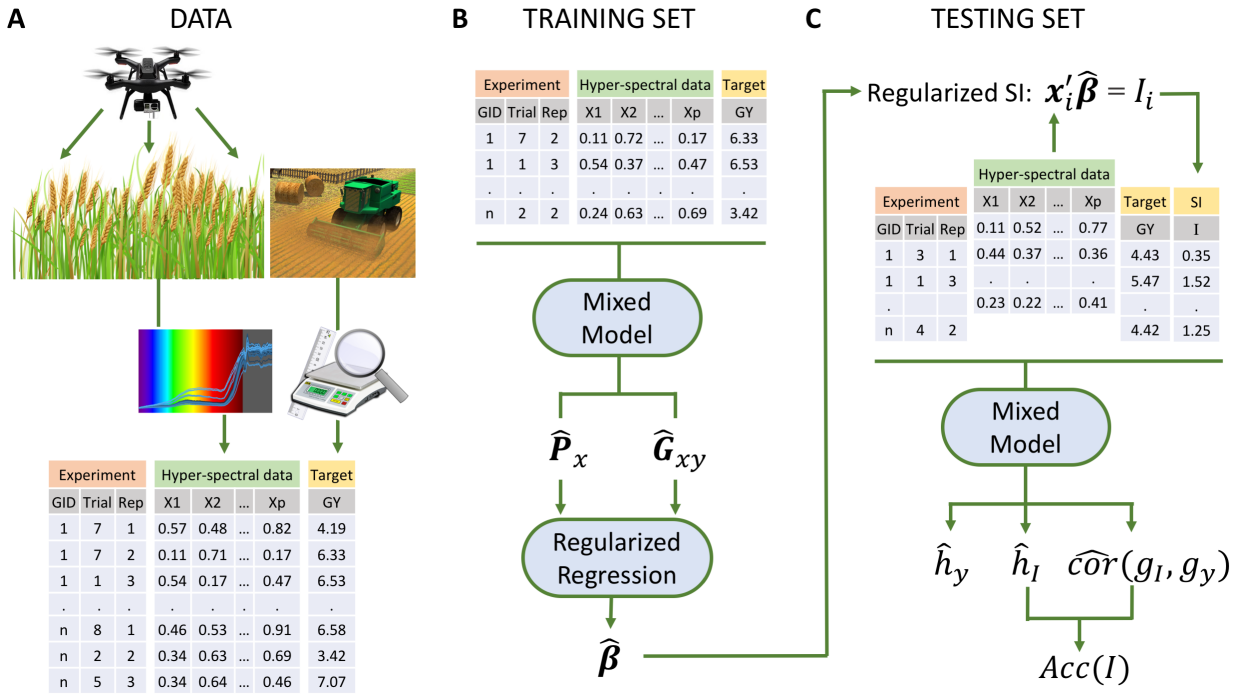

Fig 1. Prediction of the genetic merit for grain yield using hyper-spectral crop image data. (A) Data consists of hyper-spectral reflectance data ($\boldsymbol{x}_i$) and phenotypic measurements of the target trait (y_i, e.g., grain yield). (B) A subset of the data (the training set) is used to derive the coefficients ($\boldsymbol{\beta}$) of a selection index. (C) These coefficients are then applied to image data of individuals in the testing set to derive the index (I_i) for each individual. The predictive ability of the index is assessed by calculating the accuracy of indirect selection ($Acc(I)$) in the testing set.

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 Global Wheat Program consisting of grain yield ($\mathrm{ton\ ha^{-1}}$) 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 1 and S1 Fig for boxplots of yield by environment).

170 **Table 1 . Average grain yield and heritability by environmental condition**

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

171 SD: standard deviation.

172

173 Image data was collected using an infrared and an hyper-spectral camera and consisted
174 of reflectance of electromagnetic power at 250 wavelengths (or bands) within the visible and
175 near-infrared spectrums (392-850 nm). Separate images were collected at 9 time-points covering
176 vegetative (VEG), grain filling (GF), and maturity (MAT) stages of the crop (see S2 Fig). Grain yield
177 and image data were pre-adjusted using mixed-effects model that accounted for genotype, trial,
178 replicate, and sub-block (see *Methods* section).

179 **Regularization improves the heritability and the accuracy of the index**

180 To assess the effect of regularization on the accuracy of indirect selection we fitted an L1-PSI over
181 a grid of values of the regularization parameter ($\lambda^{(1)} > \lambda^{(2)} > \cdots > 0$ in expression (3), using $\lambda =$
182 0 renders a canonical SI). For each of the solutions ($\widehat{\boldsymbol{\beta}}(\lambda^{(1)})$, $\widehat{\boldsymbol{\beta}}(\lambda^{(2)}), \cdots$) we estimated the
183 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 Fig 2 (see S3-S5 Figs for other time-points). The heritability of the L1-PSI (Fig 2A) decreased as more bands became active in the index. Likewise, the heritability of PC-SI (Fig 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 (S3-S5 Figs) 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 (S1 Table). Importantly, in all cases the accuracy of indirect selection of the optimal regularized SIs was considerably higher than that of the canonical SI, which is the one corresponding to 250 active bands or 250 PCs (i.e., the right-most results in the plots in Fig 2).

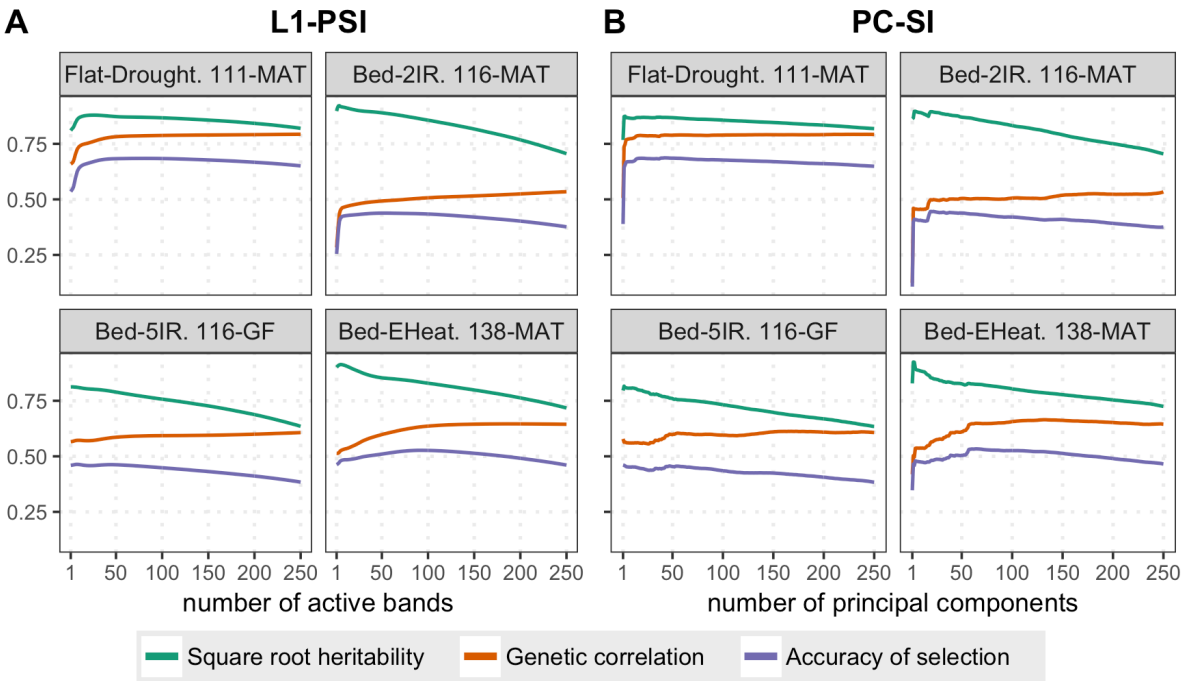

Fig 2. Accuracy of indirect selection of regularized SIs and its components. Square root heritability (green), genetic correlation (orange) and accuracy of indirect selection (purple), all averaged over 100 training-testing partitions, versus the number of predictors used to build the index: (A) number of active bands in the case of the L1-PSI, or (B) number of PCs in the PC-SI. Each panels represents one environment (latest time-point).

Fig 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 canonical SI (in green). 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 10-30% higher than the accuracy achieved by a canonical SI. 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 the canonical SI, regularized SIs had higher heritability (S6 Fig); this was achieved without compromising the genetic correlation (S7 Fig), thus leading to a001 higher accuracy of indirect selection achieved by either penalization or reduced-rank strategies.

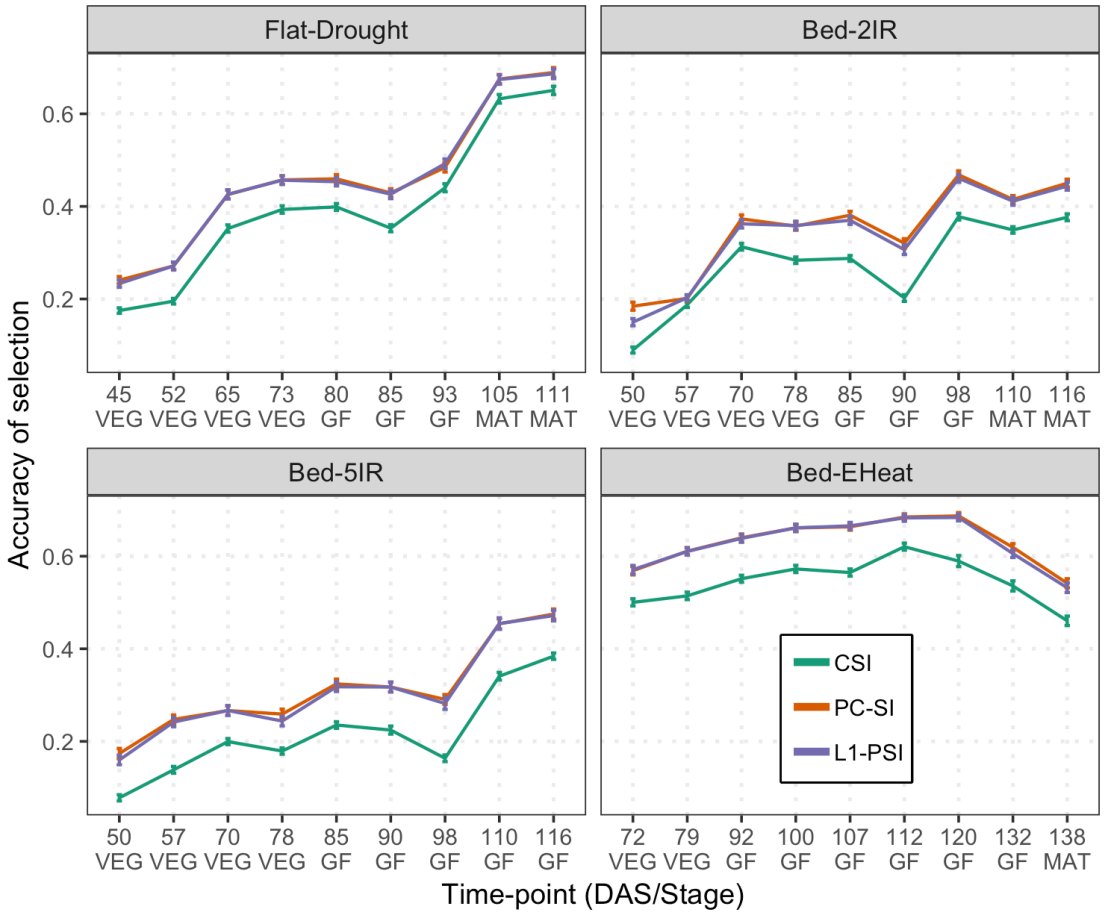

Fig 3. Accuracy of indirect selection achieved by a canonical (CSI) and by regularized (PC-SI and L1-PSI) selection indices. The lines provide the average accuracy over 100 training-testing partitions. Horizontal lines represent the 95% CI. The horizontal axis give the time-point at which images were collected and are expressed in both days after sowing (DAS) and stages (VEG=vegetative, GF=grain filling, MAT=maturity).

228 **Using data from multiple time-points further improves selection accuracy**

229 The results presented above were based on data from a single time-point. We also generated
230 selection indices using data from multiple time-points (in this case, $\boldsymbol{x}_i$ was a vector containing
231 2,250 traits, corresponding to 250 wavelengths measured at each of 9 time-points). Integrating
232 data from multiple time-points further increased the accuracy of L1-PSI by a margin that ranged
233 from 1 to 8 points on the correlation scale (Table 2). The gains in selection accuracy obtained
234 using data from multiple time-points were more evident in environments with lower accuracy;
235 similar results were obtained for the PC-SI and L2-PSI (S1 Table).

236

237 **Table 2. Accuracy and relative efficiency of indirect selection of an L1-penalized SI using data**
238 **from one and nine time-points.**

Environment	Accuracy (SD)		Relative Efficiency (SD)	
	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)

239 Values are presented as an average across 100 training-testing partitions. SD: standard deviation.
240 *For each environment we include the time-point that gave the highest accuracy of selection (see
241 Fig 3 for other time-points).

L1-penalization leads to sparse selection indices

Fig 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 bands 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 wavelengths 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 wavelengths in the green and infrared regions that were active. In all the indices, there were wavelengths 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.

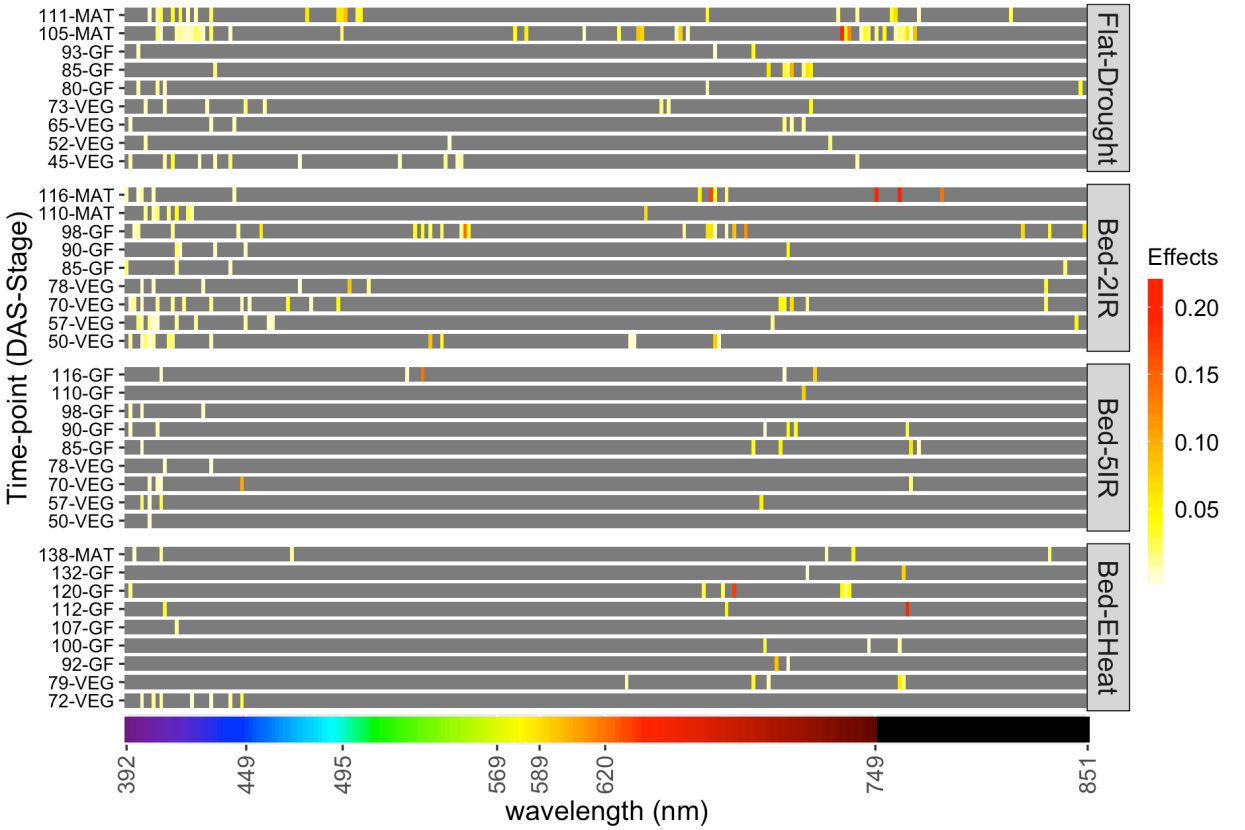

Fig 4. Heatmap of regression coefficients for L1-penalized selection indices. Separate indices were derived for each environment using multi time-point data. DAS=days after sowing, VEG, GF, MAT represent vegetative, grain-filling and maturity stages, respectively. The bottom color-bar shows the light color associated with each wavelength in the visible spectrum (≤ 750 nm); black was used to represent the near-infrared spectrum (wavelength > 750 nm).

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 (e.g., [22]), biomass, and chlorophyll content (e.g., [23,24]). We considered two vegetation indices: the Red

and Green Normalized Difference Vegetation Indices (RNDVI [25] and GNDVI [26] respectively). For each of these indices we estimated the genetic correlation with grain yield, as well as their heritability and accuracy of indirect selection (S1 Table). 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, S1 Table, 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 (S6 Fig); 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 (S7 Fig). 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 (S1 Table); however, in a few environments at some time-points, the L1-PSI achieved a higher accuracy of indirect selection than the phenotypic prediction.

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 phenotypic predictor is the sum of the best predictor of the genetic merit (g_y) plus the best 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 S8 Fig). 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 (S1 Table).

A canonical SI (expression (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 expression (2)) are known without error. However, when the measured phenotype is high-dimensional, estimation errors in the phenotypic (co)variance matrix ($\boldsymbol{P}_x$), as well as in the genetic covariances ($\boldsymbol{G}_{x,y}$), can make the canonical SI sub-optimal. Our empirical results confirm this: canonical 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 canonical 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 canonical 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 [27] or genomic-enabled [28]) 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, Misztal & Gengler [29] 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,

329 Dagnachew, Meuwissen & Ådnøy [30] suggested predicting the genetic merit for milk fatty acids
330 using FTIR-derived PCs as 'traits' in a genetic evaluation. However, when mapping from genetic
331 predictions of PC-lodgings onto genetic predictions for the selection objective the authors used
332 coefficients derived from a phenotypic (partial least squares) regression. This does not guarantee
333 that the resulting index is maximally correlated with the genetic merit of the selection target. The
334 penalized and PC-SI presented in this study address that problem by using coefficients that are
335 derived using genetic (and not phenotypic) covariances.

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

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

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

346 One possibility is to use **two-steps** whereas in the first step a PSI or a PC-SI is used to
347 predict the genetic merit using within-individual information. This step can be considered as a
348 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 $\{I_1, \dots, I_n\}$ could be used as a trait in a genetic evaluation.

A **one-step** approach is also conceptually possible: the optimization problem of expression (3) can be modified by replacing $\boldsymbol{x}_i$, the vector with the measured phenotypes on the i^{th} individual, with a vector $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \dots, \boldsymbol{x}_n')'$ that contains all the available HTP data (measured on n individuals); after expanding the squared error loss and taking expectations we get

$$\widehat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}_i} \left\{ \frac{1}{2}\mathbb{E}\left(g_{y_i}^2\right) - \boldsymbol{\beta}_i' \boldsymbol{G}_{gx} + \frac{1}{2}\boldsymbol{\beta}_i' \boldsymbol{P}_x \boldsymbol{\beta}_i + \lambda J(\boldsymbol{\beta}_i) \right\},$$

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

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

370 across hundreds or thousands of phenotypes (e.g., RNA-abundance across thousands of genes)
371 may critically compromise power. The PSI and PC-SI presented in this study could be used to
372 extract genetic patterns from high dimensional phenotype data such as brain imaging or whole-
373 genome gene expression profiles and these patterns can then be used as traits in genetic studies.

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

383

384 ## Methods

385 ### Canonical selection index

386 The weights on a SI are derived as the solution to the optimization problem of expression (1):

387 $$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \frac{1}{2}\mathbb{E}\left(g_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2.$$

388 The right-hand side can be expressed as $\mathbb{E}\left(g_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 = \mathbb{E}\left(g_{y_i}^2\right) - 2\,\mathbb{E}\left(g_{y_i}\boldsymbol{x}_i\right)'\boldsymbol{\beta} +$
389 $\boldsymbol{\beta}'\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')\boldsymbol{\beta}$. The first term, $\mathbb{E}\left(g_{y_i}^2\right)$, does not involve $\boldsymbol{\beta}$; therefore, it can be dropped from the

390 objective function. Furthermore, if $\boldsymbol{x}_i$ has null mean, and assuming that the environmental
391 effects on $\boldsymbol{x}_i$ are orthogonal to g_{y_i}, then $\mathbb{E}(g_{y_i}\boldsymbol{x}_i) = \boldsymbol{G}_{x,y}$ is a vector containing the genetic
392 covariances between the selection target and each of the measured phenotypes. Likewise,
393 $\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i') = \boldsymbol{P}_x$ is the phenotypic (co)variance matrix of $\boldsymbol{x}_i$. Therefore, the problem in expression
394 (1) can be written as

395 $$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta}\left\{-\boldsymbol{G}_{x,y}'\boldsymbol{\beta} + \tfrac{1}{2}\boldsymbol{\beta}'\boldsymbol{P}_x\,\boldsymbol{\beta}\right\}.$$

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

398 $$\widehat{\boldsymbol{\beta}} = \boldsymbol{P}_x^{-1}\boldsymbol{G}_{x,y}.$$

399 **Reduced-rank selection index**

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

407 $$\widehat{\boldsymbol{\gamma}}^{(q)} = \arg\min_{\gamma}\tfrac{1}{2}\mathbb{E}\left(g_{y_i} - \widetilde{\boldsymbol{w}}_i'\boldsymbol{\gamma}^{(q)}\right)^2,$$

where $\widetilde{\boldsymbol{w}}_i$ 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 $\widehat{\boldsymbol{\gamma}}^{(q)} = \boldsymbol{P}_{\widetilde{w}}^{-1}\boldsymbol{G}_{\widetilde{w},y}$, where $\boldsymbol{P}_{\widetilde{w}}$ is the phenotypic (co)variance matrix of the first q PCs and $\boldsymbol{G}_{\widetilde{w},y}$ 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 $\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{D}^2 = \text{diag}(d_1^2, \ldots, d_p^2)$. Hence, a method-of-moments estimate of the phenotypic (co)variance matrix of $\widetilde{\boldsymbol{W}}$ contains only the first q elements $\widetilde{\boldsymbol{D}}^2 = \text{diag}(d_1^2, \ldots, d_q^2)$; this is

$$\widehat{\boldsymbol{P}}_{\widetilde{w}} = \tfrac{1}{n-1}\widetilde{\boldsymbol{D}}^2.$$

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

Penalized selection indices

The objective function of the penalized SI is given by expression (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, $J(\boldsymbol{\beta}) = \frac{1}{2}\sum_{j=1}^{p}\beta_j^2 = \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$, in expression (3) leads to the following optimization problem:

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

Therefore:

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

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

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

where $\boldsymbol{I}$ 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: $(\boldsymbol{P}_x + \lambda\boldsymbol{I})\widehat{\boldsymbol{\beta}}^{L2} = \boldsymbol{G}_{x,y}$; therefore:

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

EN-PSI. The coefficients for the elastic-net family are obtained by considering an objective function as in expression (3), with $J(\boldsymbol{\beta}) = \frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|$; therefore,

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

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 [14]). These algorithms can be implemented either by 'partial residuals' or using 'covariance updates' [31]. In our case, the objective function is entirely based on (co)variance terms. The objects $\boldsymbol{P}_x$ and $\boldsymbol{G}_{x,y}$ enter in the objective function of the PSI in the same way that $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{y}$ enter in a standard elastic-net regression. Therefore, to obtain solutions, we implemented the standard LARS algorithm (e.g., Hastie *et al*. [14]) entirely based on covariance updates.

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 ($\mathrm{ton\ ha^{-1}}$, total plot yield after maturity) was recorded. Reflectance phenotypic 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 and March 27^{th}, 2014. During each flight, data from $p = 250$ wavelengths ranging from 392 to 850 nm were collected for each pixel in the pictures. The average reflectance of all the pixels for each wavelength was calculated from each of the geo-referenced trial plots and reported as each line reflectance. 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.

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

$$y^*_{jklm} = y_{jklm} - \hat{\mu} - \hat{t}_k - \hat{r}_{l(k)} - \hat{b}_{m(kl)} = \hat{g}_j + \hat{\varepsilon}_{jklm}. \quad (4)$$

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 wavelengths. 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.

Heritability estimation

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

$$y_{ij} = g_j + \varepsilon_{ij}, \quad (5)$$

where y_{ij} is the phenotype for the i^{th} observation (i here is a single index for indices k, l, and m in expression (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

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

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 trials to training or testing sets. Twenty-six trials ($n_{trn} \approx 2{,}184$ observations) were randomly assigned to the training set, and the remaining 13 trials ($n_{tst} \approx 1{,}092$) were used as the testing set. The regression coefficients of the indices (the $\boldsymbol{\beta}$'s for the canonical 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 $I_{ij} = \boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}}$, for each observation i in the testing set ($i = 1, \ldots, n_{tst}$). The heritability of the index and the genetic correlation between the index and the selection goal were estimated in the testing set.

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

512 **Estimation of phenotypic and genetic parameters**

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

518 The genetic covariance ($G_{x_j,y}$) between grain yield and the j^{th} measured trait ($j = 1, \dots, p$)
519 was estimated using a sequence of univariate genetic models as in expression (5). We fitted that
520 model with grain yield phenotypes as response, then for each of the reflectance bands and then
521 for the sum of grain yield and each of the bands. The genetic covariances between the bands and
522 grain yield were then estimated using

523 $$\widehat{\boldsymbol{G}}_{y,x_j} = \frac{1}{2}\left(\hat{\sigma}^2_{g_{y+x_j}} - \hat{\sigma}^2_{g_y} - \hat{\sigma}^2_{g_{x_j}}\right)',$$

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

526

527

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 $I_{ij} = \boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}}$. Then, using a mixed model analysis like that described in the previous section we estimated the heritability of the SI (h_I^2), the heritability of grain yield (h_y^2), and the genetic correlation between the SI and grain yield ($cor(g_{I_i}, g_{y_i})$). From these estimates, we derived the accuracy of indirect selection, $Acc(I) = h_I\, cor(g_{I_i}, g_{y_i})$, and the relative efficiency, $RE = \frac{h_I}{h_y} cor(g_{I_i}, g_{y_i})$.

Software

All the aforementioned analyses were implemented in the R software environment [32], version 3.5.1. Linear mixed models were implemented using the 'lmer' function from the LME4 [33] R-package. Code that implements LARS in the context of SIs was programed based on the LARS [34] R-package.

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. Data sets and R-scripts to perform all the analyses are publicly available upon request to the corresponding author.

Acknowledgments

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

Author contributions

RS, SD, JC and SM were involved in the design of the field experiments and data collection. SM performed the HTP data correction and georeferencing. MLC and GDLC conceived the idea, performed the analyses and produced a first draft, all the authors contributed to the final manuscript.

References

1. Nagel KA, Putz A, Gilmer F, Heinz K, Fischbach A, Pfeifer J, et al. GROWSCREEN-Rhizo is a novel phenotyping robot enabling simultaneous measurements of root and shoot growth for plants grown in soil-filled rhizotrons. Funct Plant Biol. 2012;39(11):891–904.

2. Araus L, Cairns JE. Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci. 2014;19(1):52–61.

3. Montes JM, Utz HF, Schipprack W, Kusterer B, Muminovic J, Paul C, et al. Near-infrared spectroscopy on combine harvesters to measure maize grain dry matter content and quality parameters. Plant Breed. 2006;125:591–5.

4. White JW, Andrade-Sanchez P, Gore MA, Bronson KF, Coffelt TA, Conley MM, et al. Field Crops Research Field-based phenomics for plant genetics research. F Crop Res.

2012;133:101–12.

5. Ferrio JP, Villegas D, Zarco J, Aparicio N, Araus JL, Royo C. Assessment of durum wheat yield using visible and near-infrared reflectance spectra of canopies. F Crop Res. 2005;94:126–48.

6. Spielbauer G, Armstrong P, Baier JW, Allen WB, Richardson K, Shen B, et al. High-throughput near-infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. Cereal Chem. 2009;86(5):556–64.

7. Garriga M, Romero-Bravo S, Estrada F, Escobar A, Matus IA, del Pozo A, et al. Assessing wheat traits by spectral reflectance: do we really need to focus on predicted trait-values or directly identify the elite genotypes group? Front Plant Sci. 2017;8(280):1–12.

8. Weber VS, Araus JL, Cairns JE, Sanchez C, Melchinger AE, Orsini E. Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes. F Crop Res. 2012;128:82–90.

9. Aguate FM, Trachsel S, González-Pérez L, Burgueño J, Crossa J, Balzarini M, et al. Use of hyperspectral image data outperforms vegetation indices in prediction of maize yield. Crop Sci. 2017;57:2517–24.

10. Garnsworthy PC, Wiseman J, Fegeros K. Prediction of chemical, nutritive and agronomic characteristics of wheat by near infrared spectroscopy. J Agric Sci. 2000;135:409–17.

11. Oblath EA, Isbell TA, Berhow MA, Allen B, Archer D, Brown J, et al. Development of near-infrared spectroscopy calibrations to measure quality characteristics in intact Brassicaceae

589 germplasm. Ind Crop Prod. 2016;89:52–8.

590 12. Smith HF. A discrimant function for plant selection. Ann Eugen. 1936;7:240–50.

591 13. Hazel LN. The genetic basis for constructing selection indexes. Genetics. 1943;28(6):476–
592 90.

593 14. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining,
594 inference, and prediction. 2nd ed. New York, USA: Springer; 2009. 745 p.

595 15. Bulmer MG. The mathematical theory of quantitative genetics. New York, USA: Oxford
596 University Press; 1985.

597 16. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. Essex, UK: Prentice
598 Hall; 1996.

599 17. Fu WJ. Penalized regressions: the Bridge versus the LASSO. J Comput Graph Stat.
600 1998;7(3):397–416.

601 18. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc B.
602 2005;67(2):301–20.

603 19. Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc B.
604 1996;58(1):267–88.

605 20. Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. Ann Appl
606 Stat. 2007;1(2):302–32.

607 21. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat.
608 2004;32(2):407–99.

609 22. Tattaris M, Reynolds MP, Chapman SC. A direct comparison of remote sensing approaches
610 for high-throughput phenotyping in plant breeding. Front Plant Sci. 2016;7:1–9.

611 23. Babar MA, Reynolds MP, Ginkel M Van, Klatt AR, Raun WR, Stone ML. Spectral reflectance
612 to estimate genetic variation for in-season biomass, leaf chlorophyll, and canopy
613 temperature in wheat. Crop Sci. 2006;46:1046–57.

614 24. Haboudane D, Miller JR, Tremblay N, Zarco-Tejada PJ, Dextraze L. Integrated narrow-band
615 vegetation indices for prediction of crop chlorophyll content for application to precision
616 agriculture. Remote Sens Environ. 2002;81:416–26.

617 25. Tucker CJ. Red and photographic infrared linear combinations for monitoring vegetation.
618 Remote Sens Environ. 1979;8(2):127–50.

619 26. Gitelson AA, Kaufman YJ, Merzlyak MN. Use of a green channel in remote sensing of global
620 vegetation from EOS-MODIS. Remote Sens Environ. 1996;58(3):289–98.

621 27. Henderson CR, Quaas RL. Multiple trait evaluation using relatives' records. J Anim Sci.
622 1976;43(6):1188–97.

623 28. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-
624 wide dense marker maps. Genetics. 2001;157(4):1819–29.

625 29. Soyeurt H, Misztal I, Gengler N. Genetic variability of milk components based on mid-
626 infrared spectral data. J Dairy Sci. 2010;93(4):1722–8.

627 30. Dagnachew BS, Meuwissen THE, Ådnøy T. Genetic components of milk Fourier-transform
628 infrared spectra used to predict breeding values for milk composition and quality traits in

dairy goats. J Dairy Sci. 2013;96(9):5933–42.

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

32. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.

33. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67(1):1–48.

34. Hastie T, Efron B. lars: least angle regression, Lasso and forward stagewise [Internet]. 2013. Available from: https://cran.r-project.org/package=lars

Supporting information

S1 Fig. Box-plot of grain yield phenotypic records by environmental condition. $n \approx 3200$. SD: standard deviation.

S2 Fig. Light reflectance patterns as function of the wavelength. Each line represents the mean (across $n \approx 3200$ observations) for each wavelength, within time-point (flight date). Within each environment, means were scaled to lie within 0 and 1 by dividing them by the maximum average.

S3 Fig. 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.

649 **S4 Fig. Accuracy of indirect selection of L2-PSI and its components.** Square root heritability,
650 genetic correlation and accuracy of indirect selection, all averaged over 100 training-testing
651 partitions versus the penalization parameter (λ, logarithm scale) used to build the index; by time-
652 point (DAS=days after sowing, Stage: VEG=vegetative, GF=grain filling, or MAT=maturity) within
653 environment.

654 **S5 Fig. Accuracy of indirect selection of PC-SI and its components.** Square root heritability,
655 genetic correlation and accuracy of indirect selection, all averaged over 100 training-testing
656 partitions versus the number of principal components used to build the index; by time-point
657 (DAS=days after sowing, Stage: VEG=vegetative, GF=grain filling, or MAT=maturity) within
658 environment.

659 **S6 Fig. Square root of heritability of the canonical (CSI), of the regularized (PC-SI and L1-PSI)**
660 **selection indices, and of the RNDVI.** The lines provide the average square root heritability over
661 100 training-testing partitions. Horizontal lines represent the 95% CI. The horizontal axis give the
662 time-point at which images were collected and are expressed in both days after sowing (DAS)
663 and stages (VEG=vegetative, GF=grain filling, MAT=maturity).

664 **S7 Fig. Genetic correlation between grain yield and all: the canonical (CSI), the regularized (PC-**
665 **SI and L1-PSI) selection indices, and the RNDVI.** The lines provide the average genetic correlation
666 over 100 training-testing partitions. Horizontal lines represent the 95% CI. The horizontal axis
667 give the time-point at which images were collected and are expressed in both days after sowing
668 (DAS) and stages (VEG=vegetative, GF=grain filling, MAT=maturity).

669 **S8 Fig. Phenotypic, genetic, and environmental covariances between wavelengths and grain**
670 **yield**. 'D': discrepancy between phenotypic and genetic covariances as measured by the sum of
671 the absolute differences; by time-point (DAS: days after sowing, Stage: VEG=vegetative, GF=grain
672 filling, MAT=maturity) within environment.

673 **S1 Table. Accuracy of indirect selection (average over 100 training-testing partitions) for best**
674 **phenotypic prediction (principal components (PCR), L1-penalized prediction (L1-Phen), RNDVI,**
675 **and GNDVI) and for best genotypic prediction (canonical SI (CSI), optimal PC-SI, L1-PSI, and L2-**
676 **PSI).** Each row contains results for each environment and time-point (DAS: days after sowing,
677 Stage: VEG= vegetative, GF=grain filling, MAT=maturity). Models with the same letter (within
678 each row) are not significantly different from each other (α=0.05, ANOVA followed by Tuckey
679 test).

680

681