Phenomic selection in wheat breeding: identification and optimisation of factors influencing prediction accuracy and comparison to genomic selection

P. Robert[1,2], J. Auzanneau[3], E. Goudemand[4], F.X. Oury[2], B. Rolland[5], E. Heumez[6], S. Bouchet[2], J. Le Gouis[2], R. Rincent[1,2, #]

[1] Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE - Le Moulon, 91190, Gif-sur-Yvette, France

[2] INRAE - Université Clermont-Auvergne, UMR1095, GDEC, 5 chemin de Beaulieu, 63000 Clermont-Ferrand, France

[3] Agri-Obtentions, Ferme de Gauvilliers, 78660, Orsonville, France

[4] Florimond-Desprez Veuve & Fils SAS, 3 rue Florimond-Desprez, BP 41, 59242, Cappelle-en-Pévèle, France

[5] INRAE–Agrocampus Ouest-Université Rennes 1, UMR1349, IGEPP, Domaine de la Motte, 35653 Le Rheu, France

[6] INRAE, UE 972, Grandes Cultures Innovation Environnement, 2 Chaussée Brunehaut, 80200 Estrées-Mons, France

Corresponding authors: renaud.rincent@inrae.fr

ORCID : P. Robert : 0000-0002-6314-180X / J. Le Gouis : 0000-0001-5726-4902 / S. Bouchet : 0000-0001-5868-3359 / R. Rincent : 0000-0003-0885-0969

Abstract

Key message

Phenomic selection accurately predicts heading date and grain yield of wheat breeding lines. Combining spectra from different environments and optimising the training set maximise the predictive ability of the method.

Phenomic selection (PS) is a recent breeding approach similar to genomic selection (GS) except that genotyping is replaced by near infrared (NIR) spectroscopy. PS can potentially account for non-additive effects and has the major advantage of being low cost and high throughput. Factors influencing GS predictive abilities have been intensively studied, but little is known about PS. We tested and compared the abilities of PS and GS to predict grain yield and heading date from several datasets of bread wheat lines corresponding to the first or second years of trial evaluation from two breeding companies and one research institute in France. We evaluated several factors affecting PS predictive abilities including the possibility of combining spectra collected in different environments. A simple H-BLUP model predicted both traits with prediction ability from 0.26 to 0.62, and with an efficient computation time. Our results showed that the environments in which lines are grown had a crucial impact on predictive ability based on the spectra acquired and was specific to the trait considered. Models combining NIR spectra from different environments were the best PS models and were at least as accurate as GS in most of the datasets. Furthermore, a GH-BLUP model combining genotyping and NIR spectra was the best model of all (prediction ability from 0.31 to 0.73). We demonstrated also that as for GS, the size and the composition of the training set has a crucial impact on predictive ability. PS could therefore replace or complement GS for efficient wheat breeding programs.

Key words:

Bread wheat, Genomic selection (GS), Genomic-like omics-based (GLOB) prediction, Near infrared spectroscopy (NIRS), Phenomic selection (PS), Plant breeding.

Declarations

Funding: This work was funded by Agri-Obtentions, Florimond-Desprez, and the Association Nationale de la Recherche et de la Technologie (ANRT, grant number 2019/0060)

Conflicts of interest/Competing interests: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Availability of data and material: The datasets generated during and/or analysed during the current study are not publicly available due breeding programs privacy but are available from the corresponding author on reasonable request.

Code availability: Code used to lead the analysis of this study is available from the corresponding author on request.

Author contribution statement: Author contribution statement JA, FXO, BR and EH designed the field trials and collected the phenotypic data from Agri-Obtentions and INRAE. EGD provided the phenotypic data and genotyping data from Florimond Desprez company. SB provided the genotyping data from Agri-Obtentions and INRAE and participate in discussions of this study. RR initiated the project and with JLG supervised the study and helped improving the manuscript. PR analysed the data and wrote the manuscript. All authors approved the final manuscript.

INTRODUCTION

In plant breeding, phenotyping is a key element for breeders to select candidates. The creation of new varieties that outperform the control varieties rely on the ability of breeders to evaluate numerous candidates to maximise the chance of success. It is particularly difficult for complex traits interacting with the environmental conditions, as in this case the best variety is not the same from one environment to another (Allard and Bradshaw 1964). Phenotyping is costly, time consuming and requires a large amount of seed, so the number of candidates phenotyped is one of the main limiting factors. Prediction approaches such as genomic selection (GS) have been proposed in the last two decades to overcome this constraint. The aim of GS is to predict the performances of unphenotyped candidates based on molecular markers (Whittaker et al. 2000; Meuwissen et al. 2001). In GS, a genotyped and phenotyped training set (TS) of varieties is used to train a statistical model. This model can then be applied to predict the genomic estimated breeding values of a predicted set using genotyping information alone. GS can improve genetic gain by increasing the intensity of selection with more candidates or by reducing the duration of the breeding cycles. The efficiency of GS depends on the prediction accuracy of the model, which can be influenced by multiple factors: the genetic architecture of the trait (Daetwyler et al. 2010) and its heritability (Hayes et al. 2009b), linkage disequilibrium (Zhong et al. 2009), marker density (Solberg et al. 2008; de los Campos et al. 2013; Hickey et al. 2014), the size and composition of the training set (Albrecht et al. 2011; Heffner et al. 2011; Pszczola et al. 2012; Rincent et al. 2012; Daetwyler et al. 2013), and the statistical model (Charmet et al. 2020).The efficiency of GS has been evaluated in breeding (Hayes et al. 2009a; Jannink et al. 2010; Crossa et al. 2010, 2017) and pre-breeding programs (Gorjanc et al. 2016; Yu et al. 2016) with promising results, but implementation of the approach remains difficult for some species due to the expense of genotyping tools. Small breeding companies may not have the budget for genotyping thousands of candidates each year (R2D2 Consortium et al. 2021). Furthermore, in GS, the biological processes occurring between the genotype level and the phenotype level are ignored. One can then wonder if other types of information could better link genotypes to phenotypes considering non-additive effect for example.

Endophenotype refers to an intermediate trait level between the genome and the phenotype, resulting from the association of a genotype, an environment, and the interaction between them (Fernandez et al. 2016). Some studies have shown that molecular markers can be replaced by endophenotypes such as transcripts (Frisch et al. 2010; Fu et al. 2012; Zenke-Philippi et al. 2017; Azodi et al. 2020), metabolites (Riedelsheimer et al. 2012; Xu et al. 2016), or both (Guo et al. 2016; Westhues et al. 2017; Schrag et al. 2018), or small RNA (Schrag et al. 2018; Seifert et al. 2018) when predicting complex traits in maize hybrids. By depicting interactions between genes and the environment and the plant regulatory system, these types of omics data are thought to capture non-additive relationships that are difficult to model with molecular markers alone, and the resulting predictions are often at least as accurate. The omics data are usually collected at an early stage of plant development from different tissues (leaves,

roots) of genotypes grown in controlled conditions. However, like in GS, the controlled conditions of omics acquisition may be radically different from the field trial environment in which the training set is phenotyped. This genomic-like omics-based (GLOB) prediction approach is however effective in capturing genetic similarity potentially related to both additive and non-additive effects.

The routine acquisition of endophenotypic data is expensive, so it is still challenging to implement them in breeding. Rincent et al. (2018) proposed a low-cost and efficient alternative, called phenomic selection (PS), in which genotyping or other omics are replaced by near infrared spectra. Near infrared spectroscopy (NIRS) is a non-destructive method to measure the reflectance or absorbance of a biological sample at different wavelengths in the visible and near-infrared spectrum. According to the Beer-Lambert law, reflectance and absorbance are directly linked to the molecular composition of the sample, which for plants is influenced by genetics, environmental factors and their interaction. NIR spectra are already routinely collected in cereal breeding, for instance, to predict grain composition and quality (Osborne 2006). The accuracies of PS and GS in predicting different traits in bread wheat and poplar populations were compared in two scenarios (Rincent et al. 2018). In the first scenario, predictions were made within one environment, that is, the plants or plots that were predicted are those from which the NIR spectra were acquired. The second scenario was a GLOB approach in which spectra were used instead of endophenotypes, so spectra were collected in at least one environment that may have differed from the environments in which the training set was phenotyped. Spectra were used to estimate genetic similarities between varieties. PS outperformed GS in predicting wheat grain yield (GY) in both scenarios. A simple mixed model was used in which the similarity matrix was computed with molecular marker information for GS (G-BLUP) and with NIR spectra for PS (H-BLUP). As molecular markers are considered as quantitative variables in GS (with regression according to allele counts), most GS models including RR-BLUP, G-BLUP, BayesA, B, C and Cπ can be used in PS by replacing genotyping data by NIRS measurements.

There are only a few studies of PS so far. In some PS studies, spectra were acquired in the same environment as the predicted trait (as in scenario 1 above). For the best statistical model based on scenario 1, PS results were promising with GY prediction accuracies ranging from 0.37 to 0.51 in wheat (Rincent et al. 2018; Krause et al. 2019; Cuevas et al. 2019). By comparison, a GLOB approach produced wheat GY prediction accuracies ranging from 0.28 to 0.53 (Rincent et al. 2018). Galán et al. (2020) adjusted mean spectra for each genotype by correcting environment and spatial effects and predicted dry matter yield in winter rye with prediction accuracies ranging from 0.48 to 0.59. Lane et al. (2020) also applied a GLOB prediction approach in which a set of spectra from one year was used to predict maize GY for another year using partial least square regression. The resulting prediction accuracies ranged from 0.19 to 0.69. They also averaged the relationship matrices computed on NIR spectra from several environments to predict maize grain yield. Predictive accuracies of H-BLUP model ranged 0.07 to 0.70 for all environments and two cross-validation scenarios.

As for GS, some factors have already been identified as impacting PS accuracy. Specifically, PS predictive ability (PA) using NIRS was strongly influenced by the trait and tissue considered, including the environment in which plants were grown (Rincent et al. 2018). Drought environments appeared to be more useful than irrigated environments, for example. Galán et al. (2020) compared different training population sizes to make predictions with H-BLUP and G-BLUP and found that prediction abilities increased the larger the training set. Cuevas et al. (2019) pre-processed NIR spectra with different mathematical transformations and found the predictive abilities varied according to predictive model used.

Here we investigated the ability of PS to predict the traits of GY and HD in elite breeding lines in winter wheat produced by two breeding companies (Agri-Obtentions and Florimond-Desprez) and one research institute (INRAE) in France. The accuracy of any predictive model used in breeding will have an impact on genetic gain and selection efficiency, so we explored which factors alter the predictive abilities of PS. The objectives were to investigate how the predictive ability of statistical models was affected by the assumptions made, or by how the NIR spectra were collected (environment or generation of plant sample) and used (singly or in combination). The second objective was to compare the predictive abilities of PS, GS, and models combining variables related to both NIR spectra and molecular markers. The effects of the size and the composition of the training set on the GS and PS predictive abilities were also assessed.

Materials and Methods

Phenotyping, genotyping and NIRS data

Genetic material and phenotypic data

We used four different datasets corresponding to the first or second year of trial evaluation of bread wheat candidate breeding lines. These lines (pedigree and double haploid) were developed in the breeding programs of Florimond Desprez (France), and Agri-Obtentions (France) together with INRAE (France). Data were grouped according to breeding program and according to the year of trial evaluation. For each dataset, the year, site, trial treatments, the number of lines, the number of lines genotyped and the traits evaluated are listed (Table 1).

Set1 is data from the first year of trial evaluation of breeding lines developed by Florimond Desprez between 2016 and 2018. Data were collected from six environments (two sites in three years). Grain yield (GY) and heading date (HD) were evaluated for each breeding line in an augmented design with a control repeated every five microplots. NIR spectra are available for all lines in all the environments between 2016-2018 and also for the preceding year when breeding lines were in nursery, for the 2017-2018 years.

Set2 is data from the first year of trial evaluation of breeding lines developed by Agri-Obtentions and INRAE in 2018 and 2019. It is composed of eight environments, involving four sites and two years. EM site is a reference site for evaluating all the lines that were evaluated in the same year at the three other sites. GY was evaluated in an augmented design in the reference site and in randomized complete blocks for the other sites. HD data was missing for some of the environments, so to simplify the analysis, HD was not predicted for Set2. NIR spectra are available for all the genotypes and all the repetitions in all the environments.

Set3 is data from the second year of trial evaluation of breeding lines developed by Florimond Desprez between 2016 and 2018. It is composed of nine environments, involving four sites and three years. GY and HD were evaluated in lattice or randomised block designs. NIR spectra are available for all the lines in all the environments, and also in the preceding year when breeding lines were in nursery, for 2017-2018 years.

Set4 is data from the second year of trial evaluation of breeding lines developed by Agri-Obtentions and INRAE in 2018 and 2019. This dataset is composed of seven environments, involving two sites, two treatments (intensive practices versus low input) and two years. One environment was removed from the analysis due to mis-association between genotypes and phenotypes in the data that could not be corrected. GY and HD were evaluated in each environment in three complete randomised blocks. NIR spectra are available for all the lines in all the environments and also in the preceding year when breeding lines were in nursery, for 2019 year.

GY data was adjusted to a humidity rate of 15% and HD was scored on each plot as the date when 50% of the ears were half emerged. All the trial details can be found in Supplemental data Tables S1 & S2.

Genotypic data

Lines in Set1 and Set3 were genotyped with the 35K breeder Bristol array (Axiom™ Wheat Breeder's Genotyping Array). Lines in Set2 and Set4 were genotyped with the 35K BreedWheat array (Axiom™ BreedWheat Genotyping Array) which is a subset of the 280K SNP array (Rimbert et al. 2018). This subset corresponds to a selection of markers minimizing linkage disequilibrium between them (Ben-Sadoun et al. 2020). Molecular markers were eliminated if the minor allele frequency was less than 5%, or if the heterozygosity rate or missing value rate was greater than 5%. After filtering for high-resolution SNPs, we obtained 5,824 SNPs in both Set1 and Set3, 12,303 SNPs in Set4-2018, and 19,512 SNPs for both Set2 and Set4-2019. On average over the sets, a proportion of 1.1% of missing values were imputed with the average allele frequency of the corresponding marker. Minor allele frequency, missing value and mean imputation were computed with the sommer R package (Covarrubias-Pazaran 2016)

Table 1 Summary of the four datasets resulting from the trial evaluations. All lines evaluated were either double haploid or self-pollinated pedigree lines. **T**, treated (equivalent to intensive practices). **LI**, low input. **NT**, not treated (equivalent to intensive practices without fungicide treatment). Summary statistics were calculated on the adjusted means (models M1-M3 in Table 2). **GY**, grain yield in t/ha. **HD**, heading date in number of days after the 1st of January. **μ**, mean. $\boldsymbol{\sigma_\epsilon}$, residual standard deviation. **CV**, coefficient of variation (%) as $CV = \frac{\sigma_\epsilon}{\mu} \times 100$. See Supplementary Table S1 for the description of the sites.

Panel	Year	Site	Treatment	Number of lines	Lines genotyped	Trait					
						GY			HD		
						μ	σ_ϵ	CV	μ	σ_ϵ	CV
Set1	2016	HV	T	234	199	7.7	0.4	5.2	143	0.7	0.5
		LC		269	209	8.5	0.8	9.4	121	1.4	1.2
	2017	HV	T	280	157	10.7	0.3	2.8	140	0.8	0.6
		LC		184	90	7.9	0.5	6.3	124	1.2	1.0
	2018	HV	T	125	-	10.4	0.4	3.8	137	0.4	0.3
		LT		113		8.2	0.5	6.1	125	1.3	1.0
Set2	2018	CF	LI	328	-	7.5	0.3	4.0	-	-	-
		GL	T	566		9.5	0.3	3.2	-	-	-
		RE	LI	131		6.1	0.3	4.9	-	-	-
		EM	LI / NT	1351		8.3	0.5	6.0	-	-	-
	2019	CF	LI	264	-	5.3	0.4	7.5	-	-	-
		GL	T	433	325	10.5	0.4	3.8	-	-	-
		RE	LI	113	-	7.2	0.4	5.6	-	-	-
		EM	LI / NT	1418	-	8.2	0.5	6.2	-	-	-
Set3	2016	CP	T	95	-	9.0	0.3	3.3	145	0.9	0.6
		HV		103		7.7	0.4	5.2	140	0.7	0.5
		LC		141		9.1	0.5	5.5	120	4	3.3
	2017	CP	T	77	49	12.6	0.3	2.4	144	0.7	0.5
		HV		82	49	10.9	0.2	1.8	140	0.6	0.4
		LC		94	57	7.8	0.4	5.1	122	1.2	1.0
		VB		104	57	8.8	0.2	2.3	133	0.8	0.6
	2018	CP	T	101	69	11.0	0.2	1.8	145	2.2	1.5
		HV		102	70	10.6	0.3	2.8	136	3.1	2.3
Set4	2018	ML	T	82	71	9.7	0.5	5.2	141	0.2	0.1
		EM	LI			8.0	0.4	5.0	139	0.1	0.1
		ML				5.8	0.5	8.6	141	0.1	0.1
	2019	EM	T	111	100	10.4	0.3	2.9	144	0.1	0.1
		LM		112	101	11.5	0.4	3.5	145	0.5	0.3
		EM	LI	112	101	8.2	0.4	4.9	144	0.1	0.1
		LM		112	101	10.4	0.4	3.8	145	0.2	0.1

NIRS data

For NIRS acquisition, all flour and grain samples were stored in dry conditions at ambient temperature. NIRS of Set1 and Set3 were collected with the NIRS 6500 FOSS spectrometer (FOSS NIR Systems, Silver Spring, MD, USA) over the range 400 to 2500 nm in steps of 2 nm. All the NIRS were collected on 10 g of wheat flour except for Set1-2019 where 40 g of grain was used. Final spectra are the average of 32 repetitions measured by the spectrometer. NIRS of Set2 were collected on 150 g of grain for each genotype with the XDS NIR Analysers FOSS spectrometer (FOSS NIR Systems, Silver Spring, MD, USA) in the range 400 to 2500 nm in steps of 2 nm. Final spectra are the average of 16 repetitions measured by the spectrometer. For these three sets, NIRS were exported with ISIscanTM and WINisiTM 4.20 (Infrasoft International). NIRS of Set4 were collected on 350 g of grain for each genotype with the MPA II FT-NIR analyser (Bruker Optics, Ettlingen, Germany) ranging from 3594.92 cm^{-1} to 12489.60 cm^{-1} in steps of 7.7cm^{-1}. Final spectra are the average of 64 repetitions measured by the spectrometer. NIRS were exported with the OPUS LAB software provided by Bruker. To harmonise the data, NIRS data of Set4 were converted into nm in steps of 2 nm, so the final spectra ranged from 802 to 2492 nm. Spectra were visualised to filter out any spectra with abnormal absorbances resulting from technical errors. Where possible, we computed adjusted means of NIR absorbance for each wavelength when NIRS were measured on each microplot of the experimental design using the corrective model applied for GY and HD (Table 2). It was necessary to transform raw spectra because of inherent noise in the absorbance measurement. We first applied the standard normal variate transformation, which scales and centres each spectrum. Then we applied the Savitzky-Golay filter (Savitzky and Golay 1964) to compute the derivative with a window length of 61, using the R package signal (Signal developers 2013). All the NIRS were processed in R software (R Core Team 2019).

Statistical models for adjusting data and estimating heritability

Different statistical models adapted to each experimental design were used to correct traits for environmental effects, and to obtain the adjusted mean and broad-sense heritability of GY and HD. In environments with an augmented design, a two-dimensional P-spline mixed model was used to compute adjusted means, simultaneously accounting for global and local spatial trends across row and column effects (Model M1). In environments with a randomised complete block, global spatial effects were corrected for block and sub-block effects using the checks and replicated lines (M2). For the other environments with a randomised complete block with row and column information, a two-dimensional P-spline mixed model including block and sub-block effects was used to adjust for global and local trends (M3). Details of the traits and absorbance adjustments are summarised in Table S2.

After correcting for spatial effects, heritability within each environment was calculated for GY and HD. The following mixed model (Equation 1) was fitted to the corrected phenotypes in order to estimate the genetic (σ_G^2) and error (σ_e^2) variances. The controls were excluded from this step to focus on the variance of the breeding lines.

$$\hat{Y}_{ik} = \mu + G_i + \epsilon_{ik} \quad (1)$$

$$G_i \sim N(0, I\sigma_G^2) \text{ and } \epsilon_{ik} \sim N(0, I\sigma_e^2)$$

where $\hat{Y}_{ik}$ is the trait corrected for spatial effects for genotype i and rep k, G_i the random genetic effect for genotype i, and ϵ_{ik} the residual effect.

Broad sense heritability (H^2) was estimated as follows:

$$H^2 = \frac{\hat{\sigma}_G^2}{\hat{\sigma}_G^2 + \frac{\hat{\sigma}_\epsilon^2}{Nrep}} \quad (2)$$

where $\hat{\sigma}_G^2$ and, $\hat{\sigma}_\epsilon^2$ are the Restricted Maximum Likelihood of the genetic and residual variances derived by fitting Equation 1, respectively, and $Nrep$ is the average number of replicates in the corresponding environment. Details of heritability and variances are given in Table S3. Models M1 and M3 were fitted with the SpATS R package (Rodríguez-Álvarez et al. 2018) and Model M2 and Equation 1 were fitted with the breedR package (Facundo Muñoz and Leopoldo Sanchez 2020).

Table 2 Statistical models for computing adjusted means for GY, HD, and absorbance at each wavelength. y is a vector of the trait observations. β is a vector of the intercept and genotype fixed effect and X is the corresponding design matrix for the fixed effect. $f(r.c)$ represents a smooth bivariate function on a vector of rows r and a vector of columns c. $\boldsymbol{R}$ is the random effect of rows such that $\boldsymbol{R}\sim N(0, I\sigma_r^2)$ with the corresponding design matrix $\boldsymbol{Z_r}$. $\boldsymbol{C}$ is the random effect of columns such that $\boldsymbol{C}\sim N(0, I\sigma_c^2)$ with the corresponding design matrix $\boldsymbol{Z_c}$. $\boldsymbol{B}$ is the random effect of the block design such that $\boldsymbol{B}\sim N(0. I\sigma_B^2)$ with the corresponding design matrix $\boldsymbol{Z_j}$. $\boldsymbol{SB}$ is the random effect of sub-blocks in corresponding blocks such that $\boldsymbol{SB}\sim N(0. I\sigma_{Sb}^2)$ with the corresponding design matrix $\boldsymbol{Z_k}$. Finally, in all models, $\boldsymbol{\epsilon}\sim N(0, I\sigma_e^2)$ is the random residual effect.

Model	Equation
M1	$y = X\beta + f(r.c) + \boldsymbol{Z_r R} + \boldsymbol{Z_c C} + \boldsymbol{\epsilon}$
M2	$y = X\beta + \boldsymbol{Z_j B} + \boldsymbol{Z_k SB} + \boldsymbol{\epsilon}$
M3	$y = X\beta + \boldsymbol{Z_j B} + \boldsymbol{Z_k SB} + f(r.c) + \boldsymbol{Z_r R} + \boldsymbol{Z_c C} + \boldsymbol{\epsilon}$

Calculation of the genetic and hyperspectral relationship matrices

Relationship between individuals was estimated with data from molecular markers (kinship matrix *K*) and from NIRS (hyperspectral relationship matrix *H*). For markers, the relationship matrix was estimated following the Endelman and Janninck (2012) equation :

$$K = \frac{AA'}{2\sum p_k(1-p_k)}$$

where A is a centred matrix with dimensions $N \times M$, N is the number of genotypes and M the number of molecular markers. For the i^{th} individual and the k^{th} marker, $A_{ik} = X_{ik} + 1 - 2p_k$ with X the genotype matrix, coded in {-1,0,1} and p_k the frequency of allele 1 at marker k. K was computed with the rrBLUP R package (Endelman 2011).

For NIRS, the relationship matrix was computed as:

$$H = \frac{S^* S^{*\prime}}{L}$$

where S^* is the centred and scaled matrix of NIRS (dimension $N \times L$), and L the number of wavelengths.

K and *H* were both scaled to have a sample variance of 1 to avoid biased parameter estimations due to different scalings (Kang et al. 2010; Forni et al. 2011).

Reference prediction models and prediction scenario

For genomic prediction, we applied two reference models (G-BLUP, GS-LASSO), suited to contrasted genetic architectures, to predict GY and HD:

G-BLUP: $\hat{Y}_i = \mu + G_i + \epsilon_i$ with $G\sim N(0, K\sigma_G^2)$ and $\epsilon\sim N(0, I\sigma_\epsilon^2)$

GS-LASSO: $\hat{Y}_i = \mu + \sum_{k=1}^{p} x_{ik}\,\beta_k + \epsilon_i$ where

$\hat{\beta}(\lambda) = \text{Argmin}_{\beta\in\mathbb{R}^p}\{\|Y - X\beta\|^2 + \lambda\|\beta\|\}$ with $\|\beta\| = \sum_{k=1}^{p}|\beta_k|$ and $\lambda > 0$.

For G-BLUP, μ is the intercept, G_i is a random polygenic effect for the variety i and K the relationship matrix calculated with molecular markers (see above). For LASSO, x_{ik} is the allele of marker k for the variety i and β_k the effect of the marker k on the trait. For all models, $\hat{Y}_i$ is the adjusted mean of the trait for variety i, and ϵ_i the residual.

These two models were adapted for phenomic selection. H-BLUP is an adaptation of G-BLUP with *H* instead of *K* as covariance matrix. PS-LASSO is an adaptation of GS-LASSO using each wavelength as predictive variable

s_{ik}, corresponding to the adjusted, transformed, scaled and centred absorbance of wavelength *k* for the variety *i*, and β_k as the effect of the wavelength *k* on the trait.

H-BLUP: $\hat{Y}_i = \mu + W_i + \epsilon_i$ with $W \sim N(0, H\sigma_G^2)$ and $\epsilon \sim N(0, I\sigma_\epsilon^2)$

PS-LASSO: $\hat{Y}_i = \mu + \sum_{k=1}^{p} s_{ik}\,\beta_k + \epsilon_i$ where

$\hat{\beta}(\lambda) = \text{Argmin}_{\beta \in \mathbb{R}^p}\{\|Y - S\beta\|^2 + \lambda\|\beta\|\}$ with $\|\beta\| = \sum_{k=1}^{p}|\beta_k|$ and $\lambda > 0$

G-BLUP and H-BLUP models were run with the R package rrBLUP and the GS-LASSO and PS-LASSO models were run with the glmnet R package (Friedman et al. 2010)

To compare models, we computed the predictions within each environment. We considered a 5-fold cross validation scheme, which consisted in randomly splitting the data in five folds of the same size. The training set (TS) was constituted of four folds and the validation set of the remaining fold. Predictive ability (PA) was computed as the Pearson's correlation between the predicted values and the adjusted means of the validation set. In total, five PA were calculated when the five folds were used once each as the validation set. This procedure was repeated 25 times to get robust results. Comparison of models was then based on the mean and quantiles of the 125 resulting PA obtained in each environment.

Evaluation of factors influencing predictive ability of PS

Data subsets were specifically defined with the properties necessary for evaluating each factor (e.g. availability of genotypic data, number of environments with NIRS). The different sets and subsets are detailed in Table S4.

Impact of the prediction models on predictive ability

We considered two types of penalised models with assumptions corresponding to contrasted genetic architectures (G-BLUP/H-BLUP and GS-LASSO/PS-LASSO). In this analysis, the NIR spectra used in each model were collected in the same environment as the predicted traits. The four datasets were analysed in this way to gain an overview of PS in different breeding materials.

Impact of NIR spectrum origin of acquisition (environment and generation)

In practice, a genotype may be characterized by multiple spectra acquired from plant material grown in different environments. We investigated whether the environment or generation of NIR spectra acquisition had an impact on PA based on NIRS. We compared two types of predictive models involving a relationship matrix *H* derived from the NIR spectrum measured on plant material from a single environment (Single-NIRS) or from multiple NIR spectra measured on plant material from different environments and/or on the previous generation in nurseries (Multi-NIRS). For the models based on a single NIR spectrum, we compared PA obtained with a NIR spectrum measured on plant material from the same environment as the predicted trait **(S)**, or with a NIR spectrum measured on plant material from a different environment than the predicted trait **(D)** or with a NIR spectrum measured on the previous generation in nurseries **(G).** Scenarios D and G correspond to a GLOB approach as the predictions are made in environments in which no NIR spectrum was acquired.

The next objective was to assess the effect of combining NIR spectra acquired from different environments or different breeding generations to predict GY and HD. To do this, we computed *H* using a matrix with the NIR spectra added one next to the other from NIR spectrum 1 matrix (N_1) to NIR spectrum p matrix (N_p)

$$S = (N_1 \quad \dots \quad N_p)$$

S had $i \times \sum_{j=1}^{p} \lambda_j$ dimensions where λ_j designates the number of wavelengths for the NIR spectrum *j*, and *i* the number of genotypes.

We combined NIR spectra from the same and the different environments as the one in which the predicted trait was phenotyped for the TS (**CbSD**), or combined only those from different environments **(CbD).** Also, we combined the NIR spectrum recorded in the same environment as the predicted trait with the NIR spectrum recorded on the previous generation in nurseries **(CbSG).** All the NIR spectra available were combined in a final model **(CbSDG).** Another method of combining information from multiple NIR spectra is to average NIR spectra measured at multiple sites for each genotype. Here we averaged the spectra from the same environment as the predicted

trait and at least a different one for each genotype (**AvSD**). In this model each genotype is characterised by a single spectrum instead of the multiple spectra collected in the different environments. All the models are detailed in Table 3. These models correspond to GLOB approach too, as at least one NIR spectrum used was from another environment than the environment for the predicted traits.

Because of differences in datasets (number of environments, availability of NIRS collected in the previous generation) these models could not be applied to all the data. We therefore divided the four sets into two subsets to address each research question. **SubSet_SD** included genotypes that had been evaluated in at least two environments. Each genotype is described by J spectra originating from J environments, compatible with testing models S, D, CbSD, CbD, AvSD. **SubSet_SDG** includes genotypes evaluated in at least two environments plus in the previous generation in nurseries, with NIRS being compatible with evaluating models S, D, G, CbSD, CbD, CbSDG, AvSD. Details about the composition of the subsets are presented in Table S4.

Comparison of models based on data from molecular markers, NIRS, or both

To compare PS and GS PA, we compared the CbSD model to the reference model G-BLUP **(M)**. Then we investigated models that combine NIR spectra and molecular marker information (GH-BLUP). Two models were developed: **CbSD+M** and **CbSG+M** which combined the hyperspectral effect of **CbSD**/**CbSG** with the polygenic effect of **M**. All the models are detailed in Table 3. GH-BLUP models were run with R package Sommer. Two subsets were created to compare these models: **SubSet_SDM** referred to genotypes that had been evaluated in more than one environment and genotyped. **SubSet_SGM** referred to genotypes that had been evaluated in the previous generation and had been genotyped. Details of the composition of subsets are given in Table S4.

Impact of the training population size and composition on predictive ability

In GS the size and composition of the training set (TS) has an impact on PA. We characterised this effect by testing six TS sizes (10, 50, 100, 150, 200, 250 genotypes). We randomly split the data in five folds of the same size. One fold was attributed to the validation set and the remaining folds are the genotypes potentially included in the TS. Among the latter, we randomly sampled a definite number of lines to constitute the final TS with the corresponding size. The same procedure was followed for all the folds and all the TS sizes, and was repeated 25 times to give 125 predictive abilities for each TS size. We thus compared the PA of the four models **M**, **S**, **CbSD**, and **CbSD+M** representing the G-BLUP, H-BLUP and GH-BLUP model types. Model comparison was possible in two Set2-2019 datasets, GL and EM.

The composition of a TS can be optimised in order to minimise its size while retaining similar PA. In one scenario the size of the TS was arbitrarily defined, and either random or optimized procedures were used to define the TS genotypes. The validation set was composed of genotypes not present in the TS. We compared three optimisation algorithms to define the TS for a particular size. We tested the CDmean algorithm (**CDmean**) (Rincent et al. 2012), developed originally for GS, and two algorithms developed to optimize NIRS calibration equations, Honigs (**HG**) (Honigs et al. 1985) and Kennard-Stone (**KS**) (Kennard and Stone 1969). CDmean was applied with custom R code, while HG and KS were applied with the prospectr R package (Stevens and Ramirez-Lopez 2020). Algorithm performance was compared to randomly selected (RD) training populations. PA for each TS size was averaged from 50 repetitions for CDmean or 125 repetitions for RD. There was no repetition for HG and KS as they are both deterministic. To compare these TS selection approaches, we used the datasets Set2-2019-EM and Set2-2019-GL, in which many varieties were genotyped.

Table 3 Description of the GS and PS models. Y_i is the trait adjusted mean for genotype i. μ is the global mean, G_i is the random polygenic effect for genotype i, W_i is the random hyperspectral effect for the genotype i, ϵ_i is the residual effect. K is the relationship matrix derived from molecular markers and H the relationship matrix derived from NIRS. **M**, model based on molecular markers. **S** model based on NIR spectrum measured on plant material from the same environment as the predicted trait. **D**, model based on NIR spectrum measured on plant material from a different environment as the predicted trait. **G** model based on NIR spectrum measured on the previous generation. **CbSD** combined NIR spectra from the same and the different environments as the predicted trait. **CbD** combined NIR spectra from several different environments. **CbSG** combined NIR spectra measured on plant material from the same environment as the predicted trait and from the previous generation. **CbSDG** combined NIR spectra collected from plant material from all environments and the previous generation. **AvSD** averaged NIR spectra from plant material from the same environment and from one different environment for each genotype. **CbSD+M** and **CbSG+M** respectively combine models **CbSD** and **CbSG** with model **M**.

Type of model	Variable	Label	Model	Random effect	Residual effect
G-BLUP	SNP	**M**	$Y_i = \mu + G_i + \epsilon_i$	$G \sim N(0, K_G\sigma_g^2)$	$\epsilon \sim N(0, I\sigma_\epsilon^2)$
H-BLUP	Single-NIRS	**S**	$Y_i = \mu + W_i + \epsilon_i$	$W \sim N(0, H_S\sigma_w^2)$	
		D		$W \sim N(0, H_D\sigma_w^2)$	
		G		$W \sim N(0, H_G\sigma_w^2)$	
	Multi-NIRS	**CbSD**		$W \sim N(0, H_{CbSD}\sigma_w^2)$	
		CbD		$W \sim N(0, H_{CbD}\sigma_w^2)$	
		CbSG		$W \sim N(0, H_{CbSG}\sigma_w^2)$	
		CbSDG		$W \sim N(0, H_{CbSDG}\sigma_w^2)$	
		AvSD		$W \sim N(0, H_{AvSD}\sigma_w^2)$	
GH-BLUP	SNP + Multi-NIRS	**CbSD+M**	$Y_i = \mu + G_i + W_i + \epsilon_i$	$W \sim N(0, H_{CbSD}\sigma_w^2)$ $G \sim N(0, K_G\sigma_g^2)$	
		CbSG+M		$W \sim N(0, H_{CbSG}\sigma_w^2)$ $G \sim N(0, K_G\sigma_g^2)$	

Results

Descriptive statistics, heritability, and correlation between traits and absorbances

Four data sets were obtained from current bread wheat breeding programs run by public and private entities in France. The data was collected from either the first or second year of trials of candidate breeding lines. We first sought to characterize the data and NIR spectra of all datasets by evaluating the dispersion of the data, the broad-sense heritability of traits and wavelengths and the correlation between wavelengths and phenotypic traits.

Across all datasets, the general mean for GY ranged from 5.3 t/ha to 12.9 t/ha and for HD from 122 days to 145 days. We measured the dispersion of the data around the mean for each set with the coefficient of variation (CV). For GY, the CV were low to moderate ranging from 1.8% to 9.4%, and for HD the CV were low ranging from 0.1% to 2.3% (Table 1). When repetitions were available, broad-sense heritability (H^2) was calculated for both traits and for each wavelength within each environment. For GY, H^2 values were high, ranging from 0.77 to 0.95. For HD estimated only for Set3, H^2 values were high and stable from year to year ranging from 0.97 to 0.99 (Table S3). H^2 estimated for each wavelength in each environment ranged from 0.00 to 0.99. Nearly all spectra had H^2

values of between 0.88 and 0.95. Exceptions were Set4_LI_2018_EM, Set4_T_2019_LM and Set4_LI_2019_LM where the mean H^2 was poor or moderate: 0.28, 0.48, 0.66 respectively (Figure S1). Correlation estimated between adjusted wavelengths and adjusted GY)) across environments, fluctuated from -0.60 to 0.61 along spectra. Correlation estimated between adjusted wavelengths and adjusted HD across environments, fluctuated from -0.56 to 0.59 (Figure S1).

Phenomic predictions of grain yield and heading date in different bread wheat breeding programs

GY and HD do not have the same genetic architecture. GY is controlled by many genes with small effects and is highly impacted by genotype × environment (GxE) interactions. HD is mainly controlled by major genes, is very heritable (e.g. Hanocq et al. 2007; Griffiths et al. 2009) with lower levels of GxE. These differences are reflected in the H-BLUP and PS-LASSO models for predicting each trait. We first compared the PA of H-BLUP and PS-LASSO for HD and GY in different generations of two breeding schemes. The breeding sets (Set1-Set4) came from different generations of selection, with different population sizes and genetic variances (Table S3). Overall, H-BLUP and PS-LASSO were able to accurately predict both traits in each panel, but with a wide variability in PA (Figure 1). Set2 gave rise to the best PA for GY with a mean of 0.57 and a standard deviation around 0.08 for both H-BLUP and PS-LASSO. Set1 and Set4 gave rise to the lowest PA for GY with both models with PA between 0.34 and 0.39. The standard deviation of PA for Set4 GY was as high as 0.22. Conversely, Set1 gave rise to the best PA for HD with the PS-LASSO model, with an average of 0.62 and a standard deviation of 0.10. Set4 gave rise to the lowest average PA for HD with H-BLUP (0.26).

PS-LASSO showed the better PA for HD in the three sets, on average 13.4% better than H-BLUP (Figure 1). Moreover, the interquartile ranges of PS-LASSO PA were narrower than with H-BLUP. For GY, H-BLUP and PS-LASSO had similar PA with similar variances for each panel of wheat genotypes. We calculated the frequency of selection by PS-LASSO of each wavelength in all the environments to visualize the distribution of influential wavelength (Figure S2). For both traits, influential wavelengths were distributed across the entire visible and infrared spectra. Wavelengths of 2108 and 2496 nm were very often selected regardless of the trait or the environment considered. More specifically, for GY at 1202 nm and for HD at 400 nm, these wavelengths were very often selected regardless of the environment.

Only the H-BLUP model was used for the following analyses, because it is computationally more efficient than PS-LASSO.

Impact of NIR spectrum origin of acquisition on predictive ability

We tested different models relying on a single spectrum or multiple spectra collected on the same genetic material grown in different environments to calculate the similarity between genotypes (Figure 2). The environment or generation of the plant material which produced the grain or flour that was analysed by NIRS had a consequence on the PA of the H-BLUP model. For models based on a single spectrum as predictive variable, the NIRS may have been measured on plant material grown on the same site as the site for which the trait was predicted (S), on a different site (D), or from an earlier generation (G). As expected, in the respective models S was on average better than D and G for predicting GY in each panel of the two subgroups SubSet_SD and SubSet_SDG (Figures 2A, 2C). For HD, S and D models from SubSet_SD and SubSet_SDG gave rise to similar average PA (Figure 2 b). In SubSet_SDG, prediction by the S model outperformed D and G models for Set1 data, but was outperformed by G for Set4 data, with little difference for Set3 (Figure 2 d).

We also compared different models combining multiple NIR spectra. For SubSet_SD genotypes, predictions based on CbSD were as good as those based on the S model for GY (Figure 2a) and were even better for HD (Figure 2b). For SubSet_SDG genotypes, prediction based on this same model was variable but was mostly as good as the best single NIRS model for both traits (Figures 2c, 2d). In direct comparison, using the AvSD model resulted in lower PA than CbSD in 78% of cases and in particular for HD, on average 38% less of the CbSD PA. The multi-NIRS models resulted in dissimilar and variable PA. CbSDG produced the best predictions in almost all panels and was at least as good as CbSD for both traits (Figures 2c, 2d). For Set4, the CbD model was compared to that CbSD. PA were higher with CbSD than with CbD for HD. Surprisingly, using CbD resulted in similar PA for GY as with CbSD. This was unexpected because, for example, the PA reduction between S and D was almost 22% for GY (Figure 2a), but the loss of PA between CbD and CbSD was only 11%. Using CbSD resulted in better predictions for both traits than using the D model.

As the PA were averaged by set, to ensure that the tendencies found for S, D and CbSD held true for all the specific panel-environment (P-E) combinations, we compared the deviation between the PA of S and D models. For GY and HD, around 10% and 50% of the combinations respectively gave better predictions using NIR spectra originating from a trial different than the one in which the predicted trait was measured (Figure S3A). These P-E combinations were selected to compare the PA of the S, D and CbSD models (Figure S3B). In these particular P-E combinations, the PA of the S model was on average 43% less for GY and 69% less for HD than the PA of the D model. The multi-NIRS model CbSD resulted in PA as high as the best single NIRS models. For HD, the same tendency was observed, but for some P-E combinations CbSD resulted in much higher PA than using D or S models.

Influence of the number of NIRS combined in multiple NIRS models on predictive ability

Combining NIR spectra from genotypes grown in multiple situations enhanced the PA of PS models. We investigated whether increasing the number of NIR spectra in the prediction model affected the PA (Figure 3). To dissociate the effect of including the NIR spectrum collected in the same environment as the predicted trait from the effect of combining different NIR spectra, we compared first the effect of adding more NIR spectra to the NIR spectrum collected in the same environment as the predicted trait (Figure 3a), and second the effect of adding more spectra from different environments without the NIR spectrum collected in the same environment as the predicted trait (Figure 3b). In the first scenario, size one corresponds to model S and other sizes to CbSD and in second scenario, size one corresponds to model D and other sizes to CbD. For both scenario and traits, on average, adding more NIR spectra in the multi-NIRS model enhanced PA, but the marginal gain in PA decreased as more NIRS were introduced. For GY, on average for both scenarios, successive addition of one more NIR spectra to the original NIR spectrum in the model increased PA by 12.0%, 3.7%, 1.4% and 2.0%, respectively. The same tendency was observed with HD, but with higher gains of 57.0%, 22.0%, 15.0%, and 2.5%, respectively. These general trends were also observed on the other sets (Figure S4).

Comparison of GS and PS predictive abilities

We performed predictions with the H-BLUP multi-NIRS models (CbSD, CbSG) and the GS G-BLUP model based on molecular markers (M). We also tested a GH-BLUP model combining information from both markers and NIRS (CbSD+M, CbSG+M). Based on SubSet_SDM, PA for GY were on average better for CbSD than M (Figure 4). Considering Set2 with the larger training population, CbSD outperformed the M model. With Set3 and Set1, CbSD performed slightly better than M and with Set4, CbSD was outperformed by M. For HD, CbSD outperformed M in all the panels. For SubSet_SGM, PA for GY was higher for CbSD than for M with Set1, but worse with Set3 and Set4. For HD, CbSG performed similarly to M or much better with Set1. On average, considering all the subsets, in 48% of cases PS outperformed GS in predicting GY and in 62% of cases PS outperformed GS in predicting HD. Interestingly, GH-BLUP models combining both marker and NIRS data were the best models for both traits in all cases except one, and sometimes resulted in marked improvements, such as with SubSet_SDM Set4 that was 30% better than the M and CbSD models for HD prediction.

Impact of training set size and composition on predictive ability

The training set size usually has a noticeable effect on PA in GS. We analysed the effect of increasing TS size on the PA of PS and GS in two environments of the Set2 panel where the number of genotypes was substantial (Figure 5). For each model and both environments, PA increased with the TS size. For the smallest TS, the S model was slightly better than the others. For Set2_2019_EM, going from 50 to 250 breeding lines in the TS, the slight rise in PA of S and CbSD followed a similar trend, almost reaching a plateau. Regardless of the TS size, the GS model M was less accurate than the PS models. Compared to the CbSD model, M followed a similar tendency in prediction gain with the increasing TS size for Set2_2019_GL, equalling the PA of S in Set2_2019_EM with a TS of 250. Model CbSD+M, combining both marker and NIRS data, outperformed the other models with as few as 50 genotypes in Set2_2019_EM and with at least 150 genotypes from Set2_2019_GL.

To optimise TS composition for a given size, we compared three different optimisation algorithms to determine the composition of the TS for performing GS or PS (Figure 6). When applying GS (G-BLUP model), TS optimized with CDmean computed with the kinship matrix (CDmean_K) performed better than random TS (RD) for both environments. When applying PS (both H-BLUP models), TS optimized with CDmean_H computed on the NIR similarity matrix also performed better than randomly sampled TS (RD) in Set2_2019_GL. Finally, for the GH-BLUP model combining molecular marker and NIRS, CDmean_K performed even better than CDmean_H

and RD. KS and HG performed very variably as function of the TS size in Set2_2019_GL and with lower PA than RD in Set2_2019_EM when applying H-BLUP and GH-BLUP models.

Discussion

Phenomic selection is an accurate complement or alternative to genomic selection in elite breeding material

The results presented here showed that phenomic selection was able to predict two traits with dissimilar genetic architectures for different sets of elite bread wheat breeding lines from different breeding companies. PS may therefore be useful for direct applications in breeding.

To evaluate the performance of this method, we compared PS predictions to those obtained with a reference GS model. We developed two multi-NIRS models to make GLOB prediction for wheat breeding material. Predictions were at least as accurate as the GS model in 58% of the cases for GY and in 65% of cases for HD (Figure 4). These promising results were consistent with previous results using PS to predict GY and HD in a panel of registered bread wheat varieties and in a panel of poplars (Rincent et al. 2018). The few studies to have implemented PS with H-BLUP were mainly applied to cereal species (Montesinos-López et al. 2017; Rincent et al. 2018; Krause et al. 2019; Cuevas et al. 2019; Galán et al. 2020; Lane et al. 2020), facilitated by the routine NIRS use in these crops to predict grain quality traits such as protein content (Osborne 2006). In most of the studies just cited, the traits were phenotyped and NIRS data was acquired on the same plots. PS was mostly as good and sometimes better than GS (Rincent et al. 2018; Krause et al. 2019; Galán et al. 2020). Even if GS sometimes outperforms PS (e.g. Cuevas et al. 2019), PS can still be an attractive method due to the low cost of NIRS acquisition (Rincent et al. 2018).

We assessed a model that used both types of similarity matrix, that is, with genomic and hyperspectral data. Interestingly, this model systematically outperformed each of the basic models (G-BLUP and H-BLUP), meaning that molecular markers and NIR spectra bring complementary information (Figure 4). These results are consistent with studies which combined molecular marker scores and NIRS with the addition of pedigree information (Krause et al. 2019) or plant height phenotype (Galán et al. 2020). The same was not true for a model combining NIRS, molecular marker and pedigree information, as PS and GS PA were similar in Cuevas et al. (2019). In GLOB studies on maize (Riedelsheimer et al. 2012; Guo et al. 2016; Westhues et al. 2017; Schrag et al. 2018; Azodi et al. 2020), wheat (Ward et al. 2015), rice (Xu et al. 2016) or *Arabidopsis thaliana* (Gärtner et al. 2009), combining information on transcripts and/or metabolites with DNA markers, led to more accurate predictions than classic GS models. However, Riedelsheimer et al. (2012) and Xu et al. (2016), found that combining multi-omics data did not further improve genomic prediction in comparison to GLOB prediction with metabolomics data alone. Possibly GLOB prediction blending genomic and metabolomic information is better than GS when fewer DNA markers are available. However, even with numerous markers, combining metabolic data may or may not improve GY prediction as found, respectively, by Schrag et al. (2018) with around 37,000 DNA markers and 284 metabolites and by Riedelsheimer et al. (2012) with around 56,000 DNA markers and 130 metabolites.

Our results are consistent with the idea that, if multiple types of information are available to characterise a genotype, it would be preferable to include them all in the predictive model. The reasoning is that omics data other than genomics reflect the expression of genes or biological pathways, which is different information than the genotype. Being able to account for non-additive effects like epistasis or dominance may lead to improvement in complex trait prediction.

Factors influencing PS predictive ability

Our study explored different factors thought to impact the PA of PS, whether factors that have been identified in GS such as the type of predictive model or the size and composition of the training population, or factors specific to PS such as how many spectra and which spectra (from which individual plant samples) are used to compute similarity matrix.

H-BLUP was well suited to predicting both traits with a reasonable computational demand

We compared the H-BLUP model, which considers absorbances at all the wavelengths of a spectrum to compute a relationship matrix between individuals, and the PS-LASSO model, which selects wavelengths that contribute the most to the trait variance. We observed that the prediction for GY was similar for both models (Figure 1).

However, for HD, PS-LASSO was slightly more accurate than H-BLUP. These results suggest that specific wavelengths could be linked to the HD trait. We compared the wavelengths selected by the LASSO model for each environment in parallel with the PA of the model for each environment (Figure S2). As the NIR bands are signatures of specific chemical bonds, investigating the organic molecules giving rise to influential bands would help us understand which secondary traits are correlated to the trait to be predicted. NIR readouts can give an overview of the organic molecules identified (Xiaobo et al. 2010). The peaks identified in the present study are characteristic of several common atomic bonds (e.g. H_20, CH, CH_2, $CONH_2$) making interpretation difficult.

NIR spectra are composed of multiple quantitative variables, which capture some genetic variability (Posada et al. 2009; Rincent et al. 2018). Thus, models developed for GS using molecular markers as explicative variables can be transposed to using NIRS variables. In this study we found that H-BLUP was efficient at predicting both traits with a low computational demand. We have evaluated only two models here among the numerous ones available (Charmet et al. 2020). However, H-BLUP performs well for different traits with an efficient computation time and it is user-friendly. More specifically for PS, other types of models, such as functional regression, can be considered to improve PA. NIR spectra can be modelled as a curve in a continuous domain and so considered as a function (Montesinos-López et al. 2017a,b; Lane et al. 2020). In this case the wavelengths are not used to compute a kernel but as covariates in the functional model. Lane et al. (2020) showed that, in most scenarios, prediction based on functional regressions outperformed H-BLUP. Functional regression reduces dimensionality by transforming wavelengths into functional covariates. This kind of model is thus able to handle high-dimensional data, but is more difficult to implement than H-BLUP due to parametrisation choices needed to approximate the functions. Montesinos-López et al. (2017a) found that on average across all environments, a developed B-Spline and Fourrier functional regression on wavelengths outperformed the other models tested (namely Ordinary Least Square regression on vegetative index and Bayes B and Principal Component Bayes B regression on all wavelengths). Reducing the dimensionality of regressors is a major issue when implementing a predictive model with numerous regressors. Runcie et al. (2020) recently developed a mega-scale linear mixed model dealing with thousands of traits, which is interesting because their multivariate analysis leveraged information across thousands of secondary traits. They used the data from Krause et al. (2019) to compare the PA of G-BLUP, H-BLUP, and combined both (GH-BLUP) with their MegaLMM. Inspired by a multivariate model, wavelengths are used to assist the prediction of the targeted trait. They found that their method dramatically outperformed the GH-BLUP model with an improvement from 0.64 to 0.77 in PA. However, this method always requires the use of molecular markers with the secondary traits to estimate latent traits. By contrast, the PS models developed in our study are able to predict complex traits without any genotyping.

Multi-NIRS models ensured accurate predictions regardless of the trait and the environment

Combining multiple NIRS from individual plant samples from different environments or generations resulted in more robust PA in the models developed here. Unlike molecular markers, NIRS are not independent of the environment. As for GY or HD, spectra are phenotypes that are the result of the genotype, the environment and the GxE interaction. The calculation of spectra heritability within environments revealed that the heritability varies from one environment to another and from one wavelength to another (Figure S1). Spectra with a medium or low average heritability also tended to be worse for predicting GY and HD. The quality of spectra measurement should be considered too. When comparing single NIR spectra, our first assumption was that the NIR spectrum from the same environment as the predicted trait would be the best predictor (S model). We compared this approach to GLOB predictions in which the spectra are acquired from plants in another environment (D model) or generation (G) than the one in which the TS was phenotyped for the predicted trait. Our results showed that on average, this was indeed the case when using a single spectrum to compute the similarity matrix (Figure 2). However, D and G models sometimes lead to similar or higher PA than model S, especially for the prediction of HD. This is a significant result showing that according to the trait, the source of the spectra, that is the environment in which the specific plant material is assayed, can have a major impact on prediction. We know that NIR spectra are able to capture a significant portion of additive genetic variance (Rincent et al. 2018). Furthermore, we assume that the NIRS are able to indirectly capture some secondary traits correlated to the predicted trait. For example, GY is negatively correlated to the protein content and the NIR spectra is an excellent predictor of protein content, so this correlation would indirectly predict GY in some situations. Both factors can explain why PS is able to predict complex traits. GY is a difficult trait to select for, because of the impact of the GxE interaction on the trait variance. For this trait, genetic and GxE interaction information are both required to reach good prediction. We suppose that spectra contain the GxE variance information by capturing the GxE effect of the phenotype via the metabolites and other molecules in response to the different stresses and growing conditions. Thus, for GY, only the spectra from the same environment as the phenotype is expected to depict the responses of the plant through the gene expression and regulation that occurred in this particular environment. Unlike GY, HD of winter bread wheat has

a very low level of GxE when the varieties are sown in autumn. We suppose that spectra collected in any environment are useful to predict this trait because they all capture genetic variance (as for GY) and because a relationship between a spectrum and HD obtained in one environment will be useful in any other environment.

We explored two ways of dealing with spectral data, focusing on the second: identify the best spectra *a priori* to calculate the similarity matrix or consider all the available spectra together. We proposed several models combining spectra from different environments or generations. We also compared two ways of computing the similarity matrix. The multi-NIRS models (CbSD, CbD, CbSG, CbSDG) are equivalent to the one used by Lane et al. (2020) where all the similarity matrices specific to each environment are averaged when all the spectra had the same number of wavelengths. The multi-NIRS model AvSD differs as the spectra were averaged by genotype followed by computation of the similarity matrix. We showed that combining spectra gave better predictions than averaging spectra. The difference is mainly due to the loss of information characterising genotypes in different environments when averaging the spectra by genotype.

Our multi-NIR spectra models CbSD and CbSDG outperformed the best single NIRS model for GY in 47% of cases and the best single NIRS model for HD in 68% of cases (Figure 2). This suggests that adding other spectra to the predictive model, from another environment or generation, allowed a better estimate of the similarity between genotypes than a single spectrum. More investigation is still required to incorporate spectra from different tissues or at different time points of the growing cycle to bring different useful information to the predictive model. Similarly to molecular markers in GS (Norman et al. 2018), adding more and more predictors eventually leads to a plateau in PA, probably as the information added becomes redundant (Figure 3). We noticed that depending on the predicted trait, the plateau was reached with fewer predictors. For GY, the gain was marginal when more than two spectra were added, whereas for HD, the plateau was still not reached with four spectra. For HD, we suppose that adding more spectra into the model allows more genetic variance to be captured reflecting the more indirect relationships between the underlying tissue composition and HD. For GY, we suppose that the GxE which explains much of the trait variance is environment specific and can be captured only with spectra from the same environment as the phenotype. Thus, adding more spectra from different environments will not bring more useful information in the predictive model. More data are however required to fully understand the benefit of combining NIR spectra for different traits.

Size of the training set had a dramatic impact on predictive abilities and CDmean optimised TS for both PS and GS

Finally, with a view to optimising the use of PS in breeding programs, we explored the effect of the training population size and composition on the PA of PS and GS for two environments with the most genotypes (Figure 5). Our results showed that the TS size had a crucial impact on prediction, such that the PA with 200 genotypes in the TS was double that with 10 genotypes in the TS. Our results also revealed that like GS (Norman et al. 2018), the gain in accuracy reached a plateau for PS. There is a point where adding other individuals to train the model does not bring novel information. This suggests that optimisation of the TS could generate the best possible prediction while minimising TS size. Among all the models, for TS as small as 10 or 50 genotypes, the single-NIRS model gave better predictions than GS models or the model combining genomic and NIRS information. One interesting observation is that the model using genomics and NIRS data gave the best prediction when we considered larger TS and the plateau was reached with larger TS than in the other models. This may be because useful genomic information and phenomic information are not borne by the same genotypes and thus increasing the TP size still increases the PA.

We then compared different algorithms to select the individuals which would constitute the TS of imposed sizes. We compared the CDmean algorithm (Rincent et al. 2012) for GS with two optimisation algorithms usually used in chemometrics to select the TS, the H and KS algorithms (Honigs et al. 1985; Kennard and Stone 1969). In the optimisation scenario, the algorithm selects the genotypes to be phenotyped to train the model. Our results revealed that CDmean gave better PA than random selection, with both molecular marker and NIRS data (Figure 6). TS determined by the H and KS algorithms gave rise to variable PA and were usually less accurate than random selection and CDmean. Because the CDmean maximised the expected PA, this criterion is better suited to predictive models than the H and KS deterministic algorithms which select the individuals to cover the absorbance variability across the spectrum. Finally, we compared the CDmean optimisation based on information from molecular marker or NIR spectra in the prediction model using both genomic data and NIR spectra. On average CDmean based on molecular markers gave better predictions than CDmean based on NIR spectra. It would be interesting to find a criterion which could handle molecular markers and NIRS at the same time to select the individuals in

the TS. We tested the optimisation based on spectra from the same generation as the phenotyping of the target trait. Another possibility would be to use the NIR spectra from the N-1 generation to define the TS and to predict the target trait to be phenotyped in the N generation. For example, it would be useful for choosing which genotypes planted in a trial should be finely phenotyped.

Other factors such as the scanned organ and spectral acquisition method should be considered

In our study we could not compare the effect of the tissue used for NIRS collection or the effect of the spectrometer and protocol used to collect NIRS. However, it is interesting that other studies used hyperspectral imaging of the canopy (UAV) and obtained PS predictions even better than GS (Krause et al. 2019; Galán et al. 2020). In this case the number of wavelengths compared to lab measurements is halved, with wavelengths from the visible part of spectrum only. The nature of the plant material in canopies is quite different from grain samples, particularly as the canopy is still growing in the environment or the trial. The tissue sample for NIRS acquisition can have a substantial impact on PA, for example, using leaf spectra was less accurate than using grain spectra for the prediction of GY in Rincent et al. (2018). Furthermore in PS, or in any application of NIRS, spectra need to be pre-treated with specific filters to remove irrelevant information due to additive and multiplicative external effects (Blanco and Villarroya 2002). Mathematical transformations are applied to spectra correcting the absorbance value of each wavelength. Such transformations impact the PA of the method used downstream and cannot be generalised to other data (Cuevas et al. 2019). In practice, different filters should be tested in a cross-validation to select the best set of pre-treatments.

Conclusion

We explored whether PS can be implemented in wheat breeding programs by identifying factors which impact the PA of the method. Our main results showed that the choice of the spectra used in the predictive model is crucial and is specific to the predicted trait. The H-BLUP models developed here which combined multiple spectra to compute the similarity matrix are meaningful and can be readily handled by breeders. As in GS, the size of the training set had a crucial impact on PA and optimisation of the TS with CDmean criterion giving better PA than random selection. Finally, amid all these factors, PS was mostly at least as accurate as GS for both traits. PS is a good alternative to GS, in particular when NIRS are already routinely measured (e.g. in all wheat breeding programs). For these species, PS could be implemented without any additional cost. In breeding programs in which genotypic data are already routinely collected, PS combining NIRS and molecular marker data could potentially be advantageous and more accurate as we found. Integrating different kinds of high-throughput phenotyping and multi-omics information in statistical models is a promising approach.

Acknowledgements

The authors thank the work in experimental units by INRAE (Clermont-Ferrand, Estrées-Mons, Le Moulon, Rennes), breeders from Agri-Obtentions and Florimond Desprez. The authors are grateful to Agri-Obtentions, Florimond-Desprez, and the Association Nationale de la Recherche et de la Technologie (ANRT, grant number 2019/0060) which supported this PhD word. The authors also thank Bastian Alexandre and Rachel Carol (Bioscience Editing, France) for the proofreading of this work.

References

Albrecht T, Wimmer V, Auinger H-J, et al (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339–350. https://doi.org/10.1007/s00122-011-1587-7

Allard RW, Bradshaw AD (1964) Implications of Genotype-Environmental Interactions in Applied Plant Breeding1. Crop Sci 4:cropsci1964.0011183X000400050021x. https://doi.org/10.2135/cropsci1964.0011183X000400050021x

Azodi CB, Pardo J, VanBuren R, et al (2020) Transcriptome-Based Prediction of Complex Traits in Maize. Plant Cell 32:139–151. https://doi.org/10.1105/tpc.19.00332

Ben-Sadoun S, Rincent R, Auzanneau J, et al (2020) Economical optimization of a breeding scheme by selective phenotyping of the calibration set in a multi-trait context: application to bread making quality. Theor Appl Genet 133:2197–2212. https://doi.org/10.1007/s00122-020-03590-4

Blanco M, Villarroya I (2002) NIR spectroscopy: a rapid-response analytical tool. TrAC Trends Anal Chem 21:240–250. https://doi.org/10.1016/S0165-9936(02)00404-1

Charmet G, Tran L-G, Auzanneau J, et al (2020) BWGS: A R package for genomic selection and its application to a wheat breeding programme. PLOS ONE 15:e0222733. https://doi.org/10.1371/journal.pone.0222733

Covarrubias-Pazaran G (2016) Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLOS ONE 11:e0156744. https://doi.org/10.1371/journal.pone.0156744

Crossa J, Campos G de los, Pérez P, et al (2010) Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186:713–724. https://doi.org/10.1534/genetics.110.118521

Crossa J, Pérez-Rodríguez P, Cuevas J, et al (2017) Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci 22:961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Cuevas J, Montesinos-López O, Juliana P, et al (2019) Deep Kernel for Genomic and Near Infrared Predictions in Multi-environment Breeding Trials. G3amp58 GenesGenomesGenetics 9:2913–2924. https://doi.org/10.1534/g3.119.400493

Daetwyler HD, Calus MPL, Pong-Wong R, et al (2013) Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193:347–365. https://doi.org/10.1534/genetics.112.147983

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185:1021–1031. https://doi.org/10.1534/genetics.110.116855

de los Campos G, Hickey JM, Pong-Wong R, et al (2013) Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193:327–345. https://doi.org/10.1534/genetics.112.143313

Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome 4:250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Endelman JB, Jannink J-L (2012) Shrinkage Estimation of the Realized Relationship Matrix. G3 GenesGenomesGenetics 2:1405–1413. https://doi.org/10.1534/g3.112.004259

Fernandez O, Urrutia M, Bernillon S, et al (2016) Fortune telling: metabolic markers of plant performance. Metabolomics 12:158. https://doi.org/10.1007/s11306-016-1099-1

Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43:1. https://doi.org/10.1186/1297-9686-43-1

Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33:. https://doi.org/10.18637/jss.v033.i01

Frisch M, Thiemann A, Fu J, et al (2010) Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. Theor Appl Genet 120:441–450. https://doi.org/10.1007/s00122-009-1204-1

Fu J, Falke KC, Thiemann A, et al (2012) Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. Theor Appl Genet 124:825–833. https://doi.org/10.1007/s00122-011-1747-9

Galán RJ, Bernal-Vasquez A-M, Jebsen C, et al (2020) Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye. Theor Appl Genet 133:3001–3015. https://doi.org/10.1007/s00122-020-03651-8

Gärtner T, Steinfath M, Andorf S, et al (2009) Improved Heterosis Prediction by Combining Information on DNA- and Metabolic Markers. PLoS ONE 4:e5220. https://doi.org/10.1371/journal.pone.0005220

Gorjanc G, Jenko J, Hearne SJ, Hickey JM (2016) Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17:30. https://doi.org/10.1186/s12864-015-2345-z

Griffiths S, Simmonds J, Leverington M, et al (2009) Meta-QTL analysis of the genetic control of ear emergence in elite European winter wheat germplasm. Theor Appl Genet 119:383–395. https://doi.org/10.1007/s00122-009-1046-x

Guo Z, Magwire MM, Basten CJ, et al (2016) Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theor Appl Genet 129:2413–2427. https://doi.org/10.1007/s00122-016-2780-5

Hanocq E, Laperche A, Jaminon O, et al (2007) Most significant genome regions involved in the control of earliness traits in bread wheat, as revealed by QTL meta-analysis. Theor Appl Genet 114:569–584. https://doi.org/10.1007/s00122-006-0459-z

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009a) Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92:433–443. https://doi.org/10.3168/jds.2008-1646

Hayes BJ, Visscher PM, Goddard ME (2009b) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91:47–60. https://doi.org/10.1017/S0016672308009981

Heffner EL, Jannink J-L, Iwata H, et al (2011) Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. Crop Sci 51:2597–2606. https://doi.org/10.2135/cropsci2011.05.0253

Hickey JM, Dreisigacker S, Crossa J, et al (2014) Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. Crop Sci 54:1476–1488. https://doi.org/10.2135/cropsci2013.03.0195

Honigs DE, Hieftje GM, Mark HL, Hirschfeld TB (1985) Unique-sample selection via near-infrared spectral subtraction. Anal Chem 57:2299–2303. https://doi.org/10.1021/ac00289a029

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–177. https://doi.org/10.1093/bfgp/elq001

Kang HM, Sul JH, Service SK, et al (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–354. https://doi.org/10.1038/ng.548

Kennard RW, Stone LA (1969) Computer Aided Design of Experiments. Technometrics 11:137. https://doi.org/10.2307/1266770

Krause MR, González-Pérez L, Crossa J, et al (2019) Hyperspectral Reflectance-Derived Relationship Matrices for Genomic Prediction of Grain Yield in Wheat. G3amp58 GenesGenomesGenetics g3.200856.2018. https://doi.org/10.1534/g3.118.200856

Lane HM, Murray SC, Montesinos-López OA, et al (2020) Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels. Plant Phenome J 3:. https://doi.org/10.1002/ppj2.20002

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Montesinos-López A, Montesinos-López OA, Cuevas J, et al (2017a) Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper-spectral image data. Plant Methods 13:. https://doi.org/10.1186/s13007-017-0212-4

Montesinos-López OA, Montesinos-López A, Crossa J, et al (2017b) Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. Plant Methods 13:. https://doi.org/10.1186/s13007-016-0154-2

Norman A, Taylor J, Edwards J, Kuchel H (2018) Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. G3amp58 GenesGenomesGenetics 8:2889–2899. https://doi.org/10.1534/g3.118.200311

Osborne BG (2006) Applications of near Infrared Spectroscopy in Quality Screening of Early-Generation Material in Cereal Breeding Programmes. J Infrared Spectrosc 14:93–101. https://doi.org/10.1255/jnirs.595

Posada H, Ferrand M, Davrieux F, et al (2009) Stability across environments of the coffee variety near infrared spectral signature. Heredity 102:113–119. https://doi.org/10.1038/hdy.2008.88

Pszczola M, Strabel T, Mulder HA, Calus MPL (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci 95:389–400. https://doi.org/10.3168/jds.2011-4338

R2D2 Consortium, Fugeray-Scarbel A, Bastien C, et al (2021) Why and How to Switch to Genomic Selection: Lessons From Plant and Animal Breeding Experience. Front Genet 0: https://doi.org/10.3389/fgene.2021.629737

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, et al (2012b) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44:217–220. https://doi.org/10.1038/ng.1033

Rimbert H, Darrier B, Navarro J, et al (2018) High throughput SNP discovery and genotyping in hexaploid wheat. PLOS ONE 13:e0186329. https://doi.org/10.1371/journal.pone.0186329

Rincent R, Charpentier J-P, Faivre-Rampant P, et al (2018) Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar. G3amp58 GenesGenomesGenetics g3.200760.2018. https://doi.org/10.1534/g3.118.200760

Rincent R, Laloe D, Nicolas S, et al (2012) Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (Zea mays L.). Genetics 192:715–728. https://doi.org/10.1534/genetics.112.141473

Rodríguez-Álvarez MX, Boer MP, van Eeuwijk FA, Eilers PHC (2018) Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spat Stat 23:52–71. https://doi.org/10.1016/j.spasta.2017.10.003

Runcie DE, Qu J, Cheng H, Crawford L (2021) MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. bioRxiv 2020.05.26.116814. https://doi.org/10.1101/2020.05.26.116814

Savitzky Abraham, Golay MJE (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal Chem 36:1627–1639. https://doi.org/10.1021/ac60214a047

Schrag TA, Westhues M, Schipprack W, et al (2018) Beyond Genomic Prediction: Combining Different Types of *omics* Data Can Improve Prediction of Hybrid Performance in Maize. Genetics 208:1373–1385. https://doi.org/10.1534/genetics.117.300374

Seifert F, Thiemann A, Schrag TA, et al (2018) Small RNA-based prediction of hybrid performance in maize. BMC Genomics 19:371. https://doi.org/10.1186/s12864-018-4708-8

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447–2454. https://doi.org/10.2527/jas.2007-0010

Ward J, Rakszegi M, Bedő Z, et al (2015) Differentially penalized regression to predict agronomic traits from metabolites and markers in wheat. BMC Genet 16:19. https://doi.org/10.1186/s12863-015-0169-0

Westhues M, Schrag TA, Heuer C, et al (2017) Omics-based hybrid prediction in maize. Theor Appl Genet 130:1927–1939. https://doi.org/10.1007/s00122-017-2934-0

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252. https://doi.org/10.1017/S0016672399004462

Xiaobo Z, Jiewen Z, Povey MJW, et al (2010) Variables selection methods in near-infrared spectroscopy. Anal Chim Acta 667:14–32. https://doi.org/10.1016/j.aca.2010.03.048

Xu S, Xu Y, Gong L, Zhang Q (2016) Metabolomic prediction of yield in hybrid rice. Plant J 88:219–227. https://doi.org/10.1111/tpj.13242

Yu X, Li X, Guo T, et al (2016) Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat Plants 2:. https://doi.org/10.1038/nplants.2016.150

Zenke-Philippi C, Frisch M, Thiemann A, et al (2017) Transcriptome-based prediction of hybrid performance with unbalanced data from a maize breeding programme. Plant Breed 136:331–337. https://doi.org/10.1111/pbr.12482

Zhong S, Dekkers JCM, Fernando RL, Jannink J-L (2009) Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics 182:355–364. https://doi.org/10.1534/genetics.108.098277

Figure captions

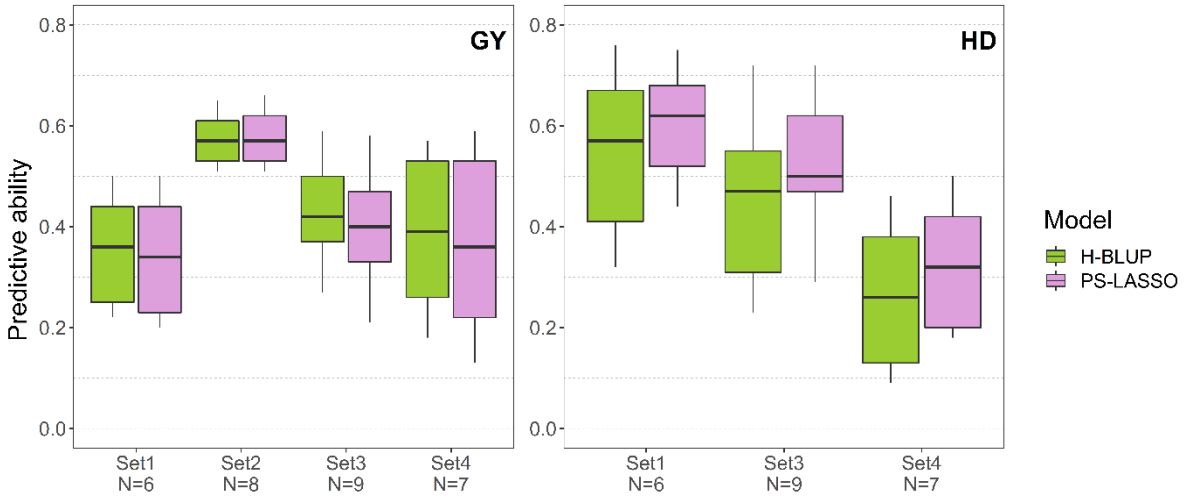

Fig. 1 Boxplot of predictive abilities of phenomic selection for grain yield (GY) and heading date (HD) based on four different wheat breeding datasets (cf. Table1) with two different models (H-BLUP, PS-LASSO). Predictions were based on a 5-fold cross-validation with 25 repetitions. Each boxplot indicates the mean (bold horizontal line), the 1st to 3rd quartiles (box), the 1st to 9th deciles (whiskers). N is the number of environments of prediction used to calculate the boxplot.

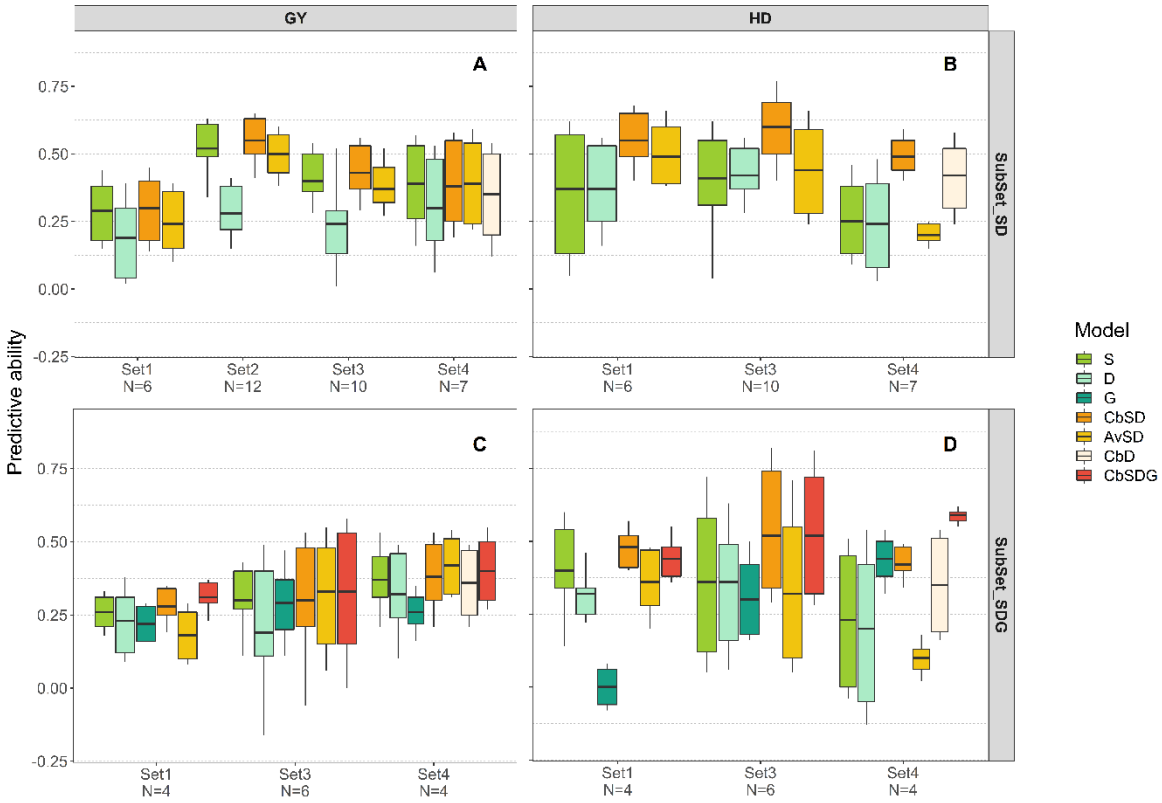

Fig.2 Boxplot of the predictive abilities of phenomic selection for grain yield (GY: **a,c**) and heading date (HD: **b,d**) in subgroup panels of wheat genotypes SubSet_SD (**a,b**) and SubSet_SDG (**c,d**) (cf. Table1). Single-NIRS models S, D, and G and multi-NIRS models CbSD, AvSD, CbD, and CbSDG were compared (cf. Table 3). Predictive abilities were obtained from a 5-fold cross-validation with 25 repetitions. Each boxplot indicates the mean (bold horizontal line), the 1st to 3rd quartile (box) and the 1st to 9th deciles (whiskers). N was the number of environments of prediction used to calculate the boxplot.

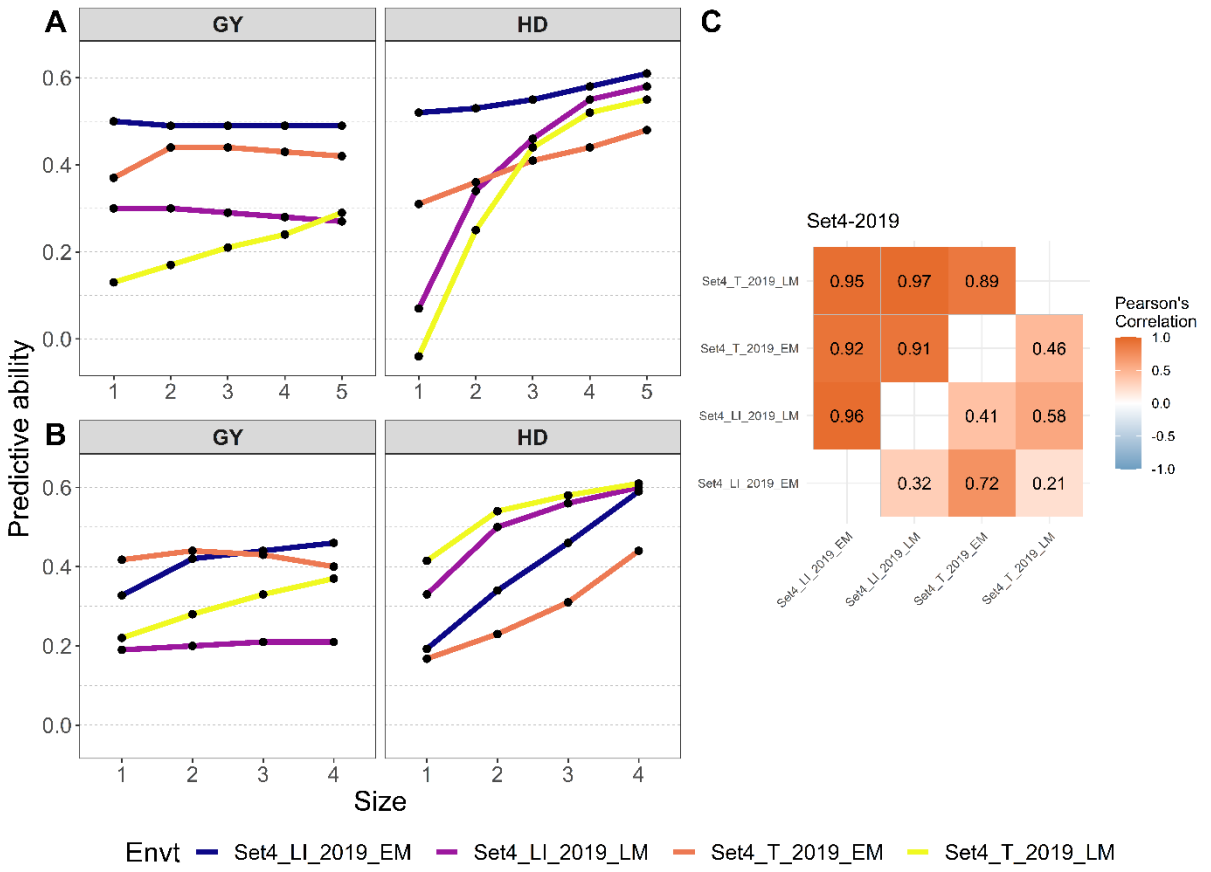

Fig. 3 Predictive ability plotted as a function of the number of combined NIR spectra for grain yield (GY) and heading date (HD) for the panel-environment combinations of Set4 genotypes from year 2019 (**a,b**). Size corresponds to the number of NIR spectra combined in making the prediction. In plot **a**, size 1 corresponds to model S and sizes 2-5 correspond to CbSD. In plot **b**, size 1 corresponds to model D and sizes 2-4 correspond to CbD. Each dot corresponds to the mean of all the possible NIRS combinations of the same size. Predictive abilities were obtained from a 5-fold cross-validation with 25 repetitions. Matrices of correlation between environments for Set4-2019 (**c**), showing the Pearson correlation for heading date (upper matrix) and grain yield (lower matrix).

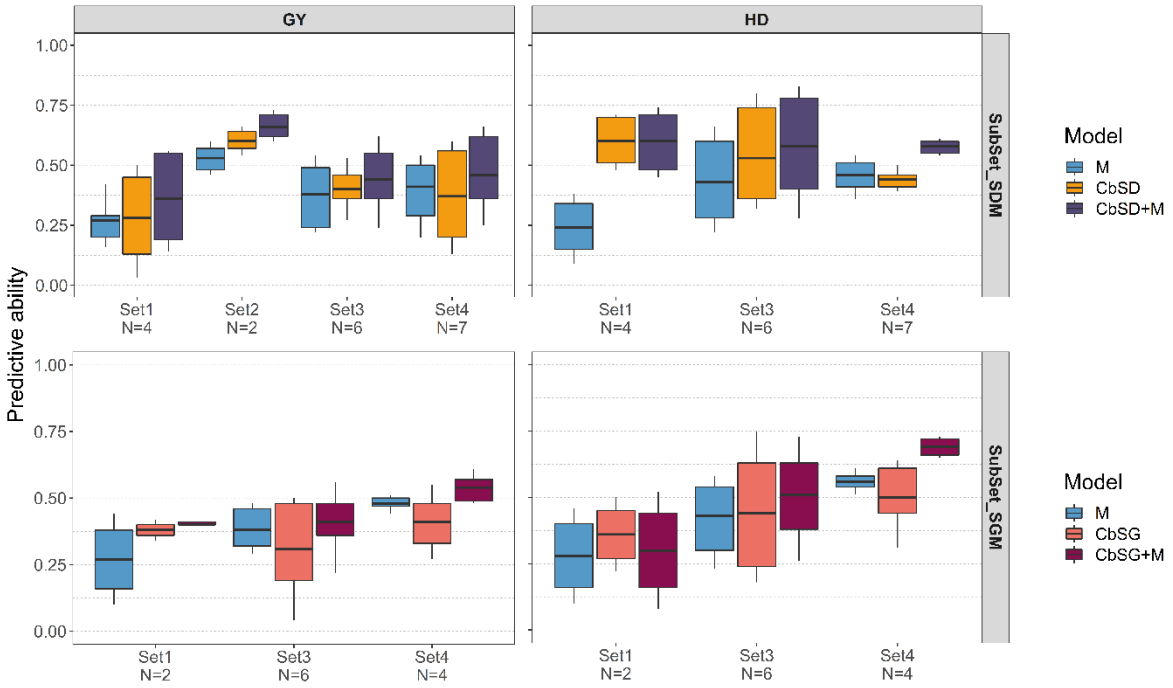

Fig. 4 Boxplots of the predictive abilities of GS and PS for grain yield (GY) and heading date (HD) using different wheat breeding datasets (SubSet_SDM and SubSet_SGM) (cf. Table1). Phenomic selection H-BLUP models CbSD and CbSG were compared to the G-BLUP genomic selection model M. GH-BLUP models including both marker and NIRS data were also tested (CbSD+M, CbSG+M). Predictive abilities were obtained from a 5-fold cross-validation with 25 repetitions. Each boxplot indicates the mean (bold horizontal line), the 1st to 3rd quartile (box), and the 1st to 9th decile (whiskers).

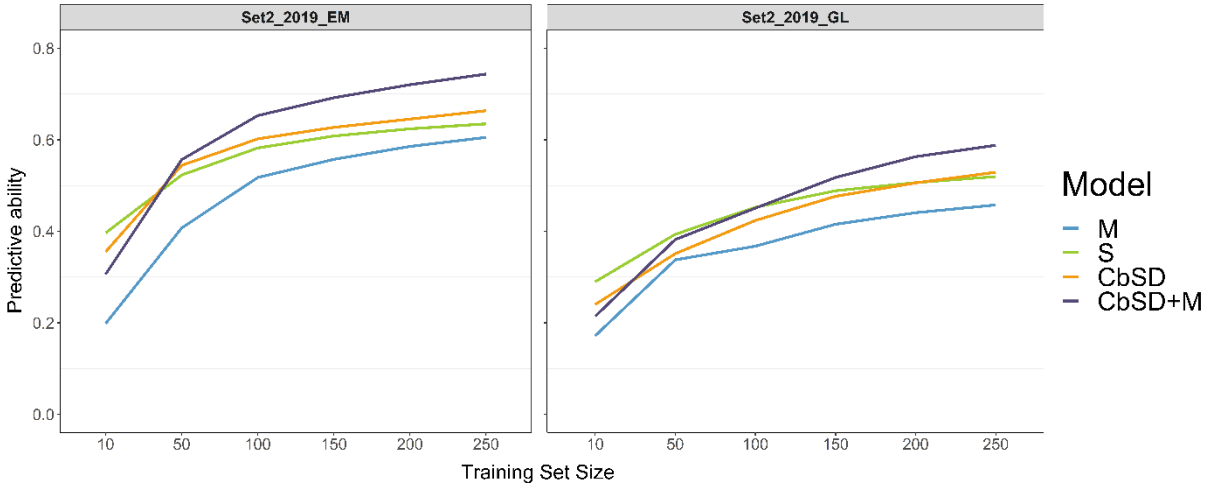

Fig. 5 Comparison of predictive abilities of GS and PS for GY as a function of the size of the training population for four predictive models. M refers to a G-BLUP GS model, S and CbSD to H-BLUP models, and CbSD+M to a GH-BLUP model. CbSD combined NIRS from two environments. Predictions were performed with Set2 data from two environments, Set2_2019_EM and Set2_2019_GL for each TP size. Lines indicate mean predictive abilities of a 5-fold cross-validation with 25 repetitions.

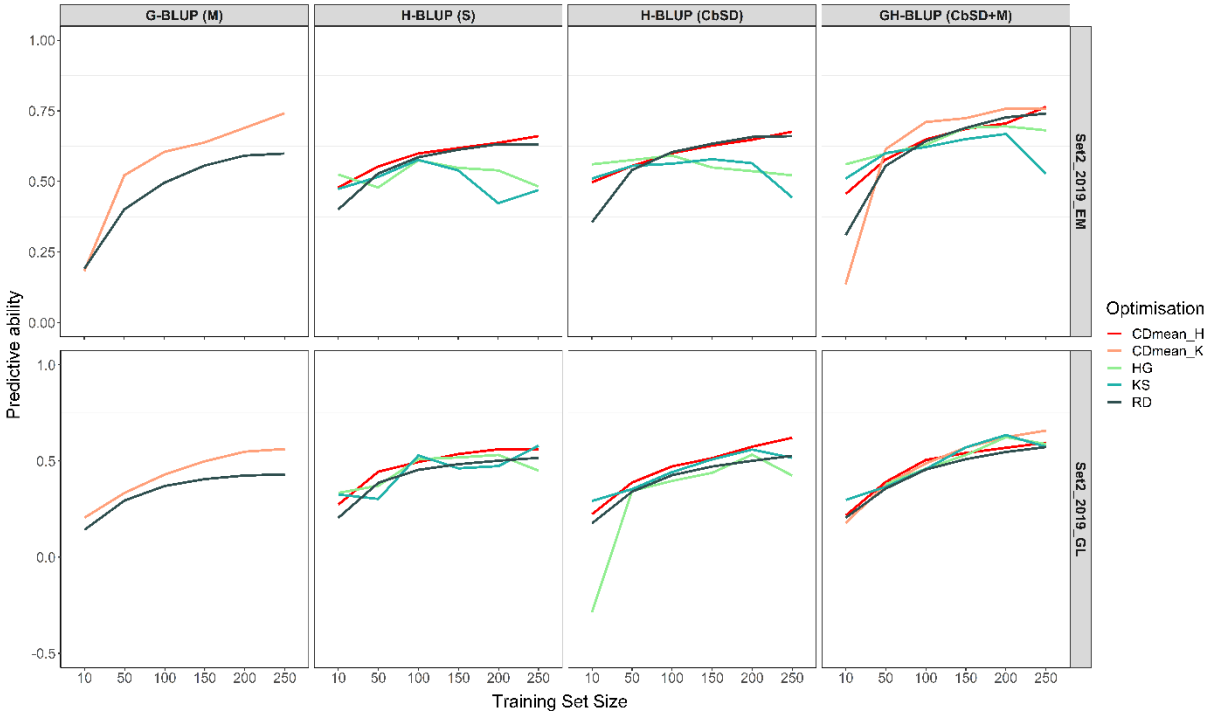

Fig. 6 Comparison of predictive abilities for GY as a function of the size of the training population for four predictive models and three optimisation algorithms. Algorithms using NIR spectra were Honigs (HG), Kennard-Stone (KS) and CDmean_H using the hyperspectral similarity matrix (*H*). CDmean_K was conducted on kinship (*K*) based on molecular markers. Optimisations are compared to a random selection (RD). Predictions were performed using Set2 data from two environments: 2019_EM and 2019_GL. Predictive models used were M, S, CbSD, CbSD+M (cf. Table3). Lines represent the predictive ability with respect to TP size. Predictive ability was averaged on 50 repetitions for CDmean_K and CDmean_H, and on 125 repetitions for RD. Predictive ability for HG and KS were obtained once because they are deterministic criteria.