

Genomic and phenomic predictions help capture low-effect alleles promoting seed germination in oilseed rape in addition to QTL analyses

Marianne Laurençon

INRAE Bretagne-Normandie: Institut National de Recherche pour l'Agriculture l'Alimentation et l'Environnement Centre Bretagne-Normandie

Julie Legrix

INRAE Bretagne-Normandie: Institut National de Recherche pour l'Agriculture l'Alimentation et l'Environnement Centre Bretagne-Normandie

Marie-Hélène Wagner

GEVES

Didier Demilly

GEVES

Cécile Baron

INRAE Bretagne-Normandie: Institut National de Recherche pour l'Agriculture l'Alimentation et l'Environnement Centre Bretagne-Normandie

Sophie Rolland

INRAE Bretagne-Normandie: Institut National de Recherche pour l'Agriculture l'Alimentation et l'Environnement Centre Bretagne-Normandie

Sylvie Ducournau

GEVES

Anne Laperche

`anne.laperche@institut-agro.fr`

L'Institut Agro Rennes-Angers <https://orcid.org/0000-0002-4930-8407>

Nathalie Nesi

INRAE Bretagne-Normandie: Institut National de Recherche pour l'Agriculture l'Alimentation et l'Environnement Centre Bretagne-Normandie

Research Article

Keywords: Germination, Brassica napus, GWAS, NIRS, Genomic Prediction, Phenomic Prediction

Posted Date: March 5th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-3985482/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Theoretical and Applied Genetics on June 11th, 2024. See the published version at <https://doi.org/10.1007/s00122-024-04659-0>.

Title: Genomic and phenomic predictions help capture low-effect alleles promoting seed germination in oilseed rape in addition to QTL analyses

Author information

Marianne Laurençon¹, Julie Legrix¹, Marie-Hélène Wagner², Didier Demilly², Cécile Baron¹, Sophie Rolland¹, Sylvie Ducournau², Anne Laperche^{1, a*}, Nathalie Nesi^{1, a}

¹ Institute of Genetics, Environment and Plant Protection (IGEPP), INRAE – Institut Agro Rennes-Angers – Université de Rennes, 35650 Le Rheu, France

² Groupe d’Etude et de contrôle des Variétés Et des Semences (GEVES), 49070, Beaucouzé, France

^a Co-last authors

* Correspondence: anne.laperche@institut-agro.fr

Keywords: Germination, *Brassica napus*, GWAS, NIRS, Genomic Prediction, Phenomic Prediction

Key message (<30words)

Phenomic prediction implemented on a large diversity set can efficiently predict seed germination, capture low-effect favorable alleles that are not revealed by GWAS and identify promising genetic resources.

Abstract (150-250 mots)

Oilseed rape faces many challenges, especially at the beginning of its developmental cycle. Achieving rapid and uniform seed germination could help to ensure a successful establishment, and therefore enabling the crop to compete with weeds and tolerate stresses during the earliest developmental stages. The polygenic nature of seed germination was highlighted in several studies, and more knowledge is needed about low- to moderate-effect underlying loci in order to enhance seed germination effectively by improving the genetic background and incorporating favorable alleles. A total of 17 QTL were detected for seed germination-related traits, for which the favorable alleles often corresponded to the most frequent alleles in the panel. Genomic and phenomic predictions methods provided moderate to high predictive abilities, demonstrating the ability to capture small additive and non-additive effects for seed germination. This study also showed that phenomic prediction better estimated breeding values than genomic prediction. Finally, as the predictive ability of phenomic prediction was less

influenced by the genetic structure of the panel, it is worth using this prediction method to characterize genetic resources, particularly with a view to design prebreeding populations.

Introduction

Oilseed rape (*Brassica napus* L.) is the leading oilseed crop in Europe (FAO 2023), but it faces numerous constraints, especially in a context of climate change and reduced chemical inputs. One of the keys to overcome these challenges would be to ensure successful plant establishment, enabling the crop to compete with weeds and tolerate stresses during the earliest developmental stages in order to guarantee a high yield potential (Nelson et al. 2022). Indeed, oilseed rape establishment is affected by a wide range of biotic and abiotic stresses, such as seed- and soil-borne pathogens causing damping-off, pest attacks reducing young plantlet biomass (slugs, flea beetles), and weed competition as well as environmental limiting factors, especially water (lack or excess) and extreme temperatures that can occur after sowings or during plant emergence (Haj Sghaier et al. 2022). These stresses can dramatically reduce plant density and biomass thus, ultimately affecting yield. Successful plant establishment results from a combination of developmental processes of which germination is the first step (Rajjou et al. 2012). As a matter of fact, rapid and highly efficient germination combined to uniform plant emergence strongly increase the likelihood of plant establishment, which is expected especially under adverse environmental conditions.

Seed germination is a complex trait subject to genetic and environmental controls. The environmental effect is generally referred to the so-called ‘seed-lot’ effect that encompass the conditions during seed development on the mother plant as well as the post-harvest conditions (Rajjou et al. 2012; Finch-Savage and Bassel 2016). In addition, the cultural practices also lead to variation in germination ability when considering factors such as the sowing date, seedbed preparation, intercropping and choice of previous crop (Elliott et al. 2007; Lamichhane et al. 2018). Several studies have reported the polygenic nature of seed germination, which is under the control of numerous loci with moderate effect each (Hatzig et al. 2015; Nguyen et al. 2018; Gad et al. 2021; Luo et al. 2021).

Modeling low additive effects distributed along the genome can be addressed through the recent development of prediction methods. The benchmark article of Meuwissen et al. (2001) conceptualized the prediction of complex traits using genotypic data, called genomic prediction (GP). GP combine phenotypic data and high-density molecular markers obtained on genotypes from a training population, to predict the genomic estimated breeding values (GEBV) of non-phenotyped genotypes from a testing population. Combined with GWAS approach, GP can accelerate genetic gain in breeding (Hickey et al. 2017; Araus et al. 2018). Other emerging methods have focused on using endophenotypes as trait predictors to complement or even to replace molecular data. This enables to

improve not only the modeling of additive effects, but also of non-additive and epistatic effects (Mackay et al. 2009; Patti et al. 2012; Ritchie et al. 2015). Multiple studies integrated transcriptomic and metabolomic data, as endophenotypes, into predictive models, either in combination with genomic data or not (Westhues et al. 2017; Schrag et al. 2018; Knoch et al. 2021). These integrations complemented the information provided by molecular markers and has demonstrated the ability to better capture small genetic effects. Rincet and coworkers (2018) proposed a low cost and high-throughput method as an alternative to the use of -omics data, based on near infrared spectroscopy (NIRS), called phenomic prediction (PP).

Up to now, most of the studies reporting the use of GP and PP models have focused on traits related to seed yield, seed quality or plant phenology mainly in cereal crops (Heffner et al. 2011; Albrecht et al. 2011; Rincet et al. 2018; Voss-Fels et al. 2019; Lane et al. 2020; Robert et al. 2022), with additional examples in other crops such as rapeseed (Würschum et al. 2014; Werner et al. 2018; Knoch et al. 2021) or other species such as trees (Resende et al. 2012; Muranty et al. 2015; Isik et al. 2016; Rincet et al. 2018; Brault et al. 2022). Considering seed germination phenotype, a first study reports the interest of GP to decipher the polygenic effect of seed germination capacity in barley (Rooney et al. 2022), with moderate to high predictive abilities.

As seed germination has rarely been considered as a selection trait in the past, it is very likely that certain genes and alleles of interest are present in both recent and ancient germplasm. Therefore, we investigate the genetic determinism of seed germination-related traits in rapeseed and the predictive ability of GP and PP models for these traits in a broad genetic pool including ancient and recent germplasm of winter, semi-winter and spring type. The combination of different germplasms provided potential access to favorable low-effect polygenes. Our goal was to investigate whether GP and PP were promising methods to improve germination capacity in addition to GWAS. Our results led to (i) identify relationships between germination-related traits, (ii) decipher the genetic control of these traits and (iii) compare multiple models of GP and PP, integrating one or multiple predictors, to evaluate the predictive abilities for seed germination-related traits.

Material and Methods

Plant material and genotyping data

The diversity set consisted of 223 genetically diverse *B. napus* inbred lines, including 127 winter oilseed rape (WOSR), 81 spring oilseed rape (SOSR), 13 winter fodder (WFR) and 2 swedes (Online Resource 1). Seed lots were all produced by open-pollination during the 2020/2021 season in Le Rheu, France. Samples of dry mature seeds were analyzed for germination time progress curves using RGB imaging and NIRS spectrum acquisition

(see below). All the accessions were genotyped using the *Brassica* 60K Infinium® SNP array (Illumina, Inc. San Diego, CA) (Clarke et al., 2016) and the data were visualized using Genome Studio software (Illumina, Inc. San Diego, CA). A total of 33,151 SNP was validated using thresholds of 5% for the minor allele frequency (MAF), 10% for the frequency of missing values and 10% for the heterozygosity level. The missing SNP data were imputed using BEAGLE software following the method described in Browning et al. (2018). SNP were physically anchored to the *B. napus* Darmor-bzh v10 reference genome (Rousseau-Gueutin et al. 2020).

Acquisition of germination data

Seed germination dynamics were monitored using the high-throughput phenotyping platform for germination at the National Seed Testing Station in Angers, France (PHENOTIC – Angers Seed Phenotyping Facility, Boureau, 2020). Seeds were imbibed under controlled conditions at 20°C, in the dark, during 96 h. The experimental design consisted of blocks of twenty-five seeds per genotype, repeated four times (Ducournau et al. 2004; Wagner et al. 2011). Imaging and analysis methods are described in Demilly et al. (2014), with pictures taken every two hours over the experiment.

Germination parameters were monitored for each block of twenty-five seeds as following (Table 1): seed volume increase after 20 h of imbibition (VI, in %), final germination percentage after 96 h (GP, in %), first germination time (FG, in h) corresponding to the time when the first seed germinates, mean germination time (MGT, in h) corresponding to the mean delay to germinate for each seed lot, and radicle elongation speed (ES, in mm/h). Supplementary parameters were deduced from the germination dynamics, such as germination after 36 h of imbibition (GP36, in %), time to reach 20% of germination (T20, in h), time to reach 50% of germination (T50, in h), time to reach 80% of germination (T80, in h), uniformity (UNI) assessed as the difference between T80 and T20, and the area under the curve (AUC) that represents the germination rate as a function of the time from the initiation of the imbibition. AUC was estimated for each genotype using the values for the 100 seeds of the four blocks. Thousand seed weight (TSW, in g) was also measured for each seed lot by weighting seed samples after drying at 105°C overnight.

Phenotypic heritability

Broad-sense heritability (H^2) was calculated for each phenotypic trait. Genotypic and error variance were extracted from the following linear mixed model:

$$Y_{ij} = \mu + G_i + e_{ij},$$

where Y_{ij} is the trait value obtained for the j^{th} repetition of the genotype i , G_i corresponds to the random genetic effect for genotype i , e_{ij} the residual effect. We assumed that G_i and e_i were independent, identically distributed and followed a normal distribution.

H^2 was then calculated as follows $H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_e^2}{n}}$, where: σ_G^2 is the genotypic variance, σ_e^2 is the error variance and n is the number of repetitions.

NIRS data

NIR spectra were collected on dry seed samples using the MPA FT-NIR spectrophotometer (Bruker Optic Inc, Germany) over the range of 4000 to 12000 cm^{-1} with a 16cm^{-1} optical resolution. NIRS were converted in nm in steps of 1nm, so the final spectra range from 800 to 2781 nm. Three biological replicates were run per genotype, with each replicate being the average of 64 technical repetitions measured by the spectrophotometer. Due to a lack of seeds, only 210 genotypes were screened (Online Resource 1). NIR spectra were centered and scaled to reduce noise. Then, the first derivative of the Savitzky-Golay filter (Savitzky and Golay 1964) was calculated using the R package signal (Ligges et al. 2014) to smooth the curve. The three replicates were then averaged to obtain a mean NIR spectrum per genotype.

Exploratory analyses and correlations

To explore the relationship between the different seed germination traits, a principal component analysis (PCA) was performed using FactomineR package (Lê et al. 2008). The number of dimensions was chosen according to the percentage of explained variance. Then a clustering was carried out, using the partitioning around medoids (PAM) method (Kaufman and Rousseeuw 1990). The number of clusters (k) was chosen using silhouette index and gap statistic from cluster R package (Maechler et al. 2022). Due to nonlinear relationships, Spearman correlations were calculated between traits.

Population genetic structure and diversity

Genetic structure of the germplasm was unraveled using the first three components of a principal coordinates analysis (PCoA, Gower, 1967) carried out using the SNP data. The PCoA is based on the genetic distances between genotypes estimated by the dissimilarity matrix $(1-K)$. The genomic kinship (K) was estimated using Astle and Balding (2009) algorithm as following :

$$K = \frac{WW^T}{M}$$

With W the matrix scaled on allelic frequencies with dimensions NxM, N is the number of genotypes and M the number of molecular markers and W^T is the transposed W matrix.

Nucleotidic diversity (π) per chromosome was estimated using VCFtools (Danecek et al. 2011).

Genome wide association study

For the GWAS only, a kinship matrix was re-estimated for each chromosome tested as described by Rincent et al., (2014). The estimation of the kinship matrix (K) is similar to the methodology previously described, except that SNP located on the chromosome under investigation were discarded. The pairwise Linkage Disequilibrium (LD) between SNPs (r^2) were estimated per chromosome using PLINK software v.1.9 (Purcell et al. 2007).

GWAS was performed using FaST-LMM algorithm (Lippert et al. 2011). To reduce false associations (types I and II errors) due to population structure and kinship between genotypes, GWAS was performed with a mixed model that takes these two factors into account (Yu et al. 2006). Due to high inter-marker LD (0.51), the SimpleM approach (Gao et al. 2008, 2010) was used to reduce type I error (false-positive) by estimating the effective number of independent tests (M_{eff}) based on composite LD for the GWAS. Markers were considered as significantly associated with a trait if the $-\log_{10}(\text{p-value})$ exceeded a 5% threshold (using our data, the 5% threshold was 3.51). In addition, false discovery rate (FDR) adjusted q-values were calculated to reduce type I errors using the q value R package (Storey 2002). Markers with a q-value inferior to 0.2 were retained.

QTL confidence intervals were estimated based on the method proposed by Albert et al. (2016). Briefly, LD calculation was performed between 100,000 randomly chosen pairs of unlinked loci located on different chromosomes. The critical LD threshold was chosen as the 95th percentile of the LD distribution, which equaled 0.16 using our data. Then, local pairwise LD with markers located upstream and downstream the significant marker (on the same chromosome) were calculated in the same way as pairwise LD explained above. Confidence intervals were estimated for each significant marker as the interval between the first (upstream) and last (downstream) marker that presented a LD value higher than the LD threshold when compared to the significant marker.

Effects of each single QTL (R^2 and allelic effect) were estimated at the peak marker that corresponded to the marker with the smallest p-value. Favorable alleles were identified for each QTL as the allele that improved the phenotype, that is to say increasing GP36, TSW and VI values but decreasing MGT, T20 and T50 values. The number of favorable alleles across all QTL locations for each genotype and trait was calculated, ranging from 0 to

the maximum number of QTL identified for the trait. An ANOVA and Tukey test at 5% probability level were performed to compare the effect of accumulating favorable alleles at QTLs for each trait.

Genomic and phenomic predictions

Genomic heritability along NIR-based spectra was estimated for each wavelength from a statistical model considering a random polygenic effect. Genotypic and error variance were extracted from the following linear mixed model:

$$Y_{ij} = \mu + G_i + e_{ij},$$

where Y_{ij} is the spectrum value obtained for the j^{th} repetition of the genotype i , G_i corresponds to the random genetic effect for genotype i , following a normal distribution $G \sim N(0, K\sigma_G^2)$ with K the kinship (see above), e_{ij} the residual effect. We assumed that G_i and e_i were independent, identically distributed and followed a normal distribution.

Genomic heritability was then calculated as follows: $H_{genomic}^2 = \frac{\hat{\sigma}_G^2}{\hat{\sigma}_G^2 + \hat{\sigma}_e^2}$, with $\hat{\sigma}_G^2$ and $\hat{\sigma}_e^2$ the REML estimates of σ_G^2 and σ_e^2 obtained using the R package sommer (Covarrubias-Pazaran 2016).

A spectral matrix H was calculated to represent the kinship between genotypes based on the NIR spectra similarity.

$$H = \frac{SS^T}{N_w}$$

With S the NIR spectra matrix of dimension $N \times N_w$. The S matrix gathered the value of absorbance pretreated as presented above, for each genotype and each wavelength. Values were also centered and scaled. N represents the number of genotypes and N_w the number of wavelengths. Pearson correlation was realized to compare the spectral relationship matrix H with the genomic kinship matrix K .

G-BLUP and H-BLUP (using the K or H matrix respectively) were used to predict seed germination-related traits are defined as follow:

$$(1) \text{ G-BLUP} \quad Y_i = \mu + G_i + \varepsilon_i$$

$$(2) \text{ H-BLUP} \quad Y_i = \mu + W_i + \varepsilon_i$$

With Y_i the mean phenotype value for the genotype i , μ the intercept, G_i or W_i the random genetic effect following a normal distribution $G \sim N(0, K\sigma_G^2)$ with K the genomic kinship matrix (see above) or $W \sim N(0, H\sigma_G^2)$ with H the hyperspectral matrix and ε is the random residual effect following $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

Both genomic (K) and spectral (H) matrices were included simultaneously in a GH-BLUP model by integrating two variance-covariance matrices as follow:

207 (3) GH-BLUP $Y_i = \mu + G_i + W_i + \varepsilon_i$,

208 with $G \sim N(0, K\sigma_G^2)$, and $W \sim N(0, H\sigma_W^2)$.

209 GP and PP models were assessed for each trait using cross-validation (CV). Training set was composed of 4 folds
210 over 5 of genotypes chosen randomly, and the testing set was composed of the remaining fold. CV was repeated
211 100 times. These models were fitted using sommer R package (Covarrubias-Pazaran 2016). PA were estimated for
212 each model by the Pearson correlation between the observed and the predicted values of the validation set.

213 By homology with the values estimated by genomic prediction, called GEBV, we called the values estimated by
214 phenomic prediction PEPV, for phenomic estimated phenotypic values.

215

216 **Results**

217 **Phenotypic distribution and correlations between traits**

218 All germination-related traits displayed moderate to high heritabilities (Table 1). The highest heritabilities (>0.60)
219 were observed for GP36, T20, T50, T80 and MGT and the lowest heritabilities were observed for VI (0.40) and
220 UNI (0.43). GP, GP36, T20, T50, T80, AUC, MGT were highly correlated as shown on Table 2. GP, AUC and
221 GP36 were positively correlated as well as T20, T50, T80, MGT, which can be partially explained by the non-
222 independence of traits. However, these two groups of variables presented negative correlations one with each
223 other. VI and ES presented significant but moderate correlations with other traits (Table 2). TSW was only
224 significantly correlated to UNI (-0.19), ES (0.19) and VI (0.52). These correlations were strengthened by the
225 correlation circle of the PCA (Fig. 1). Indeed, AUC, GP, GP36, T80, T20, T50, MGT were highly correlated to
226 the first axis that gathered 60.42% of the variability while TSW and VI were correlated to the second axis that
227 explained 12.43% of the variability. ES, UNI and FG were less correlated to these two principal components.

228

229 **Germplasm structure using genomic, phenotypic and spectral data revealed different patterns**

230 Genetic structure of the germplasm was assessed using the first two axes of the Principal Coordinates Analysis
231 using SNP (PCoA), explaining 48% of the genetic variance (Fig. 2a). The first axis that encompasses 34% of
232 genetic variance clearly discriminated winter and spring types. The second axis (14% of the genetic variance)
233 separated Asian genotypes on the one side, and European and American genotypes on the other side. As a whole,
234 four genetic clusters were identified using the PCoA clustering. Cluster 1 (n=114, red) was mostly composed of
235 European WOSR and WFR but also included some Asian WOSR. Cluster 2 (n=62, green) was mostly composed
236 of European and North-American SOSR, but also gathered some Oceanian SOSR. Cluster 3 (n=23, blue)

corresponded to Asian SOSR and WOSR. Finally, cluster 4 (n=24, purple) was composed of European winter fodder, Asian and European WOSR. The mean nucleotide diversity π was computed on the whole germplasm and equaled 7.69×10^{-5} , indicating an important genetic diversity within the germplasm used in this study.

The relatedness between genotypes was also estimated using NIRS data (Fig. 2b). Heritabilities of the NIR spectra ranged between 0.00 to 0.95 regarding the wavelengths, with 34% and 16% of the wavelengths having a heritability superior or equal to 0.50 and 0.70 respectively (Online Resource 2). No cluster of genotypes was evidenced based on the NIRS data (Fig. 2b). In addition, when the genotypes were labelled according to their genetic cluster, no specific pattern was highlighted. Indeed, the correlation between genomic (K) and hyperspectral (H) kinship matrices was equal to 0.03.

Based on the germination data, three phenotypic clusters of genotypes representing their seed germination performance were identified by PAM clustering (Fig. 2c). The first phenotypic cluster (light gray) corresponded to genotypes with a high germination capacity (high GP, GP36, AUC and low UNI, MGT, T80, T20, T50). The second phenotypic cluster (medium gray) consisted of genotypes with an intermediate behavior. Finally, the third cluster (dark gray) was composed of genotypes with poor germination ability (low GP, GP36, AUC and high MGT, T20, T50, T80, UNI). Each cluster gathered SOSR, WFR and WOSR types (Fig. 2c) and there was no obvious match with the genetic clusters previously identified on the basis of SNP. Therefore, there was no obvious and significant linkage between rapeseed type or origin and their seed germination behaviour.

Identification of multiple QTL confirmed the polygenic determinism of seed germination

A total of 17 unique regions (QTL) were detected by GWAS for seed germination traits (GP36, MGT, T20, T50, VI), and thousand seed weight (TSW), with an explained variance (R^2) ranging from 6.30% to 8.78% (Table 3). Among these 17 QTL, 3 were identified for four germination traits (GP36, MGT, T20, T50) on A06 (QTL.A06.1), A09 (QTL.A09.1) and C02 (QTL.C02.2). QTL.C01.1 was identified for GP36 and MGT, QTL.A06.2 for MGT and T50 and QTL.C04.2 for MGT and T20. Few QTL detected were trait-specific: 3 for GP36 (QTL.A07.1, QTL.C02.1, QTL.C07.1), 3 for MGT (QTL.A10.1, QTL.C04.1, QTL.C05.1), 2 for T50, (QTL.A07.2 and QTL.C07.2), 1 for TSW (QTL.A10.2) and 2 for VI (QTL.C03.1 and QTL.C09.1) (Table 3). TSW QTL and VI QTL did not collocate with any QTL detected for the other traits. For most of the QTL, the favorable allele was the major allele in the germplasm (Table 3). Indeed, this was the case for one of the two QTL for VI, 5 QTL out of 9 for MGT, 3 QTL out of 4 for T20 and 5 QTL out of 6 for T50. On the contrary, for GP36, 5 QTL out of 7 had the minor allele as the favorable one. On the same line, the only QTL detected for TSW presented the minor allele as

the favorable one. For all traits, a positive effect of the accumulation of favorable alleles was observed (Online Resource 3).

Relative performances of PP and GP at the whole panel level

Predictions of seed germination traits were run either using genomic (GBLUP) or phenomic (HBLUP) prediction models (Fig. 3). Models that used a combination of the genomic and the spectral kinships were also tested (GHBLUP). In a first approach, we consider a first cross-validation (CV) scenario where the whole panel was used as training and prediction set (CV_{whole}). More precisely, the training set included 80% of the whole population chosen randomly, the validation set being the remaining 20% of the whole population. Overall, low to medium predictive abilities (PA) were obtained, varying between 0.40 (FG, UNI) and 0.59 (T20) for GBLUP, 0.32 (FG) and 0.50 (T20, MGT) for HBLUP and between 0.13 (GP36) and 0.34 (GP) for GHBLUP. For all traits, the PA of GBLUP appeared to be slightly higher or similar to the PA of HBLUP (Fig. 3). For instance, for GP36, the overall predictive ability of the genomic prediction was moderate (0.59), whereas the PA of HBLUP equaled 0.49. Furthermore, GHBLUP, was less performant than GBLUP and HBLUP alone for all traits. Interestingly, when genotypes were labelled according to their genetic cluster affiliation, a separation between the different genetic clusters was revealed when looking at the GBLUP results, each genetic cluster corresponding to a specific stratum (Fig. 4). This result was observed for all traits (Online Resource 4). This highlighted the impact of the genetic structure on the genomic prediction model. Such pattern was less observed for the phenomic prediction model (Fig. 4, Online Resource 4). In conclusion, when the CV_{whole} scenario was considered, we found that PP was less performant but also less impacted by population structure than GP.

Relative performances of PP and GP for unstructured population

As genetic structure influenced genomic predictions, particular attention was paid to investigating the most efficient way of predicting a specific genetic cluster. For this purpose, the genotypic cluster 1 was chosen as target population, as it is the only cluster that reached a sufficient size to allow model calibration ($n=114$, but only 103 genotypes used due to missing NIRS, Fig. 2a). The cluster 1 gathered most of the winter oilseed rape genotypes. Therefore, two scenarios were compared (Fig. 3). In the first scenario, CV_{local} , cluster 1 genotypes were used as calibration and validation sets. More precisely, 80% of cluster 1 genotypes were included in the training set and the remaining 20% composed the validation set. In the second scenario, $CV_{\text{whole/local}}$, 80% of the whole panel was used in the training set, while the validation set was composed of the genotypes of the cluster 1 that were not

present in the calibration set. Each scenario was repeated 100 times. Some PA could not be calculated due to non-convergent model repetitions. A first observation was that PA calculated for these two scenarios were lower than PA obtained for the whole panel (Fig. 3), which can be explained by the fact that the genomic structure of the panel caused a bias in the PA estimation. Furthermore, PA of GBLUP or HBLUP models for scenarios $CV_{\text{whole/local}}$ and CV_{local} were similar. But we observed that HBLUP model provided higher PA than the other two models, for all traits. In addition, GBLUP models outperformed GHBLUP models or obtained similar results (Fig. 3). Consequently, in case of unstructured population, PP led to higher PA values than GP.

Choosing genotypes based on their PEPV rather than on their GEBV provided higher selection differentials

The performance of GBLUP and HBLUP models in selecting the top 10% of genotypes from the whole panel was investigated (Table 4). The mean of the BLUE values of the whole panel for each trait was compared to the mean BLUE value of the top 10% genotypes. These top 10% genotypes were chosen according to three methods: i) to their BLUE value, that approximates their true breeding values (TBV), ii) to their genomic estimated breeding values (GEBV) or iii) to their phenomic estimated phenotypic values (PEPV). The selection differential (S) was estimated for each sampling method as the difference between the mean whole population TBV and the top 10% mean TBV. The selection differentials were higher when genotypes were chosen on their PEPV than when they were chosen according their GEBV (average increase in S between S estimated for GEBV and S for PEPV = 29%), allowing potentially greater genetic gain. Concordances between best and worst (Online Resource 5) selected genotypes according their TBV and GEBV or between TBV and PEPV were studied in order to evaluate GP and PP ranking, for each scenario. Average concordances of 14.4% (TBV vs PEPV) and 7.2% (TBV vs GEBV) were obtained. Similar concordances were obtained for the 10% worst genotypes, with a rate of 13.6% between TBV and PEPV, and 8.2% between TBV and GEBV.

Discussion

In an attempt to get insights into the genetic architecture of seed germination in oilseed rape, we used a combination of genetic methods based on molecular as well as phenotypic and spectral predictors. The main results led to the highlight of 17 genomic regions that control seed germination related traits in the large genetic diversity used. Moreover, genomic and phenomic prediction methods provided moderate to high predictive abilities, demonstrating the capacity to capture small additive and non-additive effects for seed germination. This study also

provided the first application of phenomic prediction in oilseed rape and demonstrated the higher ability of phenomic prediction to estimate breeding values closer to True Breeding Values compared to genomic prediction.

Alleles promoting seed germination are almost all fixed in oilseed rape

A total of 17 QTL with moderate effects were detected confirming the polygenic nature of seed germination as illustrated in previous studies (Bettey et al. 2000; Basnet et al. 2015; Hatzig et al. 2015; Nguyen et al. 2018). Six of these genomic regions controlled several traits (e.g., GP36, MGT, T20 and T50), which was consistent with the high correlations between these traits ($r < -0.88$ or $r > 0.96$). This suggest pleiotropy or linkage drag. Therefore, it would be interesting to study the underlying genes of these six specific regions. We found no common QTL between those identified in our study and those identified in the studies deciphering the genetic control of seed germination *sensus stricto* in *Brassica napus* (Hatzig et al. 2015; Nguyen et al. 2018; Boter et al. 2019; Gad et al. 2021; Luo et al. 2021), *Brassica oleracea* (Bettey et al. 2000) or *Brassica rapa* (Basnet et al. 2015) under optimal or stressed conditions. This lack of correspondence could be explained by the fact that these different studies (except Hatzig et al. (2015)) used restricted genetic diversity populations, limited either to WOSR or SOSR genotypes. Whereas we analyzed a larger genetic diversity.

To estimate the ability to improve seed germination in breeding by Marker Assisted Selection, we looked at the effect of QTL stacking for each trait. The accumulation of several (>3) favorable alleles increased seed germination speed. However, for most of the traits used in this study, the favorable alleles were also the most frequent alleles in our germplasm. Consequently, a low genetic gain is expected. Similarly, Hatzig et al. (2015) showed an improvement of T50 when stacking multiple favorable haplotypes. They also showed that the favorable haplotypes for this trait were the most frequent in the population, suggesting a strong selection for T50. Interestingly, some minor alleles were identified as favorable, for some traits involved in seed germination capacity, such as GP36. Therefore, improving seed germination capacity could still be achieved by targeting these favorable alleles presenting a low frequency in the population.

SNP-based genetic structure of oilseed rape diversity did not match to the structure observed using phenotypic nor spectral data

As shown in previous studies of oilseed rape genetic diversity (Diers and Osborn 1994; Hasan et al. 2006; Bus et al. 2011), we observed two specific genomic clusters for winter and spring accessions using SNP data. However, this clustering was not observed using seed germination phenotypic data nor NIRS data, supporting that the

genomic structure has no influence on seed germination capacity. Furthermore, kinship estimated based on NIRS data (H matrix) was not correlated to the one based on SNP data (K matrix). Similar results were obtained for soybean RIL populations (Zhu et al. 2021) or for Dent and Flint maize populations (Weiß et al. 2022). Brault *et al.* (2022) also reported a low correlation between NIRS-based H matrix and SNP-based K matrix using a diversity panel of grapevine. A single study for triticale revealed a H matrix well correlated to the genomic (Zhu et al. 2022). The low correlations between H and K matrices, might result from different histories of domestication and selection or from different range of considered genetic diversity.

Genomic and phenomic predictions allowed capturing weak polygenic effects

After seeking to highlight moderate additive effects by GWAS, genomic and phenomic prediction models were used to consider polygenic background. The combination of spectral and genomic data in the model called GHBLUP performed worse than HBLUP and worse or similar than GBLUP. Similarly, Brault et al. (2022), who identified weak correlations between K and H matrices, obtained no gain in PA by combining NIRS and SNP information. However, these results differed from previous studies using GHBLUP model, which showed superior PA compared to GBLUP or HBLUP models on yield traits (Krause et al. 2019; Galán et al. 2020; Zhu et al. 2021; Robert et al. 2022). In the case of soybean, the rrBLUP model combining SNP and NIRS data showed a higher PA than the genomic rrBLUP, even if K and H matrices were not correlated (Zhu et al. 2021). The type of trait studied, its genetic architecture and, in particular, the proportion of phenotypic value explained by non-additive genetic effects, could explain these differences in GHBLUP performance. Therefore, this information could help to choose between genomic, phenomic and combined predictions models. In particular, the use of NIRS data and PP models could be favored for traits moderately to strongly shaped by non-additive genetic effects. It is also necessary to remain cautious regarding the absorbance values for each wavelength. They can be influenced (i) by the environment in which the seeds were formed and stored, which can have an impact on the PA when predicting in independent environments, and (ii) by the environment in which the spectra were acquired (humidity, temperature), resulting in heterogeneous data over time. In addition, PA overestimation can result from the fact that the same samples are used both for the acquisition of NIRS data and for the acquisition of the phenotypic data. Dallinger et al. (2023) also warn against PA overestimation due to the correlation between traits to be predicted and seed compounds as well against unintentional selection of seed compounds in the selected population on the basis of its PEPV. Finally, it is necessary to remain vigilant regarding the traits chosen for prediction.

In this study, both GP and PP methodologies predicted well germination traits in the CV_{whole} scenario (up to 0.59 for GBLUP on GP36 and 0.50 for HBLUP on T20). However, we showed an impact of the genomic structure on the PA of GP but not with PP models (Fig. 3). This genomic structure impact on genomic prediction models was also reported previously (Riedelsheimer et al. 2013; Schopp et al. 2017; Werner et al. 2020). The absence of impact of genomic structure on PP is in accordance with Zhu et al. (2021, 2022) and Weiß et al. (2022) studies. Therefore, HBLUP models should be preferred to predict complex traits for genetically diverse germplasm.

Interest of phenomic prediction in breeding to incorporate genetic diversity in elite pools

Our results confirm that seed germination performance could be increased through genomic and/or phenomic prediction, highlighting the interest in improving the genetic background. We showed that PP provided PEPV closer to TBV than GEBV. Therefore, selecting according to PEPV is more efficient and thus increase the genetic gain. It would be necessary to confirm these results using independent phenotyping trials. According to the breeder equation (Lush, 1943), the impact of PP could be interpreted according to the different terms of the equation: concerning the (i) the generation interval, no impact is expected as seed germination phenotype is acquired at the beginning of the cycle. Considering (ii) the intensity i , NIRS acquisition is simple, fast, and inexpensive, which would allow to screen a higher number of genotypes. Therefore, more resources could be allocated to construct the training population, which then could lead to a better PA and therefore to a better identification of the best promising individuals. This resources reallocation could subsequently also contribute to phenotype these promising individuals more reliably and in different environments. As a whole, PP is a mean to increase the selection intensity. However, a specific attention has to be paid to maintain genetic diversity within the breeding population to assure long-term genetic gain. This can be achieved, (i) by limiting or optimizing the selection intensity, (ii) by considering genetic diversity in the choice of parents to cross, using for example optimal contribution selection (Cowling et al. 2017) or (iii) by integrating exotic material into elite genotypes (Simmonds 1993). For this last point, we propose to use PP to efficiently identify genetic diversity to be valorized. Integrating this genetic diversity into elite germplasm would increase the frequency of favorable alleles in the population. Indeed, integrating exotic diversity into elite germplasm is one of the most important challenges in field crops (Cowling 2013), especially in oilseed rape. Oilseed rape has a narrow genetic diversity and could take advantage of this gain of diversity to overcome biotic and abiotic stresses while maintaining or increasing yields. For this purpose, the creation of pre-bridging and bridging populations (i.e. progeny obtain from elite x exotic crosses to be integrated into elite germplasm) have been proposed (Gorjanc et al. 2016; Allier et al. 2020; Sanchez et al.

2023). The construction of these (pre-)bridging populations could benefit from phenomic prediction. PP could help identifying material from a broad range of exotic resources for the creation and maintenance of pre-bridging and bridging populations.

References

- Albert E, Segura V, Gricourt J, et al (2016) Association mapping reveals the genetic architecture of tomato response to water deficit: focus on major fruit quality traits. *Journal of Experimental Botany* 67:6413–6430. <https://doi.org/10.1093/jxb/erw411>
- 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>
- Allier A, Teyssèdre S, Lehermeier C, et al (2020) Optimized breeding strategies to harness genetic resources with different performance levels. *BMC Genomics* 21:349. <https://doi.org/10.1186/s12864-020-6756-0>
- Araus JL, Kefauver SC, Zaman-Allah M, et al (2018) Translating High-Throughput Phenotyping into Genetic Gain. *Trends in Plant Science* 23:451–466. <https://doi.org/10.1016/j.tplants.2018.02.001>
- Astle W, Balding DJ (2009) Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* 24:451–471. <https://doi.org/10.1214/09-STS307>
- Basnet RK, Duwal A, Tiwari DN, et al (2015) Quantitative Trait Locus Analysis of Seed Germination and Seedling Vigor in Brassica rapa Reveals QTL Hotspots and Epistatic Interactions. *Frontiers in Plant Science* 6:1032. <https://doi.org/10.3389/fpls.2015.01032>
- Betty M, Finch-Savage WE, King GJ, Lynn JR (2000) Quantitative genetic analysis of seed vigour and pre-emergence seedling growth traits in Brassica oleracea. *New Phytol* 148:277–286. <https://doi.org/10.1046/j.1469-8137.2000.00760.x>
- Boter M, Calleja-Cabrera J, Carrera-Castaño G, et al (2019) An Integrative Approach to Analyze Seed Germination in Brassica napus. *Frontiers in Plant Science* 10:1342. <https://doi.org/10.3389/fpls.2019.01342>
- Boureau T (2020) PHENOTIC Platform
- Brault C, Lazerges J, Doligez A, et al (2022) Interest of phenomic prediction as an alternative to genomic prediction in grapevine. *Plant Methods* 18:108. <https://doi.org/10.1186/s13007-022-00940-9>
- Browning BL, Zhou Y, Browning SR (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics* 103:338–348. <https://doi.org/10.1016/j.ajhg.2018.07.015>
- Bus A, Körber N, Snowdon RJ, Stich B (2011) Patterns of molecular variation in a species-wide germplasm set of Brassica napus. *Theor Appl Genet* 123:1413–1423. <https://doi.org/10.1007/s00122-011-1676-7>
- Clarke WE, Higgins EE, Plieske J, et al (2016) A high-density SNP genotyping array for Brassica napus and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theor Appl Genet* 129:1887–1899. <https://doi.org/10.1007/s00122-016-2746-7>
- Covarrubias-Pazarán 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>
- Cowling WA (2013) Sustainable plant breeding. *Plant Breeding* 132:1–9. <https://doi.org/10.1111/pbr.12026>

454 Cowling WA, Li L, Siddique KHM, et al (2017) Evolving gene banks: improving diverse populations of crop
455 and exotic germplasm with optimal contribution selection. *Journal of Experimental Botany* 68:1927–
456 1939. <https://doi.org/10.1093/jxb/erw406>

457 Dallinger HG, Löschenberger F, Bistrich H, et al (2023) Predictor bias in genomic and phenomic selection.
458 *Theor Appl Genet* 136:235. <https://doi.org/10.1007/s00122-023-04479-8>

459 Danecek P, Auton A, Abecasis G, et al (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156–
460 2158. <https://doi.org/10.1093/bioinformatics/btr330>

461 Demilly D, Ducournau S, Wagner M-H (2014) Digital imaging of seed germination. In: *Plant Image Analysis:
462 Fundamentals and Applications*, Gupta, S. D., Ibaraki, Y. CRC Press, Boca Raton, pp 147–162

463 Diers BW, Osborn TC (1994) Genetic diversity of oilseed *Brassica napus* germ plasm based on restriction
464 fragment length polymorphisms. *Theoret Appl Genetics* 88:662–668.
465 <https://doi.org/10.1007/BF01253968>

466 Ducournau S, Feutry A, Plainchault P, et al (2004) An image acquisition system for automated monitoring of the
467 germination rate of sunflower seeds. *Computers and Electronics in Agriculture* 44:189–202.
468 <https://doi.org/10.1016/j.compag.2004.04.005>

469 Elliott RH, Mann LW, Johnson EN, et al (2007) Vigor Tests for Evaluating Establishment of Canola under
470 Different Growing Conditions and Tillage Practices. *Seed Technol* 29:21–36

471 FAO (2023) FAOSTAT Food and agriculture data

472 Finch-Savage WE, Bassel GW (2016) Seed vigour and crop establishment: extending performance beyond
473 adaptation. *J Exp Bot* 67:567–591. <https://doi.org/10.1093/jxb/erv490>

474 Gad M, Chao H, Li H, et al (2021) QTL Mapping for Seed Germination Response to Drought Stress in *Brassica
475 napus*. *Frontiers in Plant Science* 11:629970. <https://doi.org/10.3389/fpls.2020.629970>

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

479 Gao X, Starmer J, Martin ER (2008) A multiple testing correction method for genetic association studies using
480 correlated single nucleotide polymorphisms. *Genetic Epidemiology* 32:361–369.
481 <https://doi.org/10.1002/gepi.20310>

482 Gao X, Becker LC, Becker DM, et al (2010) Avoiding the high Bonferroni penalty in genome-wide association
483 studies. *Genetic Epidemiology* 34:100–105. <https://doi.org/10.1002/gepi.20430>

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

487 Gower JC (1967) Multivariate Analysis and Multidimensional Geometry. *Journal of the Royal Statistical Society
488 Series D (The Statistician)* 17:13–28. <https://doi.org/10.2307/2987199>

489 Haj Sghaier A, Tarnawa Á, Khaeim H, et al (2022) The Effects of Temperature and Water on the Seed
490 Germination and Seedling Development of Rapeseed (*Brassica napus* L.). *Plants* 11:2819.
491 <https://doi.org/10.3390/plants11212819>

492 Hasan M, Seyis F, Badani AG, et al (2006) Analysis of Genetic Diversity in the *Brassica napus* L. Gene Pool
493 Using SSR Markers. *Genet Resour Crop Evol* 53:793–802. <https://doi.org/10.1007/s10722-004-5541-2>

494 Hatzig SV, Frisch M, Breuer F, et al (2015) Genome-wide association mapping unravels the genetic control of
495 seed germination and vigor in *Brassica napus*. *Frontiers in Plant Science* 6:221.
496 <https://doi.org/10.3389/fpls.2015.00221>

497 Heffner EL, Jannink J-L, Sorrells ME (2011) Genomic Selection Accuracy using Multifamily Prediction Models
 498 in a Wheat Breeding Program. *The Plant Genome* 4:65–75.
 499 <https://doi.org/10.3835/plantgenome2010.12.0029>

500 Hickey JM, Chiurugwi T, Mackay I, Powell W (2017) Genomic prediction unifies animal and plant breeding
 501 programs to form platforms for biological discovery. *Nat Genet* 49:1297–1303.
 502 <https://doi.org/10.1038/ng.3920>

503 Isik F, Bartholomé J, Farjat A, et al (2016) Genomic selection in maritime pine. *Plant Science* 242:108–119.
 504 <https://doi.org/10.1016/j.plantsci.2015.08.006>

505 Kaufman L, Rousseeuw P (1990) Partitioning Around Medoids (Program PAM). *Finding Groups in Data* 68–
 506 125. <https://doi.org/10.1002/9780470316801.ch2>

507 Knoch D, Werner CR, Meyer RC, et al (2021) Multi-omics-based prediction of hybrid performance in canola.
 508 *Theor Appl Genet* 134:1147–1165. <https://doi.org/10.1007/s00122-020-03759-x>

509 Krause MR, González-Pérez L, Crossa J, et al (2019) Hyperspectral Reflectance-Derived Relationship Matrices
 510 for Genomic Prediction of Grain Yield in Wheat. *G3 Genes|Genomes|Genetics* 9:1231–1247.
 511 <https://doi.org/10.1534/g3.118.200856>

512 Lamichhane JR, Debaeke P, Steinberg C, et al (2018) Abiotic and biotic factors affecting crop seed germination
 513 and seedling emergence: a conceptual framework. *Plant Soil* 432:1–28. [https://doi.org/10.1007/s11104-](https://doi.org/10.1007/s11104-018-3780-9)
 514 [018-3780-9](https://doi.org/10.1007/s11104-018-3780-9)

515 Lane HM, Murray SC, Montesinos-López OA, et al (2020) Phenomic selection and prediction of maize grain
 516 yield from near-infrared reflectance spectroscopy of kernels. *The Plant Phenome Journal* 3:e20002.
 517 <https://doi.org/10.1002/ppj2.20002>

518 Lê S, Josse J, Husson F (2008) FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical*
 519 *Software* 25:1–18. <https://doi.org/10.18637/jss.v025.i01>

520 Ligges U, Short T, Kienzle P, et al (2014) signal: Signal processing

521 Lippert C, Listgarten J, Liu Y, et al (2011) FaST linear mixed models for genome-wide association studies. *Nat*
 522 *Methods* 8:833–835. <https://doi.org/10.1038/nmeth.1681>

523 Luo T, Zhang Y, Zhang C, et al (2021) Genome-Wide Association Mapping Unravels the Genetic Control of
 524 Seed Vigor under Low-Temperature Conditions in Rapeseed (*Brassica napus* L.). *Plants* 10:426.
 525 <https://doi.org/10.3390/plants10030426>

526 Lush JL (1943) *Animal Breeding Plans*. The Iowa State College Press, Ames

527 Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. *Nat Rev*
 528 *Genet* 10:565–577. <https://doi.org/10.1038/nrg2612>

529 Maechler M, Rousseeuw P, Struyf A, et al (2022) cluster: Cluster Analysis Basics and Extensions

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

532 Muranty H, Troggio M, Sadok IB, et al (2015) Accuracy and responses of genomic selection on key traits in
 533 apple breeding. *Horticulture Research* 2:15060. <https://doi.org/10.1038/hortres.2015.60>

534 Nelson MN, Nesi N, Barrero JM, et al (2022) Strategies to improve field establishment of canola: A review.
 535 *Advances in Agronomy* 175:133–177. <https://doi.org/10.1016/bs.agron.2022.05.001>

536 Nguyen TCT, Abrams SR, Friedt W, Snowdon RJ (2018) Quantitative trait locus analysis of seed germination,
 537 seedling vigour and seedling-regulated hormones in *Brassica napus*. *Plant Breed* 137:388–401.
 538 <https://doi.org/10.1111/pbr.12576>

539 Patti GJ, Yanes O, Siuzdak G (2012) Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*
540 13:263–269. <https://doi.org/10.1038/nrm3314>

541 Purcell S, Neale B, Todd-Brown K, et al (2007) PLINK: A Tool Set for Whole-Genome Association and
542 Population-Based Linkage Analyses. *Am J Hum Genet* 81:559–575. <https://doi.org/10.1086/519795>

543 Rajjou L, Duval M, Gallardo K, et al (2012) Seed Germination and Vigor. *Annual Review of Plant Biology*
544 63:507–533. <https://doi.org/10.1146/annurev-arplant-042811-105550>

545 Resende MDV, Resende Jr MFR, Sansaloni CP, et al (2012) Genomic selection for growth and wood quality in
546 Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest
547 trees. *New Phytologist* 194:116–128. <https://doi.org/10.1111/j.1469-8137.2011.04038.x>

548 Riedelsheimer C, Endelman JB, Stange M, et al (2013) Genomic Predictability of Interconnected Biparental
549 Maize Populations. *Genetics* 194:493–503. <https://doi.org/10.1534/genetics.113.150227>

550 Rincent R, Moreau L, Monod H, et al (2014) Recovering Power in Association Mapping Panels with Variable
551 Levels of Linkage Disequilibrium. *Genetics* 197:375–387. <https://doi.org/10.1534/genetics.113.159731>

552 Rincent R, Charpentier J-P, Faivre-Rampant P, et al (2018) Phenomic Selection Is a Low-Cost and High-
553 Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar. *G3*
554 *Genes|Genomes|Genetics* 8:3961–3972. <https://doi.org/10.1534/g3.118.200760>

555 Ritchie MD, Holzinger ER, Li R, et al (2015) Methods of integrating data to uncover genotype–phenotype
556 interactions. *Nat Rev Genet* 16:85–97. <https://doi.org/10.1038/nrg3868>

557 Robert P, Auzanneau J, Goudemand E, et al (2022) Phenomic selection in wheat breeding: identification and
558 optimisation of factors influencing prediction accuracy and comparison to genomic selection. *Theor*
559 *Appl Genet* 135:895–914. <https://doi.org/10.1007/s00122-021-04005-8>

560 Rooney TE, Sweeney DW, Sorrells ME (2022) Time series barley germination is predictable and associated with
561 known seed dormancy loci. *Crop Science* 62:100–119. <https://doi.org/10.1002/csc2.20638>

562 Rousseau-Gueutin M, Belser C, Da Silva C, et al (2020) Long-read assembly of the Brassica napus reference
563 genome Darmor-bzh. *GigaScience* 9:giaa137. <https://doi.org/10.1093/gigascience/giaa137>

564 Sanchez D, Sadoun SB, Mary-Huard T, et al (2023) Improving the use of plant genetic resources to sustain
565 breeding programs' efficiency. *Proceedings of the National Academy of Sciences* 120:e2205780119.
566 <https://doi.org/10.1073/pnas.2205780119>

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

569 Schopp P, Müller D, Technow F, Melchinger AE (2017) Accuracy of Genomic Prediction in Synthetic
570 Populations Depending on the Number of Parents, Relatedness, and Ancestral Linkage Disequilibrium.
571 *Genetics* 205:441–454. <https://doi.org/10.1534/genetics.116.193243>

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

575 Simmonds NW (1993) Introgression and Incorporation. Strategies for the Use of Crop Genetic Resources.
576 *Biological Reviews* 68:539–562. <https://doi.org/10.1111/j.1469-185X.1993.tb01243.x>

577 Storey JD (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*
578 (Statistical Methodology) 64:479–498. <https://doi.org/10.1111/1467-9868.00346>

579 Voss-Fels KP, Cooper M, Hayes BJ (2019) Accelerating crop genetic gains with genomic selection. *Theor Appl*
580 *Genet* 132:669–686. <https://doi.org/10.1007/s00122-018-3270-8>

- Wagner M-H, Demilly D, Ducournau S, Durr C (2011) Computer vision for monitoring seed germination from dry state to young seedlings. *Seed Testing International* 49–51
- Weiß TM, Zhu X, Leiser WL, et al (2022) Unraveling the potential of phenomic selection within and among diverse breeding material of maize (*Zea mays* L.). *G3 Genes|Genomes|Genetics* 12:jkab445. <https://doi.org/10.1093/g3journal/jkab445>
- Werner CR, Qian L, Voss-Fels KP, et al (2018) Genome-wide regression models considering general and specific combining ability predict hybrid performance in oilseed rape with similar accuracy regardless of trait architecture. *Theor Appl Genet* 131:299–317. <https://doi.org/10.1007/s00122-017-3002-5>
- Werner CR, Gaynor RC, Gorjanc G, et al (2020) How Population Structure Impacts Genomic Selection Accuracy in Cross-Validation: Implications for Practical Breeding. *Frontiers in Plant Science* 11:592977. <https://doi.org/10.3389/fpls.2020.592977>
- 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>
- Würschum T, Abel S, Zhao Y (2014) Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. *Plant Breeding* 133:45–51. <https://doi.org/10.1111/pbr.12137>
- Yu J, Pressoir G, Briggs WH, et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208. <https://doi.org/10.1038/ng1702>
- Zhu X, Leiser WL, Hahn V, Würschum T (2021) Phenomic selection is competitive with genomic selection for breeding of complex traits. *The Plant Phenome Journal* 4:e20027. <https://doi.org/10.1002/ppj2.20027>
- Zhu X, Maurer HP, Jenz M, et al (2022) The performance of phenomic selection depends on the genetic architecture of the target trait. *Theor Appl Genet* 135:653–665. <https://doi.org/10.1007/s00122-021-03997-7>

Supplementary Information

The online version contains supplementary material

Data availability

The datasets generated and analyzed in this study are available using the following link during the review process (<https://entrepot.recherche.data.gouv.fr/privateurl.xhtml?token=4050df7f-5caf-4683-8142-df8b870ef8d5>) and will be freely available with a DOI if the article is accepted.

Acknowledgements

The authors would like to thank Renaud Rincet (GQE Le Moulon, INRAE) and Christopher Sauvage (Syngenta) for many helpful discussions regarding genetic analyses, genomic and phenomic predictions; Mathieu Rousseau-Gueutin and Cyril Falentin (IGEPP, INRAE) for help with the genomic resources of *Brassica napus*. The authors would also like to thank the CRB BrACySol for providing genotypes and Elise Alix, Bernard Moulin (IGEPP, INRAE) and the “Domaine de la Motte” Experimental Unit (INRAE) for seed multiplication. A special thanks to

Solenn Guichard for NIR spectra acquisition. The GenOuest bioinformatics platform (<https://www.genouest.org/>) are acknowledged for providing informatics resources.

Funding

This work was supported by the NAVIG project funded by Promosol association. The PhD grant of M.L. was provided by INRAE and Promosol. The RAPSODYN (ANR-11-BTBR-0004) project founded the acquisition of the genomic data.

Contributions

A.L. and N.N. designed the study and obtained funding. M-H.W., D.D., M.L. and C.B. and S.D. carried out the germination assays on the PHENOTIC platform at GEVES (Angers, France). C.B. and S.R. brought expertise with genotyping and NIRS data respectively. M.L. and J.L. performed all the statistical data, developed the scripts and ran the genetic analyses. M.L., J.L., A.L. and N.N. interpreted the results. M.L, A.L and N.N wrote the manuscript. All authors read and approved the final version.

Conflict of interest

The authors declare that they have no competing interest.

636 **Table 1** Variation of the germination-related traits assessed over a diversity germplasm of 223 rapeseed genotypes.

Trait	Abb (unit)	Formula	Min	Max	Mean	SD	H ²
Thousand seed weight	TSW (g)		2.550	6.250	4.358	0.623293	/
Seed volume increase after 20h of imbibition	VI (%)	$\text{Volume}_{20\text{h}} - \text{Volume}_{0\text{h}}$	0.997	4.440	2.145	0.496137	0.40
Time of the 1st germ.	FG (h)	$\frac{\sum h_{1\text{st germination}}}{n\text{rep}}$	22.00	58.00	30.25	4.805132	0.51
Time to 20% of germ.	T20 (h)		28.50	64.00	52.26	7.252935	0.69
Time to 50% of germ.	T50 (h)		36.00	79.67	51.93	8.855827	0.65
Time to 80% of germ.	T80 (h)		43.20	96.00	63.55	11.15751	0.61
Mean Germ. Time	MGT (h)		38.12	75.43	52.26	6.849242	0.68
Uniformity	UNI (h)	T80-T20	11.53	50.00	24.04	7.472102	0.43
Area under the curve	AUC (%.h ⁻¹)	$\int_{h_2}^{h_1} f_h d_h$	1444	5888	4077	972.7179	/
Germ. percentage at 36h	GP36 (%)	$\frac{nb_{\text{germinated seeds at 36h}}}{nb_{\text{total}}} * 100$	0.00	52.00	15.49	10.35249	0.69
Germ. percentage at 96h	GP (%)	$\frac{nb_{\text{germinated seeds at 96h}}}{nb_{\text{total}}} * 100$	46.67	100	49.46	10.91978	0.46
Radicle elongation speed	ES (mm.h ⁻¹)	$\frac{l_{\text{radicule at 16h}}}{16\text{h}}$	0.080	0.209	0.149	0.026788	0.56

637 Abb, trait abbreviation; Min, Max, Mean and SD, minimum, maximum, mean and standard deviation for fitted
638 values; H², broad-sense heritability.

639

640 **Table 2** Correlation coefficients for trait pairs associated with seed germination scored on 223 rapeseed genotypes.

	AUC	TSW	MGT	T20	T50	T80	UNI	GP	GP36	FG	VI
TSW	0.08										
MGT	-0.97***	-0.06									
T20	-0.94***	-0.04	0.96***								
T50	-0.96***	-0.05	0.96***	0.92***							
T80	-0.91***	-0.12	0.91***	0.84***	0.88***						
UNI	-0.73***	-0.19**	0.67***	0.53***	0.67***	0.88***					
GP	0.89***	0.10	-0.77***	-0.76***	-0.80***	-0.78***	-0.70***				
GP36	0.89***	0.07	-0.93***	-0.95***	-0.88***	-0.81***	-0.50***	0.68***			
FG	-0.61***	-0.08	0.6***	0.63***	0.55***	0.57***	0.40***	-0.55***	-0.62***		
VI	0.27***	0.52***	-0.29***	-0.29***	-0.27***	-0.23***	-0.18**	0.18**	0.28***	-0.11	
ES	0.35***	0.19*	-0.43***	-0.40***	-0.37***	-0.39***	-0.25***	0.20**	0.42***	-0.22**	0.14*

641 *Significant values are shown in boldface. Significances were calculated at levels of: * $p < 0.05$, ** $p < 0.01$ and*

642 **** $p < 0.001$.*

643 *For details of trait abbreviations refer to Table 1 and “Material and methods”.*

644

Table 3 QTL controlling germination-related traits detected in a population of 223 rapeseed through GWAS. QTL were ordered regarding their position on the genome. For each QTL, information is given about the peak and flanking markers, the peak SNP position, QTL size as well as major and minor allele, followed by the minor allele frequency (MAF). For each trait the SNP weight indicating the favorable allele version and the phenotypic variance explained by the SNP (R^2) are represented. Positions are relative to the reference genome Darmor-*bzh* V10 (Rousseau-Gueutin et al. 2020).

QTL Name	Chr	Trait	Peak SNP Peak Position (pb)	Flanking marker 1/ Flanking marker 2	QTL size (kb)	Major /Minor allele	MAF	SNP weight	R^2
QTL.A06.1	A06	GP36						2.60	7.60
		MGT	Bn_A06_p836912	Bn_A06_p709547/	279	T/C	0.15	-1.83	8.46
		T20	(1102221)	Bn_A06_p908271				-1.81	7.64
		T50						-2.38	8.78
QTL.A06.2	A06	MGT	Bn_A06_p14121814	Bn_A06_p19574863/	4272	T/C	0.20	1.78	7.99
		T50	(34532655)	Bn_A06_p18040423				2.25	7.78
QTL.A07.1	A07	GP36	Bn_A07_p15220175 (21453055)	Bn_A07_p14390667/ Bn_A07_p15885711	1530	A/G	0.25	2.91	8.59
QTL.A07.2	A07	T50	Bn_A07_p16553147 (22869578)	Bn_A07_p16157761/ Bn_A07_p17026444	909	G/T	0.13	2.19	7.62
QTL.A09.1	A09	GP36						-4.00	8.31
		MGT	Bn_A09_p2733282	Bn_A09_p951202/	8824	G/A	0.45	2.27	6.48
		T20	(4101875)	Bn_A09_p7560188				2.43	6.95
		T50						3.00	8.03
QTL.A10.1	A10	MGT	Bn_A10_p9459037 (13661179)	Bn_A10_p7835472/ Bn_A10_p9645870	1990	C/T	0.27	-1.88	6.60
QTL.A10.2	A10	TSW	Bn_A10_p15838932 (18901950)	Bn_A10_p13817274/ Bn_A10_p16933405	3226	G/T	0.28	0.23	7.44
QTL.C01.1	C01	GP36	Bn_scaff_15838_5_p446059	Bn_scaff_19244_1_p388798/	6736	T/G	0.45	3.07	7.47
		MGT	(3600398)	Bn_scaff_17731_1_p308944				-1.91	6.73
QTL.C02.1	C02	GP36	Bn_scaff_18507_1_p957588 (35173323)	Bn_scaff_18507_1_p737065/ Bn_scaff_17067_1_p669413	6469	A/G	0.13	2.66	6.99
QTL.C02.2	C02	GP36						-3.13	7.31
		MGT	Bn_scaff_16139_1_p720716 (61523315)	Bn_scaff_15918_1_p303294/ Bn_scaff_16139_1_p41217	5361	T/C	0.32	2.15	8.30
		T20						2.16	7.75

		T50						2.47	7.04
QTL.C03.1	C03	VI	Bn_scaff_17440_1_p552871 (57791809)	Bn_scaff_19740_1_p203934/ Bn_scaff_16394_1_p2229812	5070	T/C	0.39	0.21	6.50
QTL.C04.1	C04	MGT	Bn_scaff_18440_1_p77767 (9746347)	Bn_scaff_27469_1_p45054/_sc aff_16576_1_p202308	12852	T/C	0.24	2.53	6.69
QTL.C04.2	C04	MGT	Bn_scaff_24979_1_p13103 (53828390)	Bn_scaff_19208_1_p476848/ Bn_scaff_20079_1_p289504	5714	T/C	0.36	1.88	7.05
		T20						1.93	6.73
QTL.C05.1	C05	MGT	Bn_scaff_16414_1_p1774629 (254723)	Bn_scaff_23107_1_p7608/ Bn_scaff_15712_10_p173981	5460	C/T	0.38	-2.01	6.30
QTL.C07.1	C07	GP36	Bn_A07_p3467400 (15046703)	Bn_A07_p3467400/ Bn_scaff_16721_1_p1551224	28	C/G	0.07	3.01	8.69
QTL.C07.2	C07	T50	Bn_scaff_16110_1_p1076892 (54582081)	Bn_scaff_16110_1_p2439331/ Bn_scaff_16110_1_p1076892	1509	A/G	0.07	2.22	7.91
QTL.C09.1	C09	VI	Bn_A09_p694190 (579848)	Bn_scaff_16486_1_p88236/ Bn_scaff_17297_1_p202207	6571	A/T	0.46	-0.16	6.84

SNP weight in boldface stand for minor alleles that are favorable.

651

652

653

Table 4 Average of the True Breeding Values (TBV) for the whole germplasm, and the top 10% individuals selected based on their TBV, Genomic Estimated Breeding Values (GEBV) and Hyperspectral Estimated Breeding Values (PEPV), for each trait, using 210 rapeseed genotypes. For each average, the standard deviation is associated. The selection differential (S) was calculated for each BV, as the difference between the germplasm mean and the mean of the top 10% genotypes.

	Mean whole population	Top 10% TBV	S TBV	Top 10% GEBV	S GEBV	Top 10% PEPV	S PEPV
GP36	1.07 ±0.36	1.55 ±0.07	0.48	1.33 ±0.21	0.26	1.33 ±0.22	0.26
T50	51.85 ±8.87	40.24 ±1.77	-11.61	48.55 ±4.83	-3.3	45.33 ±5.06	-6.52
T20	41.23 ±7.20	31.68 ±1.30	-9.55	36.96 ±4.57	-4.28	35.80 ±4.51	-5.43
T80	63.48 ±11.13	48.33 ±1.97	-15.15	58.89 ±6.56	-4.59	55.19 ±5.60	-8.29
AUC	40.84 ±11.79	60.14 ±2.43	19.3	53.08 ±5.50	12.24	53.55 ±7.71	12.71
FG	30.16 ±4.75	24.29 ±0.85	-5.87	27.95 ±3.27	-2.21	27.26 ±2.39	-2.90
GP	2.53 ±1.16	4.40 ±0.33	1.87	2.81 ±0.74	0.28	3.17 ±0.84	0.64
MGT	52.22 ±6.87	42.43 ±1.62	-9.79	49.21 ±4.97	-3.01	47.12 ±4.03	-5.10
UNI	24.02 ±7.50	14.49 ±1.16	-9.53	20.83 ±4.03	-3.19	20.03 ±4.20	-3.99

Figure legends

Fig.1 Plot of principal components for 12 variables associated with seed germination scored on 223 genotypes. Trait description is provided in Table 1.

Fig. 2 Genetic and phenotypic characterization of the panel used. Relatedness of 223 genotypes of rapeseed is highlighted by a principal coordinate analysis (PCoA) using the first two components using (a) SNP, (b) NIRS data acquired on seed samples. (c) PAM clustering on principal components of 12 variables associated with seed germination measured on 223 genotypes

Fig. 3 Predictive abilities of the nine seed germination-related traits using genomic (GBLUP), phenomic (HBLUP), or combined (GHBLUP) prediction models run for scenarios CV_{whole} , CV_{local} and $CV_{whole/local}$. Scenario CV_{whole} correspond to the utilization of the whole germplasm for calibration and estimation of the PA, scenario CV_{local} correspond to cluster 1 genotypes used for calibration and estimation of the PA. Scenario $CV_{whole/local}$ correspond to the utilization of the whole germplasm for calibration and only genotypes of the cluster 1 were used for PA estimation. PA were obtained from a fivefold cross-validation with 100 repetitions. Each boxplot indicates the mean (bold line), the first and third quartiles (boxes), the first and ninth deciles (whiskers).

Fig. 4 Predictive ability for GP36 using (a) genomic (GBLUP) and (b) phenomic (HBLUP) prediction models. Dots represents the observed and the predicted trait values across a fivefold cross-validation with 10 repetitions. Colors and dot types represents genomic clusters previously defined. The mean predictive ability (r) calculated across the entire dataset using a fivefold cross-validation with 100 repetitions is represented by a black line.

Figures

Laurençon et al. Fig. 1

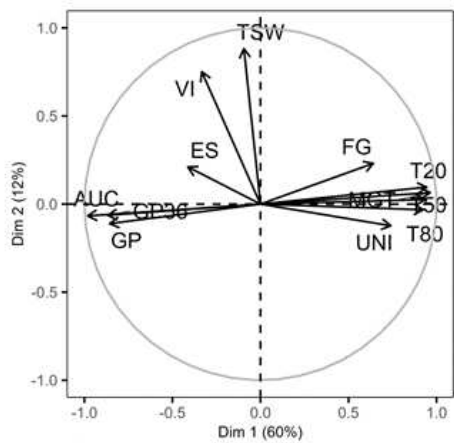


Figure 1

Plot of principal components for 12 variables associated with seed germination scored on 223 genotypes. Trait description is provided in Table 1.

Laurençon
et al. Fig. 2

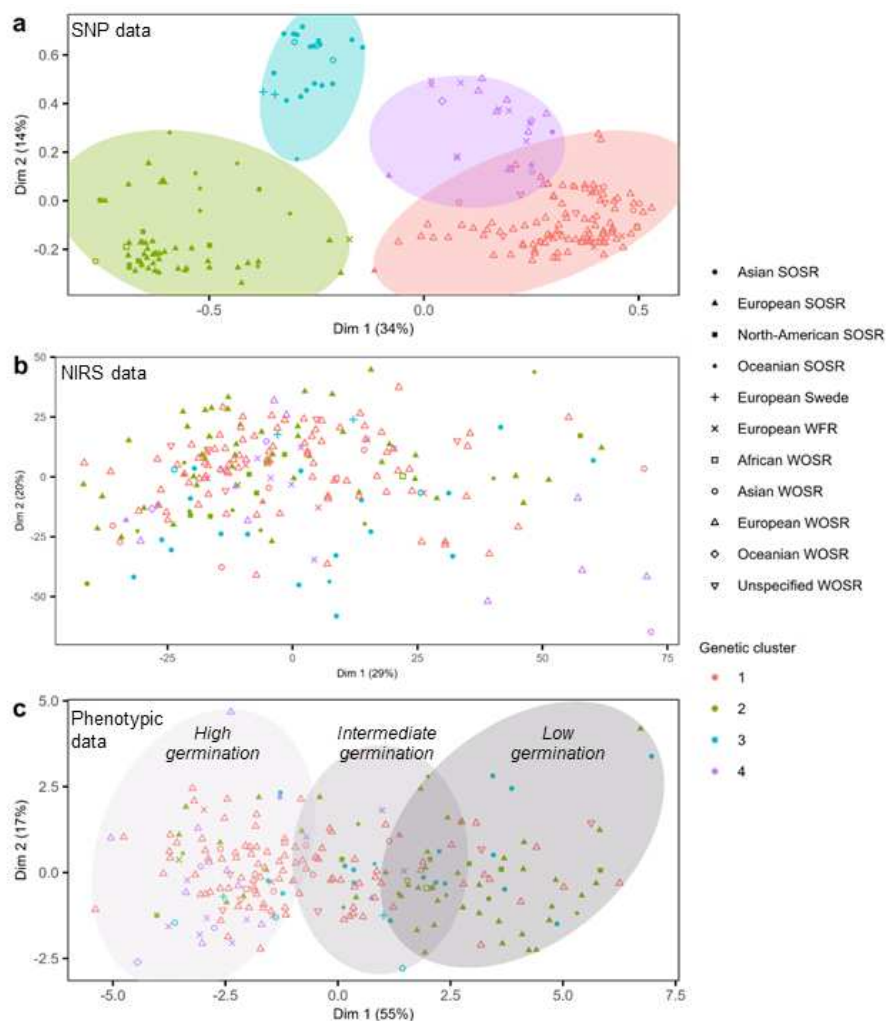


Figure 2

Genetic and phenotypic characterization of the panel used. Relatedness of 223 genotypes of rapeseed is highlighted by a principal coordinate analysis (PCoA) using the first two components using (a) SNP, (b) NIRS data acquired on seed samples. (c) PAM clustering on principal components of 12 variables associated with seed germination measured on 223 genotypes

Laurençon et al. Fig. 3

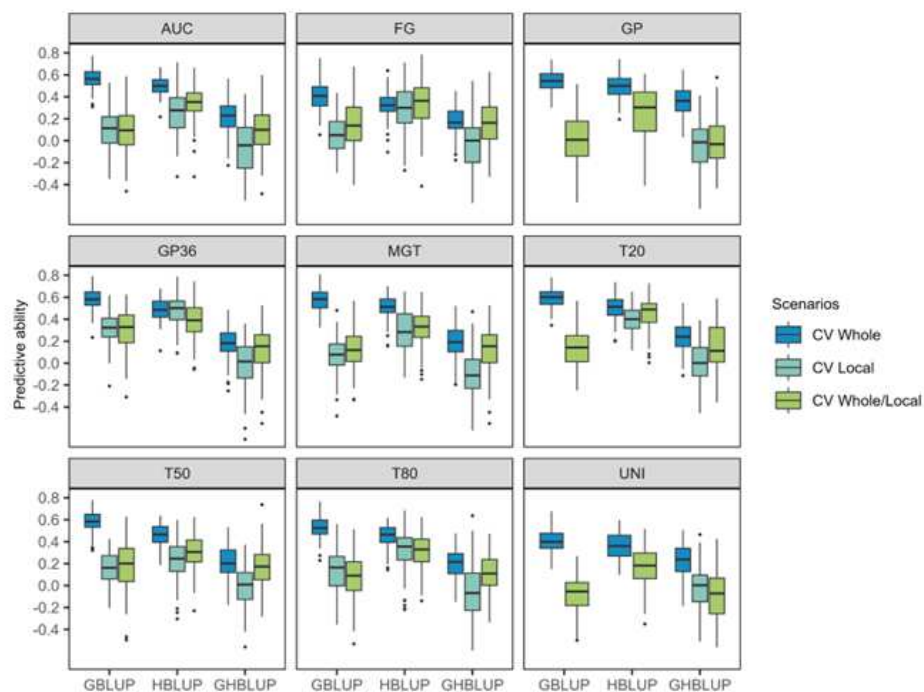


Figure 3

Predictive abilities of the nine seed germination-related traits using genomic (GBLUP), phenomic (HBLUP), or combined (GHBLUP) prediction models run for scenarios CV_{whole} , CV_{local} and $CV_{\text{whole/local}}$. Scenario CV_{whole} correspond to the utilization of the whole germplasm for calibration and estimation of the PA, scenario CV_{local} correspond to cluster 1 genotypes used for calibration and estimation of the PA. Scenario $CV_{\text{whole/local}}$ correspond to the utilization of the whole germplasm for calibration and only genotypes of the cluster 1 were used for PA estimation. PA were obtained from a fivefold cross-validation with 100 repetitions. Each boxplot indicates the mean (bold line), the first and third quartiles (boxes), the first and ninth deciles (whiskers).

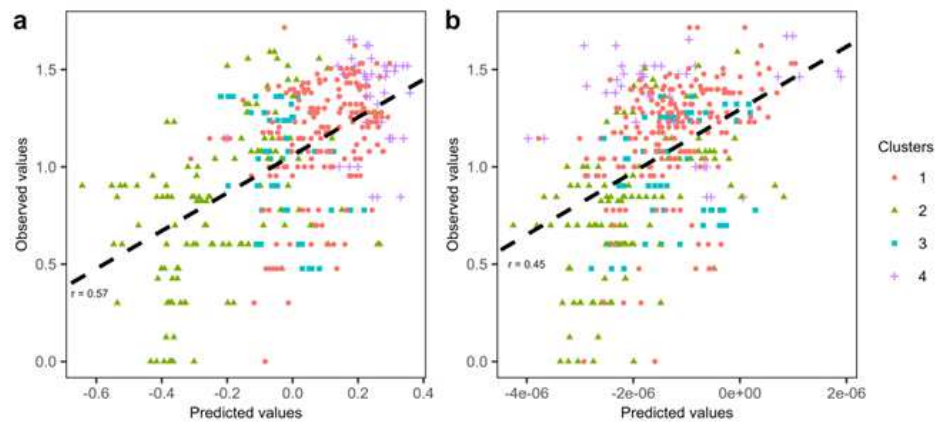


Figure 4

Predictive ability for GP36 using (a) genomic (GBLUP) and (b) phenomic (HBLUP) prediction models. Dots represents the observed and the predicted trait values across a fivefold cross-validation with 10 repetitions. Colors and dot types represents genomic clusters previously defined. The mean predictive ability (r) calculated across the entire dataset using a fivefold cross-validation with 100 repetitions is represented by a black line.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [laurenconetalsupp.xlsx](#)