Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Near-infrared spectroscopy-based phenomic prediction is competitive with genomic prediction in sorghum

Pradip Sapkota

prdp@tamu.edu

Texas A&M University

Fabian Leon

University of Kentucky

Daniel Crozier

Texas A&M University

William L. Rooney

Texas A&M University

Article

Keywords: Genomic selection, multi-environment, near-infrared spectroscopy, phenomic selection, sorghum

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11233033/v1

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

Additional Declarations: No competing interests reported.

Near-infrared spectroscopy-based phenomic prediction is competitive with genomic prediction in sorghum

Pradip Sapkota[1*], Fabian Leon[2], Daniel Crozier[1], William L. Rooney[1]

Affiliations:

[1*]Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843-2474, USA

[2]Department of Plant and Soil Sciences, University of Kentucky, Lexington, KY 40546-0312, USA

Correspondence

Pradip Sapkota, Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843-2474, USA. Email: prdp@tamu.edu ORCID: https://orcid.org/0000-0002-7430-1370

Present Address

Fabian Leon, Department of Plant and Soil Sciences, University of Kentucky, Lexington, KY 40546-0312, USA

ABSTRACT

Genomic selection is well established in many crop improvement programs, and the inclusion of phenomic data to prediction models is increasing. Applications of phenomic data for predictive methods lag because genotypes must be grown whereas genomic prediction is applied to untested genotypes. To address this drawback, we propose a low-cost approach of phenomic selection strategies using Near Infrared Spectroscopy (NIRS) data collected on inbred parental lines. For this study, 366 grain sorghum hybrids, developed from a partial factorial, were evaluated across four total environments between 2019 and 2020. The inbred parents of these hybrids were evaluated in the 2020 environments. Both hybrids and inbreds were phenotyped for grain yield, days to anthesis, and plant height and grain samples were scanned using NIRS bandwidths from 400 – 2500 nm to estimate compositional traits such as starch, protein, fat, fiber, and ash. Parents were genotyped using genotyping by sequencing which produced 38,234 high-quality SNP markers. Various prediction models incorporating these genomic, phenomic and both data were trained. Phenomic prediction models trained using NIRS obtained on parents (**M2**) and mid-parent heterosis (**M3**) of NIRS were similar, and in most cases competitive to genomic prediction (**M1**) in predicting hybrids from un-testcrossed lines (**CVa**) and untested hybrids (**CVb**). Linear discriminant analysis using NIR spectra of grain samples explained majority of variance and was able to characterize environments and heterotic groups accurately. NIRS-based phenomic prediction could be implemented as low-cost alternative to genomic prediction to predict sorghum hybrid performance across environments.

Keywords: Genomic selection, multi-environment, near-infrared spectroscopy, phenomic selection, sorghum

AUTHOR CONTRIBUTION STATEMENT

PS: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Visualization; Project administration, Software, Supervision, Validation, Visualization, Writing – original draft; Writing – review & editing. **FL:** Data curation; Investigation, Project administration, Supervision, Writing – review & editing. **DC:** Data curation; Investigation, Project administration, Supervision, Writing – review & editing. **WLR:** Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Writing- review & editing.

Key message **Phenomic prediction can accurately predict hybrid performance across environments in sorghum.**

ACKNOWLEDGMENTS

The authors thank Dr. Robert R. Klein, Dr. Patricia E. Klein and Julie McCollum for their assistance in genotyping of parents. The authors also thank all members of the sorghum improvement program at Texas A&M AgriLife Research for their assistance.

INTRODUCTION

Genomic selection (GS) approaches have dominated the field of predictive strategies in plant breeding programs for over a decade (Alemu et al. 2024). Further improvements in molecular genetics combined with novel phenomic techniques now generate a plethora of information at the level of the genome, phenome, transcriptome, and metabolome (Schrag et al. 2018). Although multi-omics information improves the understanding of biological processes, direct application in plant breeding programs is limited (Yan & Wang, 2023), which could be attributed to their variable nature and high-cost relative to the benefit to the plant breeding program (Holland, 2004). If proven efficient, phenomic data could supplement GS within a breeding program (Robert et al. 2022b). One such type of phenomic data, Near Infrared Spectroscopy (NIRS), is commonly used in plant breeding programs and is relatively easy and cheap to implement (Rincent et al. 2018).

The NIRS technology for compositional analysis is effective because it discerns organic compounds consistently (Pasquini, 2003). These organic compounds – also known as endophenotypes – are associated with standard chemical compositional traits (Houle et al. 2010). If heritable, these compositional traits could effectively capture relationships among genotypes (Rincent et al. 2025). These NIRS spectra are also influenced by polygenic effects, environment, and genotype × environment interactions (Robert et al. 2022c). NIRS-based phenomic prediction is effective in a range of crops, including coffee, corn, rye, soybean, wheat, poplar, sorghum (Adunola et al. 2024; DeSalvio et al. 2024; Galán et al. 2020; Parmley et al. 2019; Rincent et al. 2018; Sapkota et al. 2025a). To extend NIRS application, spectra obtained from parents in related environments could predict the next generation of breeding cross. Although prediction using phenomic data of parents is a relatively new concept, this approach is effective in rapeseed and corn (Rincent et al. 2025; Roscher-Ehrig et al. 2024) and warrants further investigation into small grain crops. If consistent, it could supplement GS or reduce genotyping requirements (Rincent et al. 2018).

Phenomic prediction is more effective for complex traits and less effective for traits controlled by major QTL (Zhu et al. 2022). Such models predicted agronomic performance better than GS in a population with minimal relationships among individuals and are less influenced by population structure, environment and G × E effects

(Roscher-Ehrig et al. 2024; Zhu et al. 2021). Furthermore, phenomic prediction models were also less influenced by training-testing sizes, heritability of traits, genetic relatedness and epistatic interactions (Robert et al. 2022c).

Adoption of GS can be challenging for smaller breeding programs because the initial adoption requires logistical restructuring followed by regular updates to GS models and training sets (Rincent et al. 2018). In the Texas A&M AgriLife Research sorghum breeding program, we have found that NIRS-based phenomic models were competitive to genomic prediction in predicting sorghum hybrids (Sapkota et al. 2025a). Although PS methods have merit, adoption is limited by its phenotyping requirement. Therefore, phenomic prediction methodologies must provide separate advantages if they are to be widely adopted.

In hybrid crops, breeding programs regularly evaluate the inbred lines wherein grain samples can be scanned for spectral data acquisition with minimal additional effort. This phenomic data is available to enhance selection and breeding decisions which would save time, resources and effort in the program.

In this study, genomic, phenomic and integrated prediction models were trained. The hypothesis is that NIR bands from parents could capture the heritable variation and are useful to predict hybrid performance. The objectives of this study were to 1) calculate the source of variation and heritability of phenotypic performance and NIR spectra, 2) predict hybrid performance using NIR spectral data as well as genomic data, 3) characterize environments and genotypes using NIR spectra, and finally, 4) estimate correlations between mid-parent and hybrid phenotypes and NIR bands.

MATERIALS AND METHODS

EXPERIMENTAL DESIGN

A set of 45 A-lines and 44 R-lines representing the diversity of the inbred lines used in the AgriLife Research Sorghum Breeding program were crossed in a partial factorial design to generate 366 hybrids. A total of 364 hybrids were evaluated in 2019 and 2020 in College Station, Texas (19CS and 20CS) with 362 hybrids common to both environments. Additionally, subsets of these hybrids were evaluated at two other environments: 179 hybrids in 2019

in Taft, Texas (19TA) and 187 hybrids in 2020 in Lyford, Texas (20LY). The hybrids in 19TA and 20LY were distinct except one hybrid (A11076/RTx2918) that was grown in both of those environments. Trial sets were developed in such a way that parents were equally represented in all environments. Adjacent to these hybrid trials, all 89 inbred parents were also evaluated in 20CS, and 20LY. All inbred and hybrid trials used a Randomized Complete Block Design (RCBD) replicated twice in each environment. Details on germplasm development along with field experiments are further outlined by Crozier et al. (2023b).

AGRONOMIC TRAITS

Data on three agronomic traits were collected in each environment: days to mid-anthesis (DA), plant height (PH), and grain yield (GY). DA was recorded as number of days from planting to 50% mid-anthesis. Plant height was measured on a representative plant in each plot from soil to the tip of panicle. GY was measured using a John Deere plot combine fitted with a Harvest Master GrainGage System (Juniper System, Logan, UT), and GY were calculated in kg ha^{-1} with adjustment to 14% moisture content.

GRAIN SAMPLING AND SCANNING

Prior to plot harvest, five sorghum panicles were sampled from one replication in 2019, and from both replications in 2020. The panicles from each plot were bulk threshed using an Almaco rubber-belt thresher BT14E (ALMACO). After equilibrating moisture content of the samples to 12%, a 100-gram whole grain sample from each plot was scanned twice for 2019 samples and once for 2020 samples using a FOSS XDS NIRDS 2500. NIR reflectance values were obtained at wavelengths of 400 nm to 2500 nm with intervals of 0.5 nm. In total, about 4200 reflectance values for NIR bands were collected for each sample by ISI scan Nova software. Using NIR calibration curves developed within Texas A&M AgriLife Research Sorghum Breeding Program, estimates on starch, protein, fiber, fat, ash, and moisture concentration (percent basis) were obtained.

PHENOMIC DATA: NEAR INFRARED SPECTROSCOPY

The raw NIR spectra was collected for all hybrids and inbreds from all environments. Reflectance values of these spectra collected on parents (20CS and 20LY) were fitted using mixed effect model of **equation 1** within environments.

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 g} + \boldsymbol{Z_2 r} + \boldsymbol{e} \quad (1)$$

Where y is the vector NIR reflectance values (response variable), **μ** is an intercept, **g** is a random effect of genotype **g ~ N (0, σ^2_gI)**; **r** is a vector of replications **r ~ N (0, σ^2_rI)**, and **e** is the vector of residuals, **e ~ N (0, σ^2_eI)**; **1** is a vector of ones; $\boldsymbol{Z_1}$, and $\boldsymbol{Z_2}$ are incidence matrices; σ^2_g, σ^2_r and σ^2_e are variance components for genotypes, replicates and residuals, respectively. Variance components extracted for reflectance values obtained on all bands were used to assess the quality of data using formula of heritability on an entry mean basis, $H^2 = \frac{\sigma_h^2}{\sigma_h^2 + \frac{\sigma_e^2}{r}}$.

Best Linear Unbiased Estimates (BLUEs) were calculated using reflectance values for all genotypes (hybrids and inbreds) across all wavelengths, considering hybrids and inbreds (g) as a fixed effect within each environment using **equation 2** where **r** is a vector of replications **r ~ N (0, σ^2_rI)**, and **e** is the vector of residuals, **e ~ N (0, σ^2_eI).**

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 g} + \boldsymbol{Z_2 r} + \boldsymbol{e} \quad (2)$$

In **equation 2**, y is vector of reflectance values, **μ** is a vector with the intercept, **g** is a vector of fixed genotype effects, **r** is a vector of replication effects and **e** is the vector of residuals. $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ represent incidence matrices relating reflectance values to genotypes and replication, respectively. Finally, these BLUEs of reflectance values for parents have dimensions of 89 × 4200 in a single environment. For hybrids, the BLUEs have dimensions of 364 × 4200, 179 × 4200, 364 × 4200, 187 × 4200 for environments 19CS, 19TA, 20CS, and 20LY, respectively.

GENOTYPING

Inbred parents were genotyped using Genotyping by Sequencing (GBS) following the protocol outlined by Morishige et al. (2013). DNA extraction and quality control for parents were performed to prepare the library following standard protocol. All the steps of mapping and variant calling of SNP were performed as explained in Crozier et al. (2023).

These approaches resulted in 38,234 high-quality SNPs which were then recoded in the format of 012, where "0" represents homozygous minor allele, '1' represents heterozygous allele, and "2" represents homozygous major allele.

PHENOTYPIC DATA ANALYSIS

Variance components were extracted by fitting mixed model for hybrid phenotypic data as presented in **equation 3**.

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 f} + \boldsymbol{Z_2 m} + \boldsymbol{Z_3 h} + \boldsymbol{Z_4 s} + \boldsymbol{Z_5 fs} + \boldsymbol{Z_6 ms} + \boldsymbol{Z_7 fms} + \boldsymbol{Z_8 r}\,(\boldsymbol{s}) + \boldsymbol{e} \qquad 3$$

In **equation 3**, **y** is vector of phenotypes; **μ** is an intercept; **f** represents random female effects with distribution of $\mathbf{f} \sim \mathbf{N(0, \sigma^2_f I)}$; **m** represents random male effects, $\mathbf{m} \sim \mathbf{N(0, \sigma^2_m I)}$; **h** represents random effect of hybrid, $\mathbf{h} \sim \mathbf{N(0, \sigma^2_{hyb} I)}$; **s** is a vector of environmental effects, $\mathbf{s} \sim \mathbf{N(0, \sigma^2_s I)}$; **fs** is a vector of interaction effect between GCA effect of females and environment, $\mathbf{fs} \sim \mathbf{N(0, \sigma^2_{fs} I)}$; **ms** is a vector of interaction effect between GCA effect of males and environment, $\mathbf{ms} \sim \mathbf{N(0, \sigma^2_{ms} I)}$; **fms** is a vector of interaction effect between SCA effect of hybrids and environment, $\mathbf{fms} \sim \mathbf{N(0, \sigma^2_{fms} I)}$; **r(s)** is a vector of effects of replication nested within environment, $\mathbf{r(s)} \sim \mathbf{N(0, \sigma^2_{r(s)} I)}$; **e** is the vector of residuals, $\mathbf{e} \sim \mathbf{N\,(0, \sigma^2_e I)}$; and $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$, $\boldsymbol{Z_3}$, $\boldsymbol{Z_4}$, $\boldsymbol{Z_5}$, $\boldsymbol{Z_6}$, $\boldsymbol{Z_7}$, and $\boldsymbol{Z_8}$ are incidence matrices for female, male, hybrid, environment, female × environment, male × environment, female × male × environment, and replication nested within environment effects, respectively.

Because grain composition traits had unequal numbers of replications, plot-basis heritability was estimated for all traits to ensure consistent comparisons. Extracted variance component estimates were used to compute broad-sense heritability and narrow-sense heritability on a plot basis across environments: $H^2 = \frac{\sigma^{2\,*}_{hyb}}{\sigma^{2\,*}_{hyb} + \sigma^2_{hybE} + \sigma^2_e}$ and $h^2 = \frac{\sigma^{2\,*}_a}{\sigma^{2\,*}_{hyb} + \sigma^2_{hybE} + \sigma^2_e}$. In these equations, $\sigma^{2\,*}_{hyb}$ stands for the variance component due to hybrid genotypes which is the summation of variance components due to female effects, male effects and interaction effects of female × male. $\sigma^{2\,*}_a$ represents additive genetic variance estimated by summing variance due to female and male assuming no correlation between female and male effects as well as lack of dominance, epistasis, maternal effects. σ^2_{hybE} represents variance computed for genotype × environment interaction. Residuals obtained from fitting **model 3** were used to extract the

Coefficient of Variation (CV_e) with a formula of $CV_e = \frac{\sqrt{\sigma_e^2}}{\bar{x}} \times 100\ \%$, where σ_e^2 represents residual variance extracted from the model and $\bar{x}$ is the grand mean for that particular trait.

GENOMIC AND PHENOMIC PREDICTION MODELS

The numerical formats of allele (i.e. 0,1,2) were used to calculate genomic relationship matrix using '*vanraden*' method in R (Vanraden, 2008). The genomic relationship matrix was calculated by $G = \frac{XX'}{n}$, where X is a molecular marker centered around zero, X' is transpose of X, n = number of molecular markers. These genomic relationship matrices (GRM) modeled genetic covariance of females, males, and hybrids, via the Kronecker product of females and males, per the traditional GBLUP model.

Model 1 (M1): Main genetic effects, Environment, plus Genetic × Environment effects

Model 1 was fitted by modeling effects of female, male, hybrids (male × female term), environmental effects, interaction between female and environment, interaction between male and environment and interaction between hybrid and environment;

$$\mathbf{M1}: \mathbf{y} = \mathbf{1}\mu + \boldsymbol{Z_1}\mathbf{f} + \boldsymbol{Z_2}\mathbf{m} + \boldsymbol{Z_3}\mathbf{h} + \boldsymbol{Z_4}\mathbf{s} + \boldsymbol{Z_5}\mathbf{fs} + \boldsymbol{Z_6}\mathbf{ms} + \boldsymbol{Z_7}\mathbf{fms} + \mathbf{e}$$

Where **y** is a vector of phenotypes, **f** is the vector of GCA effects of female, $\mathbf{f} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma_{fj}^2 G_f})$, where **j** represents environments; **m** represents vector of GCA effects of male, $\mathbf{m} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma_{mj}^2 G_m})$; **h** represents vector of SCA effects of hybrid combinations, $\mathbf{h} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma_{hj}^2}\mathbf{H})$; **s** represents vector of the environmental effects, $\mathbf{s} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma_s^2 I})$; **fs** represents a random vector that include the interaction between GCA of females and environment, $\mathbf{fs} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma_{fs}^2 V_f})$, $\boldsymbol{\sigma_{fs}^2}$ and $\boldsymbol{V_f}$ represents variance components and variance-covariance structures for female × environment interaction; **ms** represents a random vector that include the interaction between GCA of males and environment, $\mathbf{ms} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma_{ms}^2 V_m})$, $\boldsymbol{\sigma_{ms}^2}$ and $\boldsymbol{V_m}$ represents variance components and variance-covariance structures for male × environment interaction; **fms** represents a random vector that include the interaction between SCA of hybrids and environment, $\mathbf{fms} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma_{fms}^2 V_{fm}})$, $\boldsymbol{\sigma_{fms}^2}$ and $\boldsymbol{V_{fm}}$ represents variance components and variance-covariance

structures for female × male × environment interaction, $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^2_{ej}\mathbf{I})$, **e** is a vector of residuals. $\mathbf{G_f}$, $\mathbf{G_m}$, and **H** are the GRM for females, males, and hybrids, respectively. The elements of '**H**' were calculated by Kronecker product in silico using $\mathbf{G_f}$ and $\mathbf{G_m}$ for the available hybrid combinations. The interaction terms (G × E) were modeled by including incidence matrices for environments that relates phenotypic observation with environment following the approach of Acosta-Pech et al. (2017). Variance-covariance matrices were calculated as $\boldsymbol{V_f = Z_1 G_f Z'_1 \odot Z_S Z'_S}$, $\boldsymbol{V_m = Z_2 G_m Z'_2 \odot Z_S Z'_S}$, and $\boldsymbol{V_{fm} = Z_3 G_{fm} Z'_3 \odot Z_S Z'_S}$. **1** is a vector of ones; $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$, $\boldsymbol{Z_3}$, $\boldsymbol{Z_4}$, $\boldsymbol{Z_5}$, $\boldsymbol{Z_6}$, and $\boldsymbol{Z_7}$ are incidence matrices.

Model 2 (M2): First derivative of mid-parent reflectance of NIR ($\mathbf{NIR_{mp}}$)

Mid-parent values for NIR reflectance were calculated for each hybrid by using the formula of $\frac{(Parent\ 1 + Parent\ 2)}{2}$.

Model 2 was fitted by modeling first derivative of mid-parent reflectance values, environmental effects, interaction effects between first derivative of mid-parent reflectance and environment.

$$\mathbf{M2}: \mathbf{y} = \mathbf{1\mu} + \mathbf{NIR_{mp}} + \boldsymbol{Z_1}\mathbf{s} + \mathbf{NIR_{mp}s} + \mathbf{e}$$

where **y** is a vector of phenotypes, $\mathbf{NIR_{mp}}$ is a vector of first derivative of mid-parent reflectance values, $\mathbf{NIR_{mp}} \sim \mathbf{N}\left(\mathbf{0}, \mathbf{NIR}\boldsymbol{\sigma}^2_{\mathbf{pmp}}\right)$, **NIR** represents variance-covariance matrix using first derivative of mid-parent reflectance, and $\boldsymbol{\sigma}^2_{\mathbf{pmp}}$ represents variance components using first derivative of mid-parent reflectance; **s** represents vector of environmental effects, $\mathbf{s} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^2_{\mathbf{s}}\mathbf{I})$; $\mathbf{NIR_{mp}s}$ represents random vector that includes interaction between first derivative of mid-parent reflectance and environments, $\mathbf{NIR_{mp}s} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^2_{\mathbf{mps}}\mathbf{V_{mps}})$; $\boldsymbol{\sigma}^2_{\mathbf{mps}}$ and $\mathbf{V_{mps}}$ represents variance components and variance-covariance structures for mid-parent reflectance × environment interaction term. Variance-covariance matrices for $\mathbf{NIR_{mp}s}$ were calculated as $\mathbf{V_{mps}} = \mathbf{NIR_{mp}} \odot \mathbf{Z_S Z'_S}$. Likewise, $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^2_{ej}\mathbf{I})$, **e** is a vector of residuals.

Model 3 (M3): First derivative of mid-parent heterosis of NIR ($\mathbf{NIR_{mph}}$)

Mid-parent heterosis for NIR reflectance was estimated for each hybrid using the formula of $\frac{(Hybrid\ (F1) - Mid{-}Parent)}{Mid{-}Parent}$. **Model 3** was fitted by modeling the first derivative of mid-parent heterosis of reflectance

values, environmental effects, interaction effects between mid-parent heterosis of reflectance and environment. This model uses NIR reflectance data from hybrid grain samples and inbred grain samples to estimate final reflectance values.

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{NIR_{mph}} + \boldsymbol{Z_1}\mathbf{s} + \mathbf{NIR_{mph}}\mathbf{s} + \mathbf{e}$$

where y is a vector of phenotypes, $\mathbf{NIR_{mph}}$ is a vector of first derivative of mid-parent heterosis of reflectance, $\mathbf{NIR_{mph}} \sim \mathbf{N}(\mathbf{0}, \mathbf{NIR}\boldsymbol{\sigma}^2_{\mathbf{pmph}})$, **NIR** represents phenomic effects and $\boldsymbol{\sigma}^2_{\mathbf{pmph}}$ represents variance components using first derivative of mid-parent heterosis of reflectance. Likewise, $\mathbf{s} \sim \mathbf{N}\ (\mathbf{0}, \boldsymbol{\sigma}^2_{\mathbf{s}}\mathbf{I})$, $\mathbf{NIR_{mphs}} \sim \mathbf{N}\ (\mathbf{0}, \boldsymbol{\sigma}^2_{\mathbf{mphs}}\mathbf{V_{mphs}})$, $\boldsymbol{\sigma}^2_{\mathbf{mphs}}$ and $\mathbf{V_{mphs}}$ represent variance components and variance-covariance structures, respectively, for mid-parent heterosis of reflectance × environment interaction term. Variance-covariance matrices for $\mathbf{NIR_{mph}s}$ were calculated as $\mathbf{V_{mph}} = \mathbf{NIR_{mph}} \odot \mathbf{Z_S}\mathbf{Z}'\mathbf{s}$. Likewise, $\mathbf{e} \sim \mathbf{N}\ (\mathbf{0}, \boldsymbol{\sigma}^2_{\boldsymbol{ej}}\mathbf{I})$, **e** is a vector of residuals.

Model 4 (M4): Genomic + phenomic prediction models using parental indices

Model 4 was trained by combining **model 1** and **model 2** to integrate genomic markers with parental phenomic markers. Distributions of phenomic kernels and genomic kernels remain the same as **model 1** and **model 2**.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{NIR_{mp}} + \mathbf{NIR_{mp}}\mathbf{s} + \boldsymbol{Z_1}\mathbf{f} + \boldsymbol{Z_2}\mathbf{m} + \boldsymbol{Z_3}\mathbf{h} + \boldsymbol{Z_4}\mathbf{s} + \boldsymbol{Z_5}\mathbf{fs} + \boldsymbol{Z_6}\mathbf{ms} + \boldsymbol{Z_7}\mathbf{fms} + \mathbf{e}$$

CROSS VALIDATION SCHEMES

Two different cross-validation (**CV**) schemes were fitted for all prediction models. **CVa** predicts performance of hybrids that were not testcrossed (new inbred lines). In this case, hybrids derived from 15 parents (17%) were removed from all environments. **CVb** predicts new hybrids that have not been tested in any of the environments. In this case, the performance of 110 hybrids (30%) was predicted based on phenotypic data from the remaining 70% of hybrids for all environments. Due to availability of phenomic data from parents from two environments (20CS and 20LY), four different models were trained for these two combinations. The 20CS and 20LY model use phenomic data collected from parents in each respective environment. These four models (**M1**-**M4**) with two combinations (20CS and 20LY) used two cross validation scenarios (CVa and CVb) for eight traits (grain yield, plant height, days to

anthesis, starch, protein, fat, fiber, ash) were trained and repeated 20 times. All the prediction models were fitted into Bayesian framework using BGLR package in R (Pérez & De Los Campos, 2014). Prediction accuracy used Pearson's correlation between actual and predicted values in testing sets within environments. Mean and standard deviation were estimated for the combination of each model and cross-validation scheme for all traits. Tukey's tests were performed between four different models within combination of traits, inbred grain samples and CV schemes and significant letters were assigned for their potential comparison.

DISCRIMINANT ANALYSIS OF PRINCIPAL COMPONENTS (DAPC)

To interpret predictability of NIRS-based models, discriminant analysis of principal components (DAPC) was implemented. DAPC clusters different genotypes and environments based on prior information of heterotic groups and environments, respectively (Jombart et al. 2010). The first two linear discriminants (LDs) obtained based on sequential K-means were employed to cluster genotypes and environments. LD analysis was performed using 'lda' function available in the R package 'MASS' (Ripley et al. 2013).

CORRELATIONS BETWEEN HYBRID AND MID-PARENT PERFORMANCE

Pearson's correlations between hybrid and mid-parent performance were estimated for phenotypic performance and NIR bands. Compositional data for only 69 inbred parents were available for environment 20LY, therefore, a total of 633 observations comprising of 212 unique hybrids were used to estimate correlations for grain composition. In addition, correlations between hybrid reflectance and mid-parent reflectance were calculated.

SOFTWARE

All the phenotypic data were fitted in a mixed model using 'lmer' function to extract variance components as well as to compute Best Linear Unbiased Estimates (BLUEs) in RStudio (R Core Team, 2023). The first derivative of the matrix of NIR was computed using '*SavitzkyGolay*' with sliding window of 11 using function available in library '*prospectr*' in R (Stevens & Ramirez-Lopez, 2014). All the prediction models were fitted in BGLR package available in R with 5000 iterations, burnIn of 1000, and a thin of 10 using Gibbs sampler. All the genomic and

phenomic kernels were fitted as semiparametric Reproducing Kernel Hilbert Spaces “RKHS”, however, environment was fitted as incidence matrices that represent fixed effects using “BRR” available in the package. Interaction terms of G × E and NIR × E were computed as Hadamard product (element by element product) between matrices. Furthermore, Hadamard products between incidence matrices of environments and genomic kernels and phenomic kernels were calculated for G × E and NIR × E, respectively. The 'lda’ function available in R package ‘MASS’ was used to infer groupings for linear discriminant analysis.

RESULTS

BEST LINEAR UNBIASED ESTIMATES (BLUES) OF HYBRIDS

The average GY across environments was 6.81 t ha^{-1}, ranging from 1.47 t ha^{-1} to 10.02 t ha^{-1} (**Fig. 1**). Hybrid flowering varied among environments, ranging from 63 to 93 days with a mean of 77.5. Likewise, the average height for hybrids ranged from 92.7 – 188 cm with mean of 136.7 cm. Genotypes were generally taller, later and higher in yield in the CS environments. In composition, total starch ranged from 64.8 – 69.7% with mean of 67.5%. Total protein ranged from 7.23 – 12.7% with mean of 9.45%. Total fat ranged from 1.03 – 3.77% with mean of 2.64%. Fiber content ranged from 1.38 – 2.17% with mean of 1.78%. Lastly, total ash ranged from about 1.25 – 1.63% with mean of 1.44%.

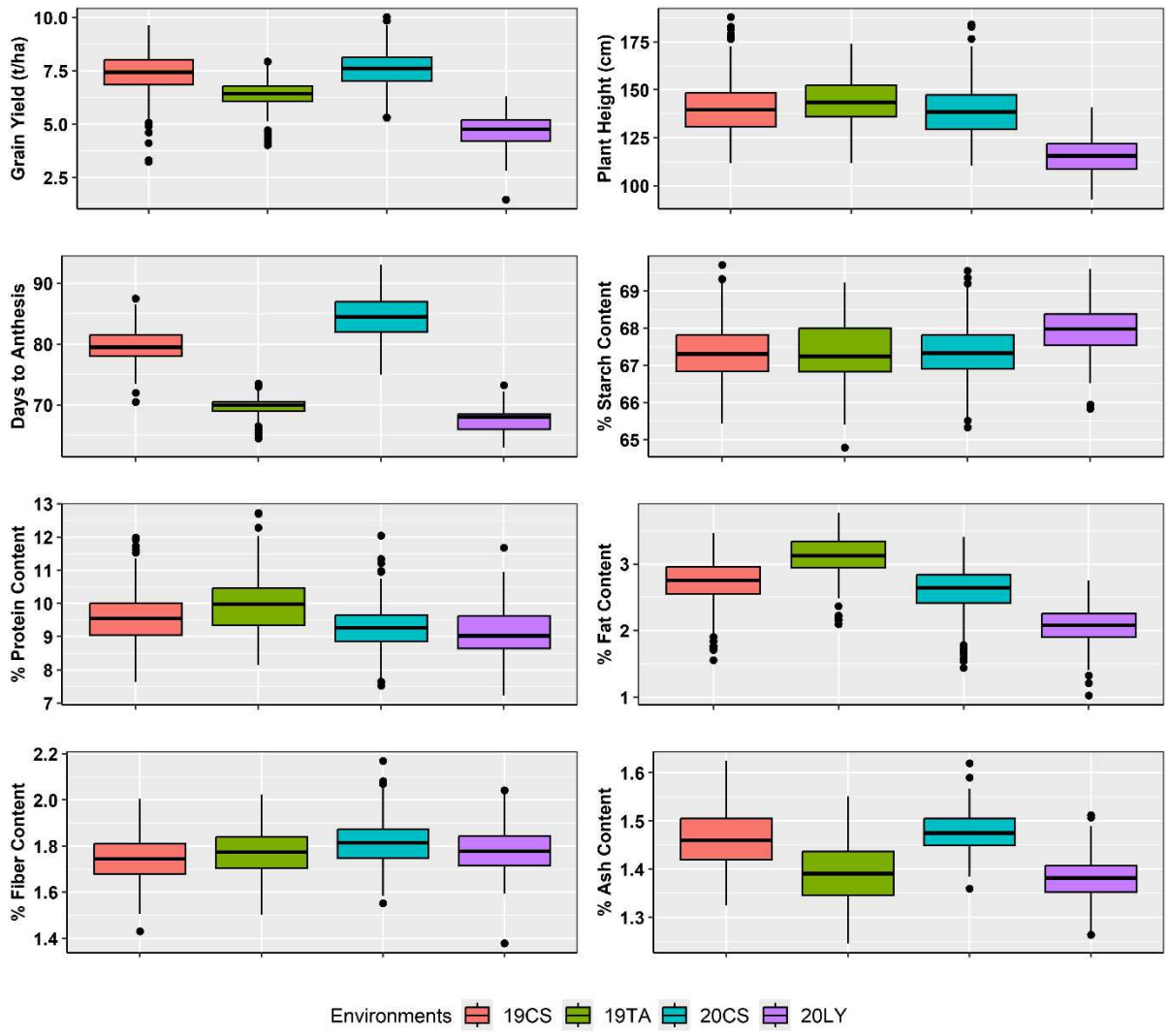


Fig. 1: Distribution of best linear unbiased estimates of agronomic traits grain yield (t/ha), days to anthesis and plant height (cm) for hybrids evaluated across eight environments 19CS (2019 College Station, TX), 19TA (2019 Taft, TX), 20CS (2020 College Station, TX), and 20LY (2020 Lyford, TX).

262

263 **VARIANCE COMPONENT ESTIMATIONS OF PHENOTYPIC PERFORMANCE**

264 Variance component estimations based on likelihood ratio test showed significant GCA effects for both females and
265 males for all traits (**Table 1**). While environment accounted for majority of variation for most traits, genetic effects
266 were predominant for starch (38.9%) and protein (39.6%). Significant G × E (GCA × E and SCA × E) were recorded
267 for all traits. Starch, protein and ash had the highest proportion of G × E among these traits. Broad-sense heritability

was low for ash (0.24) and grain yield (0.25); moderate for starch (0.45) and protein (0.49); and high for the remaining traits (>0.55). Most of the genetic effects were additive with h^2 ranging from 0.188 – 0.59.

Table 1. Variance component estimations and heritability for agronomic traits and grain composition traits (grain yield (GY), days to flowering (DA), and plant height (PH), starch, protein, fat, fiber, and ash content in 366 unique hybrids from 45 × 44 incomplete factorial design that were evaluated in four environments in Texas in 2019 and 2020.

Variance Components[a]	Grain Yield		Plant Height		Days to Anthesis		% Starch		% Protein		% Fat		% Fiber		% Ash	
	Estimate	%	Estimate	%	Estimate	%	Estimate	%	Estimate	%	Estimate	%	Estimate	%	Estimate	%
(Hybrid)	0.277	9.33	150.4	36.1	3.3	4.5	0.29	38.9	0.34	39.6	0.06	19.2	0.0069	53.5	0.0007	11.4
GCA_f	0.152***	5.12	64.3***	15.4	1.7***	2.3	0.12***	15.8	0.13***	15.0	0.03***	8.1	0.0038***	29.5	0.0003***	4.8
GCA_m	0.055***	1.85	60.3***	14.5	1.4***	1.9	0.12***	16.1	0.12***	14.0	0.02***	5.1	0.0024***	18.6	0.0002***	3.3
SCA	0.070***	2.36	25.8***	6.2	0.2^{NS}	0.3	0.05***	7.0	0.09***	10.6	0.02***	6.0	0.0007***	5.4	0.0002***	3.3
Environment	1.43^{NS}	48.15	175.3*	42.0	65.6***	87.5	0.10***	13.0	0.16***	19.2	0.21***	63.7	0.0008^{NS}	6.2	0.0010^{NS}	16.3
(Hybrid × Env)	0.159	5.35	14.4	3.5	2.2	2.9	0.19	26.6	0.21	24.2	0.03	9.9	0.0020	15.5	0.0015	25.2
GCA_f × Env	0.048***	1.62	4.6***	1.1	0.5***	0.6	0.06***	8.7	0.06***	7.0	0.01***	2.7	0.0007***	5.4	0.0007***	11.4
GCA_m × Env	0.023*	0.77	4.9***	1.2	0.5***	0.7	0.06***	8.5	0.05***	5.8	0.01***	2.7	0.0005***	3.9	0.0001*	0.8
SCA × Env	0.088**	2.96	4.9**	1.2	1.2***	1.6	0.07***	9.4	0.10***	11.4	0.02***	4.5	0.0008***	6.2	0***	0.1
Rep (Env)	0.443***	14.92	31.3***	7.5	1.5***	1.9	0.02***	2.6	0.01***	0.2	0.01***	2.1	0.0008***	6.2	0.0023***	37.0
Residual	0.661	22.26	45.7	11.0	2.4	3.2	0.14	18.9	0.14	16.8	0.02	5.1	0.0024	18.6	0.0006	9.8
H^2	0.252		0.71		0.58		0.45		0.49		0.56		0.60		0.24	
h^2	0.188		0.59		0.55		0.37		0.36		0.38		0.54		0.19	
CV_e,%	11.929		4.94		2.01		0.58		4.04		5.28		2.91		2.58	

[a]GCA_f, female general combining ability; GCA_m, male general combining ability; SCA, specific combining ability. *Significant at the 0.05 probability level. **Significant at the 0.01 probability level. ***Significant at the 0.001 probability level. Variance components of (Hybrid) were calculated by summing up GCA_f, GCA_m and SCA and (Hybrid × Env) was estimated by summing up GCA_f × Env, GCA_m × Env and SCA × Env.

HERITABILITY OF PARENTAL NIRS

Variance components extracted by fitting random effect model **equation 1** estimated entry-mean heritability of reflectance values of all wavelengths for environments 20CS and 20LY. In both environments, heritability of 0.99 was observed at around 485 – 490 nm while heritability of 0.88 – 0.91 was recorded at around 870 – 880 nm (**Fig. 2**). Since reflectance values were highly heritable, all the bands were used for predictive modeling purposes.

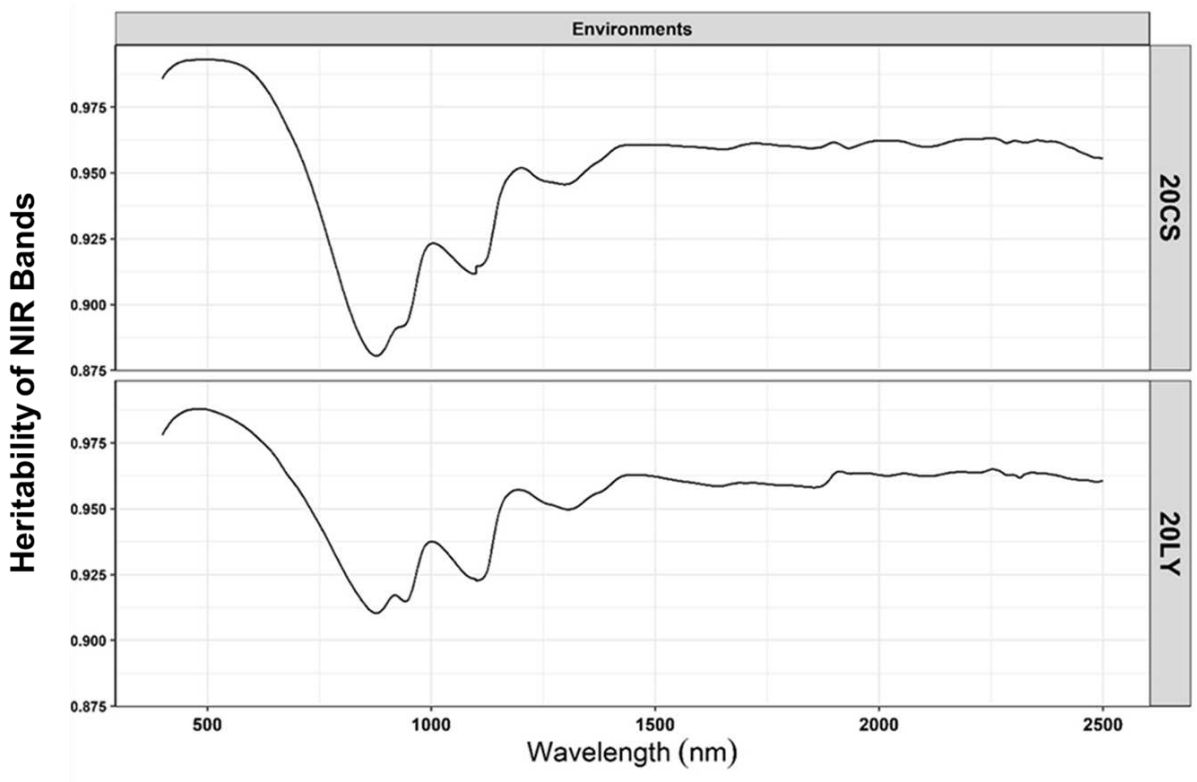


Fig. 2: Heritability (genetic values) using near infrared spectroscopy (NIRS) bands collected on all parents across all 4200 bands (400 – 2500 nm) in two environments as computed by $H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{r}}$ from random effects model fitted as $y = 1\mu + Z_1 g + Z_2 r + e$.

Prediction Abilities for Hybrid Performance of Un-testcrossed Inbred Lines (CVa)

Using inbred grain samples from environments 20CS and 20LY, **M3** numerically outperformed any other models for GY, starch, protein and fat (**Fig. 3**). Furthermore, **M3** numerically outperformed **M1** for all traits except for PH and

DA for both 20CS and 20LY grain samples. Similarly, the predictive ability of **M2** was equal or statistically better than **M1** for GY and grain composition traits for both 20CS or 20LY. Furthermore, **M2** numerically outperformed **M1** for compositional traits for 20CS inbred grain samples and same is true for GY, DA and fat while using 20LY inbred grain samples. In general, prediction accuracy improved with the inclusion of parental NIR information into genomic models (**M4**).

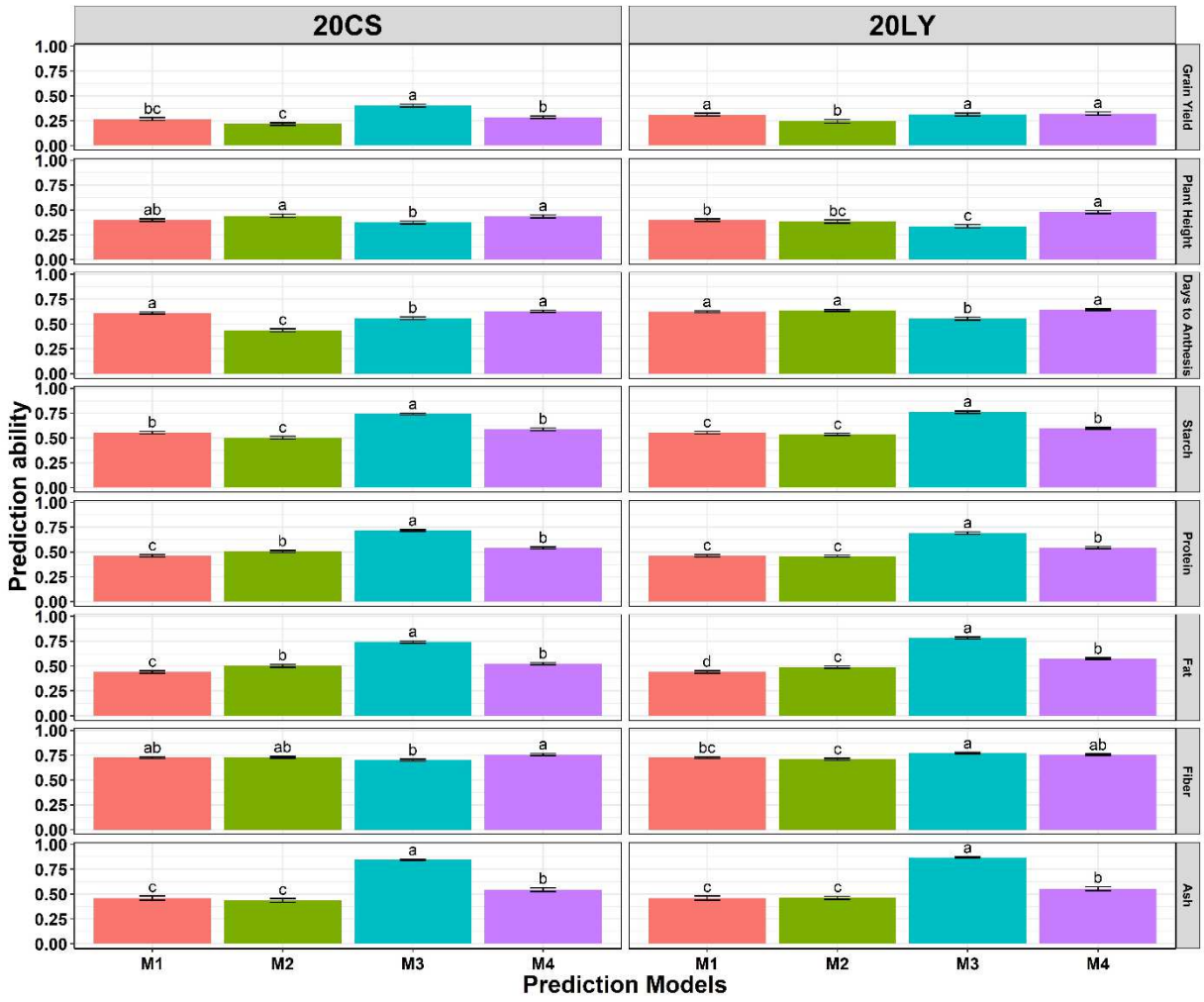


Fig. 3: Predictive ability of four different prediction models, **M1**: Genomic (G); **M2**: Mid-parent (MP); **M3**: Mid-parent heterosis (MPH); **M4**: G + MP for eight different traits Grain Yield (GY), Days to Anthesis (DA), Plant Height (PH), Starch, Protein, Fat, Fiber and Ash across four environments (19CS: 2019 College Station, 19TA: 2019 Taft, 20CS: 2020 College Station, 20LY: 2020 Lyford) under scenario of CVa. 20CS uses grain samples obtained on inbreds from 2020 College Station while 20LY uses grain samples obtained on inbreds from 2020

Lyford. Error bars and significant letters a, b, and c were assigned for statistical differences determined by Tukey's honest test between models.

Prediction Abilities for Predicting New Hybrids (CVb)

M3 outperformed **M1** in predicting starch, protein, fiber, and ash for 20CS and 20LY inbred grain samples. **M2** performed the same or better than **M1** for all the examined traits while using 20LY parental grain samples; the same occurred for 20CS parental grain samples except for GY, DA and fiber (**Fig. 4**). **M4** is numerically better than **M1**, for GY, starch, protein, fiber, fat, and ash for 20CS inbred grain samples and same is true for plant height, days to anthesis, starch, protein, fat, and ash for 20LY inbred grain samples (**Fig. 4**).

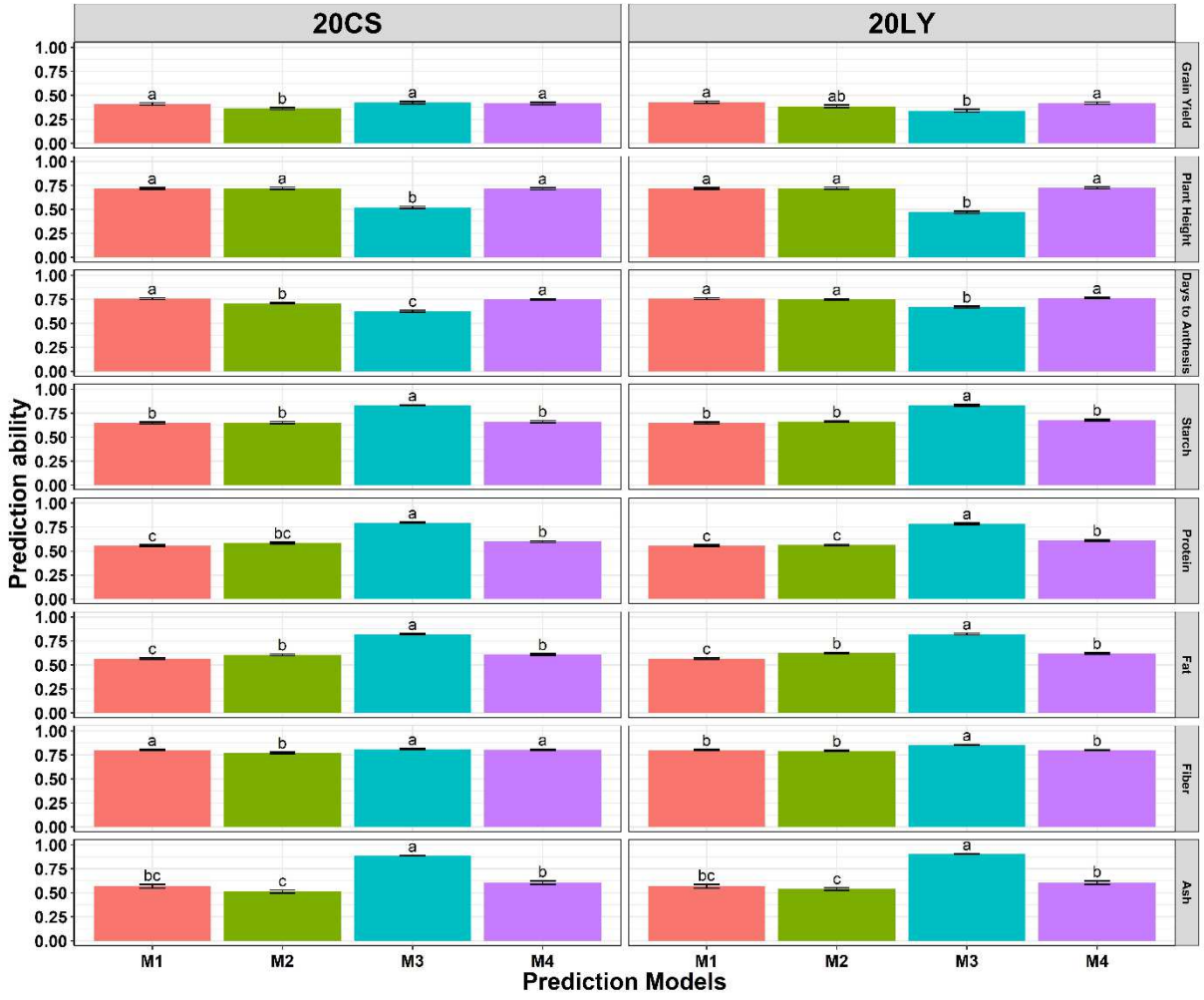


Fig. 4: Predictive ability of four different prediction models, **M1**: Genomic (G); **M2**: Mid-parent (MP); **M3**: Mid-parent heterosis (MPH); **M4**: G + MP for eight different traits Grain Yield (GY), Days to Anthesis (DA), Plant Height (PH), Starch, Protein, Fat, Fiber and Ash across four environments (19CS: 2019 College Station, 19TA: 2019 Taft, 20CS: 2020 College Station, 20LY: 2020 Lyford) under scenario of **CVb**. 20CS uses grain samples obtained on inbreds from 2020 College Station while 20LY uses grain samples obtained on inbreds from 2020 Lyford. Error bars and significant letters a, b, and c were assigned for statistical differences determined by Tukey's honest test between models.

290 **Discriminant Analysis of Principal Components to Characterize Heterotic Groups and Environments**

291 Linear Discriminant (LD) functions using grain samples of inbred parents explained 97.9% of variance using first
292 two LDs. Most of the variation was associated with the environment (20CS and 20LY), but heterotic groups were
293 distinct within each environment (**Figure. 5A**). NIR reflectance data from hybrid grain samples explained about
294 87.4% of variation using first two LDs. Environment effects varied: 2019 environments clustered together, while
295 environment 20LY and 20CS were distinct for hybrid grain samples (**Figure. 5B**).

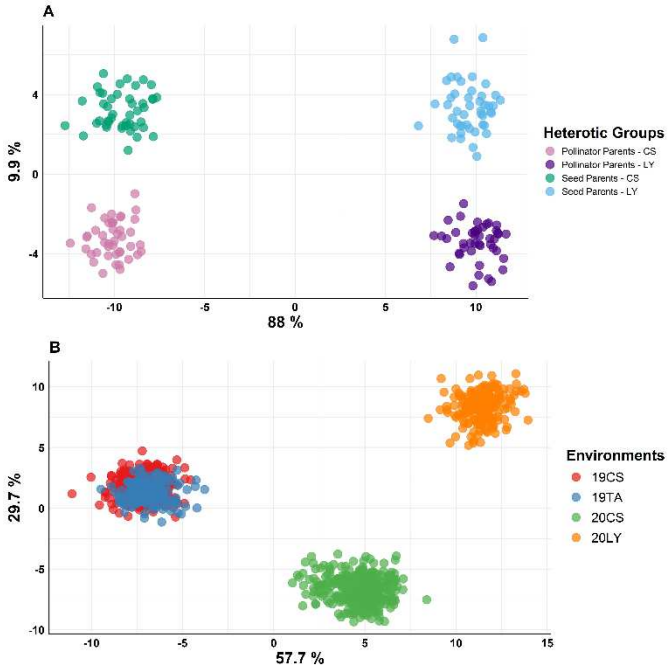


Fig. 5: Discriminant Analysis of Principal Components (DAPC) on NIR for grain samples of parents and hybrids, A) Coloring is based on combinations of heterotic groups and respective environments (CS: College Station, LY: Lyford) B) Coloring is based on environments and NIRS was based on hybrids grain samples collected from respective four environments (19CS: 2019 College Station, 19TA: 2019 Taft, 20CS: 2020 College Station and 20LY: 2020 Lyford)

296

297 **Correlation between mid-parent performance and hybrid performance**

298 Pearson correlations between mid-parent and hybrid performance varied for traits and NIR bands. Significant positive
299 correlations between mid-parent and hybrid performance were observed for GY, DA, and PH in both 20CS and 20LY
300 (**Table 2**). Significant correlations were observed for grain composition as well where NIR bands showed significantly
301 higher correlation (0.62) for 20CS than in 20LY (**Fig. 6**). All the correlations for NIR bands within or across-
302 environments were highly significant at the level of probability of 0.001 (***).

Table 2: Pearson's correlations between mid-parent performance and hybrid performance for grain yield, days to anthesis, plant height, starch, protein, fiber, fat, and ash across four environments. The environments are: 2019 College Station Hybrid (19CS H), 2019 Taft Hybrid (19TA H), 2020 College Station Hybrid (20CS H), 2020 Lyford Hybrid (2020 LY H), mid-parent performance at 2020 College Station (2020 CS MP) and mid-parent performance at 2020 Lyford (2020 LY MP) for grain yield, days to anthesis, plant height, starch, protein, fat, fiber, and ash.

Environments	**19CS H**	**19TA H**	**20CS H**	**20LY H**
		Grain Yield		
20CS MP	0.158***	0.476***	0.187***	0.173***
20LY MP	0.186***	0.308***	0.186***	0.347***
		Days to Anthesis		
20CS MP	0.519***	0.508***	0.589***	0.359***
20LY MP	0.493***	0.531***	0.300***	0.708***
		Plant Height		
20CS MP	0.589***	0.642***	0.542***	0.436***
20LY MP	0.393***	0.504***	0.354***	0.294***
		Starch		
20CS MP	0.542***	0.497***	0.544***	0.359***
20LY MP	0.166*	0.21*	0.223***	0.33***
		Protein		
20CS MP	0.515***	0.499***	0.555***	0.418***
20LY MP	0.345***	0.431***	0.403***	0.508***
		Fat		
20CS MP	0.45***	0.546***	0.509***	0.369***
20LY MP	0.404***	0.538***	0.400***	0.468***

		Fiber		
20CS MP	0.703***	0.721***	0.701***	0.636***
20LY MP	0.600***	0.661***	0.564***	0.63***
		Ash		
20CS MP	0.357***	0.495***	0.522***	0.231***
20LY MP	0.245***	0.54***	0.213***	0.491***
Significance levels are indicated by * ($\alpha = 0.05$), ** ($\alpha = 0.01$) and *** ($\alpha = 0.001$)				

303

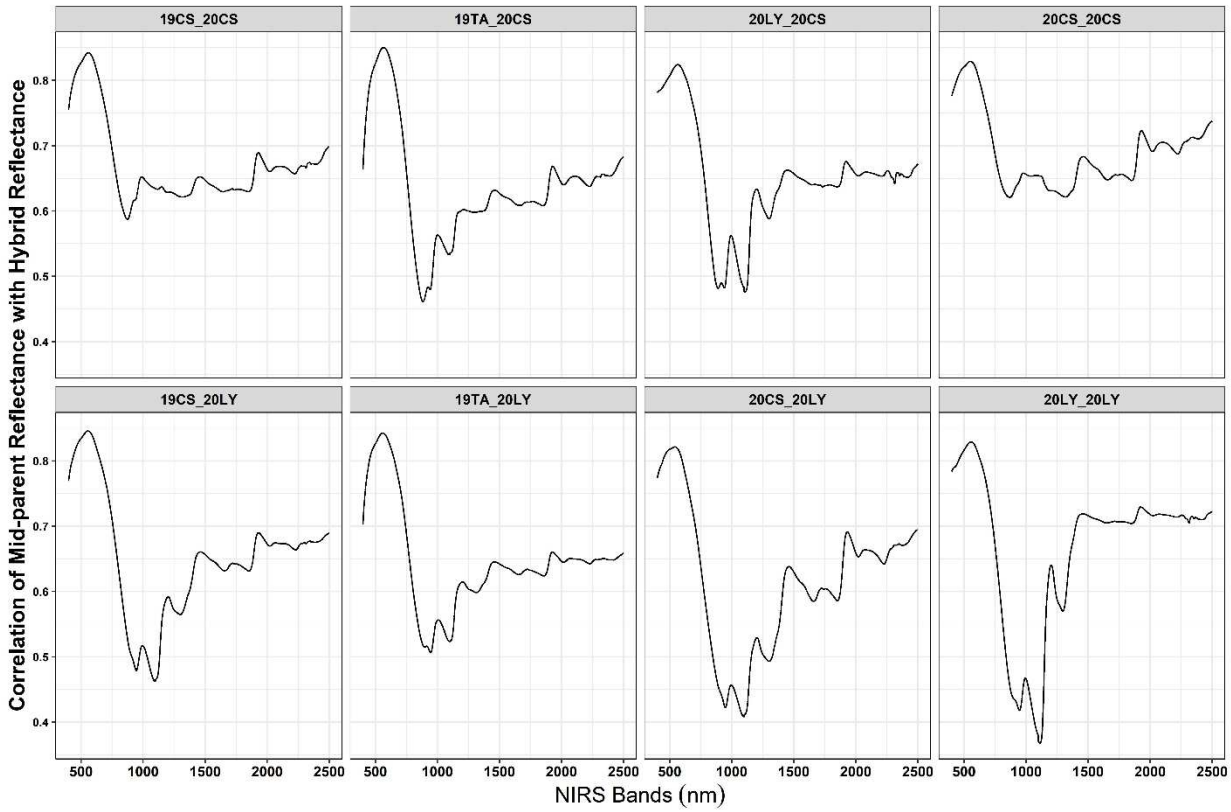


Fig. 6: Pearson's correlation between mid-parent NIR reflectance and hybrid NIR reflectance for eight different combinations. 19CS_20CS, 19TA_20CS, 20LY_20CS and 20CS_20CS represent correlation between mid-parent values for 20CS grain samples with hybrid grain samples from 19CS, 19TA, 20LY and 20CS, respectively; likewise, 19CS_20LY, 19TA_20LY, 20CS_20LY and 20LY_20LY represent correlation between mid-parent values for 20LY grain samples with hybrid grain samples from 19CS, 19TA, 20CS and 20LY, respectively.

304 **DISCUSSION**

305 Phenomic selection typically uses the same concept as genomic selection (Rincent et al. 2018) but their application

306 lags in comparison in breeding programs (Sandhu et al. 2022). Prior work demonstrate that GS and PS models were

effective at predicting hybrid performance (Sapkota et al. 2025a), and this work tests the reliability of phenomic data obtained on parents and their usefulness in predicting hybrid performances. We have presented range of prediction models and their efficacy with cross-validation scenarios relevant to the plant breeding program. If replicable, the phenomic prediction methodologies presented herein are applicable to hybrid breeding programs.

Research studies indicate that parental NIR scans can accurately predict hybrid performance in rapeseed (*Brassica napus*) and corn (Rincent et al. 2025; Roscher-Ehrig et al. 2024). Our objectives of this study were to test the effectiveness of phenomic prediction models using NIR reflectance data obtained from parents and hybrids to predict hybrid performance. Further, we have employed the concepts of quantitative genetics to model phenotypic variation in hybrid grain sorghum. This study presents unique methods of prediction modeling and specific-use case-scenarios that use NIR reflectance data obtained from parents as well as hybrids in sorghum.

Phenotypic Performance of Hybrids Across Environments

Agronomic trait data varied as typically expected in multi-environment trials. The high GY in 20CS (7.55 t ha^{-1}) is due to higher moisture availability during the growing season. Environmental differences also affected PH and DA (**Fig. 1**). Significant G × E interaction (GCA × E as well as SCA × E) for agronomic traits as well as grain composition traits revealed utility of deploying environment specific selection strategies (Sapkota et al. 2025b). Environmental effects likely reduce heritability for GY, but the opposite was observed for grain composition traits (**Table 1**). The highly heritable nature of NIR spectra of parents for both environments 20CS and 20LY illustrates consistency and reliability of phenomic data (**Fig. 2**). Typically, spectra must be collected from multi-environments to effectively capture variation. However, the high repeatability of spectrum for both environments along with higher correlation of repeatability between environments (r = 0.99) suggests they could be collected from a single representative environment (Robert et al. 2022a; Sapkota et al. 2025a).

Predicting Unobserved Inbred Lines (CVa)

To predict inbred lines not evaluated in hybrid combination (**CVa**), **M3** outperformed **M1** by about 14% for GY when using inbred grain samples from 20CS (**Fig. 3**). These differences were more pronounced for starch, protein, fat, and

ash. **M3** was trained with MPH of spectra, which models both GCA effects due to additive effects as well as SCA due to dominance effects (Endelman, 2025). Consistently higher prediction accuracy for **M3** is likely due to significant SCA effects for agronomic traits (**Table 1**). Likewise, higher prediction accuracy of grain composition trait was due to use of hybrid reflectance data in predicting hybrid grain composition traits. Although this does not seem logical, such model can reduce the need for extensive scanning of hybrid grain samples. Future studies focusing on the ability to model SCA effects using NIR bands from parents are needed.

Traditional phenomic prediction – which used phenomic data on hybrids vs genomic prediction – which used genomic data obtained on parents measure two different model properties and their direct comparison can be flawed (Wang et al. 2025). Rincent et al. (2025) reported that phenomic prediction models are more accurate than pedigree-based approaches, asserting that phenomic data improves the captures of Mendelian sampling and epistatic interactions. To address this, comparison should be made between genomic prediction models (i.e. **M1)** and phenomic prediction models using scans on parents (i.e. **M2)**. **M2** (0.218 ± 0.129) were like **M1** (0.266 ± 0.128) when inbred grain samples from 20CS were used. Inclusion of phenomic data into genomic prediction models (**M4**) significantly improved prediction ability for protein, fat and ash with numerical advantages for other traits in comparison to the baseline genomic prediction model (**M1**). Therefore, phenomic prediction provides an alternative and can complement genomic prediction for predicting performance of new sorghum lines in hybrid combinations.

Predicting Unobserved Hybrids (CVb)

When predicting new hybrids (**CVb**), **M2** produces comparable prediction accuracy to **M1** and **M3** resulted in improved prediction for grain composition traits. Although **M2** was less effective at predicting DA and PH in the **CVa** scenario, **M2** was competitive to **M1** in **CVb** although relationship between PH and grain samples were not always understood (Dallinger et al. 2023; Sapkota et al. 2025a). Higher prediction accuracy of **M2** in **CVb** in comparison to **CVa** is likely due to ability to borrow information from some genetically related hybrids. Higher prediction accuracy of 20LY grain samples (relative to 20CS grain samples) may be due to the value of grain sampling from a more stressed environment. Similar trends were previously noted in maize, wheat and poplar (Ferrio et al. 2005; Rincent et

al. 2018). Improved predictability of phenomic prediction model is attributed to the tissue-specific and environment-specific phenome unlike genomic data (Dallinger et al. 2023).

Phenomic Characterization and Clustering

Traditionally, molecular markers are implemented to cluster genotypes and phenotypic information are used in grouping environments. In a similar way, grain phenotypes harbor information on genetics, environment as well as genotype × environment interactions (Robert et al. 2022c; Sapkota et al. 2025a). Therefore, genotypes and environments could be grouped using grain samples from both inbred parents as well as hybrids (**Fig. 5**). Inbred parents were consistently grouped into two heterotic groups for two environments (**Fig. 5A**). In addition, grain samples collected from hybrid trials were represented as expected phenotypic differences as well as geographical locations (**Fig. 5B**). Geographically closer locations within the same year (19TA and 19CS) were more similar, likely due to similar moisture availability in that year which has been consistently shown to be the primary driver of productivity in sorghum (Leon et al. 2024). The separate clustering of 20CS and 20LY was due to distinct geographical locations (**Fig. 5B**). In other words, 20CS and 20LY hybrid grain samples represent high and low yielding environments.

Relationship between Inbred and Hybrid Grain Composition and Spectra

While inbred performance is moderately correlated with hybrid performance in sorghum and corn, it is not sufficient for selection in either crop (Crozier et al. 2023; Rooney, 2004; Hallauer & Miranda, 1988). While there is a plethora of information on parent-offspring correlations for agronomic performance, this study reports such correlations for grain composition and phenomic spectra for the first time in sorghum. Significant correlations of 0.166* - 0.721*** across environments revealed similar inferences could be applied for grain composition traits (**Table 2**). In a similar fashion, a highly significant and strong correlation was recorded for NIR bands from samples from 20CS inbred grain samples (>0.6***) and 20LY inbred grain samples (>0.4***). Consistently higher correlation of mid-parent reflectance with hybrid reflectance justifies their direct use in a prediction model (**Fig. 6**). Such correlations were detected even across environments. This confirms NIR reflectance obtained from parents is representative of hybrid NIR reflectance, therefore, could be used in predicting hybrid performance in similar fashion. In addition to genetic

differences, NIR bands capture relatedness between environments resulting in potentially more efficient phenomic prediction models (Bienvenu et al. 2025; Sapkota et al. 2025a). As evident by better information on NIR bands, fewer phenomic predictors can approach the prediction accuracy of genomic prediction (Zhu et al. 2021).

Applications in Plant Breeding Programs

In recent years, PS approach has evolved potential methods to improve selection intensity in a breeding program. The efficacy of PS approach is often compared with GS but direct comparison between PS and GS is flawed since GS models use molecular markers from parents, while PS models typically use reflectance values obtained in testing genotypes. Even if the prediction ability of PS is higher than that of GS, adoption in a breeding program is logistically challenging (Wang et al. 2025). All the models in this study have relative value in a plant breeding program, but deployment depends largely upon the logistics of a breeding program. We have presented more reliable comparisons between genomic prediction model (**M1**) and phenomic prediction model (**M2**) using respective indices for parents. While **M2** can model GCA effects theoretically, **M3** has ability to model SCA effects as well using mid-parent heterosis which requires phenomic data from hybrids. While both GCA effects and SCA effects are important in sorghum, early stage inbred selection primarily depends upon GCA values. In this study, about 72% of genetic variance for GY is contributed by GCA effects (**Table 1**). With such comparable prediction accuracies, **M2** can be promoted in place of **M1** which fits well with new inbred selection during early-stage testing. These can be further extended to other hybrid crops such as corn where GCA effects comprise about 93% of the total genetic variation (Westhues et al. 2017).

Phenomic markers are limited, environment subjective, and require large-scale multi-environment trials in comparison to genomic data (Delwiche et al. 2002). Potentially similar prediction accuracy for phenomic data collected from different environments suggest grain sampling can be performed in fewer environments if genetically related hybrids are evaluated (Rincent et al. 2018). Although its causal basis is still debated, phenomic prediction has strong potential in breeding programs (Feldmann et al. 2026). With NIR based infrastructure being available within the breeding programs, this phenomic prediction can be adopted alone or can co-exist with genomic prediction. This can provide additional opportunities to integrate prediction-based frameworks where there are logistical challenges of genotyping.

While genomic selection can be integrated well with single seed descent and double haploid, this technology requires phenotypic evaluation of inbreds and the collection of grain samples. Therefore, these methodologies fit well with a breeding program which still performs inbred evaluation.

While NIR-based PS systems can improve the efficiency of a breeding program, absolute reliance on it can lead to undesirable changes in grain composition since NIR-based phenomic system works based on underlying correlations of specific grain composition (chemical compounds) with targeted traits. If these correlations are negative, continuous adoption of PS could lead to decrease in such compositional traits (Dallinger et al. 2023). For example, consistent decline in protein can happen as grain yield and protein are known to have strong negative correlation in sorghum (Pfeiffer et al. 2017; Sapkota et al. 2025b). Therefore, regular assessment of grain composition and optimization of PS models is recommended to reduce such negative consequences. As most sorghum breeding programs in the US are relatively small, the adoption of PS over GS becomes more feasible in terms of logistics (Dallinger et al. 2023).

CONCLUSION

Herein, low cost-efficient NIRS data were collected on parental lines and shown to be repeatable and containing sufficient information to discriminate between genotypes and environments. Further, this NIRS data was successfully used to predict hybrid performance in sorghum for grain yield, days to anthesis, plant height, starch, protein, fat, fiber and ash; and was shown to be competitive with GS models. Although other implementations of phenomic data should be tested, PS based on inbred spectra presents a low-cost alternative to GS methods to eliminate low-performing hybrid and inbred genotypes. Successful implementation of NIRS-based prediction modeling could improve selection intensity, selection accuracy, breeding cycle and ultimately rate of genetic gain in sorghum.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets generated for current study are available from the corresponding author on reasonable request.

REFERENCES

Acosta-Pech R, Crossa J, de los Campos G, Teyssèdre S, Claustres B, Pérez-Elizalde S, Pérez-Rodríguez P (2017) Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theor Appl Genet 130:1431–1440. https://doi.org/10.1007/s00122-017-2898-0

Adunola P, Tavares Flores E, Riva-Souza EM, Ferrão MAG, Senra JFB, Comério M, et al (2024) A comparison of genomic and phenomic selection methods for yield prediction in Coffea canephora. Plant Phenom J 7:e20109. https://doi.org/10.1002/ppj2.20109

Alemu A, Åstrand J, Montesinos-López OA, Isidro y Sánchez J, Fernández-Gónzalez J, Tadesse W, Vetukuri RR, Carlsson AS, Ceplitis A, Crossa J, Ortiz R, Chawade A (2024) Genomic selection in plant breeding: key factors shaping two decades of progress. Mol Plant 17:552–578. https://doi.org/10.1016/j.molp.2024.03.007

Bienvenu C, Garin V, Salas N, Théra K, Tekete ML, Sarathjith MC, et al (2025) Factors influencing phenomic prediction: a case study on a large sorghum back cross nested association mapping population. Plant Phenom J 8:e70051. https://doi.org/10.1002/ppj2.70051

Budhlakoti N, Kushwaha AK, Rai A, Chaturvedi KK, Kumar A, Pradhan AK, Kumar U, Kumar RR, Juliana P, Mishra DC, Kumar S (2022) Genomic selection: a tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. Front Genet 13:832153. https://doi.org/10.3389/fgene.2022.832153

Crozier D, Leon F, Fonseca JMO, Klein PE, Klein RR, Rooney WL (2023) Inbred phenotypic data and non-additive effects can enhance genomic prediction models for hybrid grain sorghum. Crop Sci 63:1183–1196. https://doi.org/10.1002/csc2.20927

Cuevas J, Montesinos-López O, Juliana P, Guzmán C, Pérez-Rodríguez P, González-Bucio J, et al (2019) Deep kernel for genomic and near infrared predictions in multi-environment breeding trials. G3 Genes Genomes Genet 9:2913–2924. https://doi.org/10.1534/g3.119.400493

449 Dallinger HG, Löschenberger F, Bistrich H, Ametz C, Hetzendorfer H, Morales L, et al (2023) Predictor bias in
450 genomic and phenomic selection. Theor Appl Genet 136:235. https://doi.org/10.1007/s00122-023-04479-8

451 DeSalvio AJ, Adak A, Murray SC, Jarquín D, Winans ND, Crozier D, Rooney WL (2024) Near-infrared reflectance
452 spectroscopy phenomic prediction can perform similarly to genomic prediction of maize agronomic traits across
453 environments. Plant Genome 17:e20454. https://doi.org/10.1002/tpg2.20454

454 Delwiche SR, Graybosch RA, Nelson LA, Hruschka WR (2002) Environmental effects on developing wheat as sensed
455 by near-infrared reflectance of mature grains. Cereal Chem 79:885–891.
456 https://doi.org/10.1094/CCHEM.2002.79.6.885

457 Endelman JB (2025) Genomic prediction of heterosis, inbreeding control, and mate allocation in outbred diploid and
458 tetraploid populations. Genetics 229:193. https://doi.org/10.1093/genetics/iyae193

459 Feldmann MJ, Wang F, Runcie DE (2026) Philosophy of phenomic prediction and its incompatibility with causal
460 inference. Plant Phenom J 9:e70091. https://doi.org/10.1002/ppj2.70091

461 Ferrio JP, Villegas D, Zarco J, Aparicio N, Araus JL, Royo C (2005) Assessment of durum wheat yield using visible
462 and near-infrared reflectance spectra of canopies. Field Crops Res 94:126–148.
463 https://doi.org/10.1016/j.fcr.2004.12.002

464 Galán RJ, Bernal-Vasquez AM, Jebsen C, Piepho HP, Thorwarth P, Steffan P, et al (2020) Integration of genotypic,
465 hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye. Theor Appl Genet
466 133:3001–3015. https://doi.org/10.1007/s00122-020-03651-8

467 Hallauer AR, Miranda JB (1988) Quantitative genetics in maize breeding, 2nd edn. Iowa State University Press, Ames

468 Holland JB (2004) Implementation of molecular markers for quantitative traits in breeding programs—challenges and
469 opportunities. Proc 4th International Crop Science Congress 26:1–13

Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. Nat Rev Genet 11:855–866. https://doi.org/10.1038/nrg2897

Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 Genes Genomes Genet 4:1027–1046. https://doi.org/10.1534/g3.114.010298

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94. https://doi.org/10.1186/1471-2156-11-94

Leon F, Harvey L, Rooney WL (2024) Historical multi-environmental trials reveal trends in hybrid grain sorghum performance from 1970 to 2021 in Texas. Crop Sci 64:3014–3027. https://doi.org/10.1002/csc2.21343

Meyenberg C, Braun V, Friedrich C, Longin H, Thorwarth P (2024) Feature engineering and parameter tuning: improving phenomic prediction ability in multi-environmental durum wheat breeding trials. Theor Appl Genet 137:188. https://doi.org/10.1007/s00122-024-04695-w

Morishige DT, Klein PE, Hilley JL, Sahraeian SME, Sharma A, Mullet JE (2013) Digital genotyping of sorghum—a diverse plant species with a large repeat-rich genome. BMC Genomics 14:448. https://doi.org/10.1186/1471-2164-14-448

Parmley K, Nagasubramanian K, Sarkar S, Ganapathysubramanian B, Singh AK (2019) Development of optimized phenomic predictors for efficient plant breeding decisions using phenomic-assisted selection in soybean. Plant Phenomics 2019:5809404. https://doi.org/10.34133/2019/5809404

Pasquini C (2003) Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. J Braz Chem Soc 14:198–219. https://doi.org/10.1590/S0103-50532003000200006

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495. https://doi.org/10.1534/genetics.114.164442

Pfeiffer BK, Pietsch D, Schnell RW, Rooney WL (2019) Long-term selection in hybrid sorghum breeding programs. Crop Sci 59:150–164. https://doi.org/10.2135/cropsci2018.05.0345

R Core Team (2023) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rincent R, Charpentier JP, Faivre-Rampant P, Paux E, Le Gouis J, Bastien C, Segura V (2018) Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. G3 Genes Genomes Genet 8:3961–3972. https://doi.org/10.1534/g3.118.200760

Rincent R, Solin J, Lorenzi A, Nunes L, Griveau Y, Pirus L (2025) Using phenomic selection to predict hybrid values with NIR spectra measured on the parental lines: proof of concept on maize. Theor Appl Genet 138:28. https://doi.org/10.1007/s00122-024-04809-4

Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, Ripley MB (2013) Package 'MASS'. CRAN 538:113–120

Robert P, Auzanneau J, Goudemand E, Oury FX, Rolland B, Heumez E, et al (2022a) Phenomic selection in wheat breeding: identification and optimisation of factors influencing prediction accuracy and comparison to genomic selection. Theor Appl Genet 135:895–914. https://doi.org/10.1007/s00122-021-04005-8

Robert P, Brault C, Rincent R, Vincent S (2022b) Phenomic selection: a new and efficient alternative to genomic selection. In: Genomic prediction of complex traits: methods and protocols, pp 397–420. https://doi.org/10.1007/978-1-0716-2205-6_14

Robert P, Goudemand E, Auzanneau J, Oury FX, Rolland B, Heumez E, Bouchet S, Caillebotte A, Mary-Huard T, Le Gouis J, Rincent R (2022c) Phenomic selection in wheat breeding: prediction of the genotype-by-environment interaction in multi-environment breeding trials. Theor Appl Genet 135:3337–3356. https://doi.org/10.1007/s00122-022-04170-4

514 Rooney WL (2004) Sorghum improvement—integrating traditional and new technology to produce improved
515 genotypes. Adv Agron 83:37–109. https://doi.org/10.1016/S0065-2113(04)83002-5

516 Roscher-Ehrig L, Weber SE, Abbadi A, Malenica M, Abel S, Hemker R, Snowdon RJ, Wittkop B, Stahl A (2024)
517 Phenomic selection for hybrid rapeseed breeding. Plant Phenomics 6:0215.
518 https://doi.org/10.34133/plantphenomics.0215

519 Sandhu KS, Merrick LF, Sankaran S, Zhang Z, Carter AH (2022) Prospectus of genomic selection and phenomics in
520 cereal, legume and oilseed breeding programs. Front Genet 12:829131.
521 https://doi.org/10.3389/fgene.2021.829131

522 Sapkota P, Fonseca J, Perumal R, Crossa J, Rooney WL (2025a) Near infrared spectroscopy based on phenomic data
523 can improve genomic prediction of sorghum agronomic and kernel-related traits across environments. Plants
524 14:2871. https://doi.org/10.3390/plants14182871

525 Sapkota P, Perumal R, Fonseca J, Klein P, Klein RR, Awika J, Rooney WL (2025b) Genetic architecture and genomic
526 selection of grain quality traits in sorghum hybrids across multi-environment trials. Crop Sci 65:e70189.
527 https://doi.org/10.1002/csc2.70189

528 Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE (2018) Beyond genomic
529 prediction: combining different types of omics data can improve the prediction of hybrid performance in maize.
530 Genetics 208:1373–1385. https://doi.org/10.1534/genetics.117.300374

531 Savitzky A, Golay MJ (1964) Smoothing and differentiation of data by simplified least squares procedures. Anal Chem
532 36:1627–1639. https://doi.org/10.1021/ac60214a047

533 Stevens A, Ramirez-Lopez L (2014) An introduction to the prospectr package.

534 VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423.
535 https://doi.org/10.3168/jds.2007-0980

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

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

540 Yan J, Wang X (2023) Machine learning bridges omics sciences and plant breeding. Trends Plant Sci 28:199–210.
541 https://doi.org/10.1016/j.tplants.2022.08.018

542 Zhu X, Leiser WL, Hahn V, Würschum T (2021) Phenomic selection is competitive with genomic selection for
543 breeding of complex traits. Plant Phenom J 4:e20027. https://doi.org/10.1002/ppj2.20027