Integration of hyperspectral, genomic, and agronomic data for early prediction of biomass yield in hybrid rye (*Secale cereale* L.)

Dissertation to obtain the doctoral degree of Agricultural Sciences
(Dr. sc. agr.)

Faculty of Agricultural Sciences
University of Hohenheim

State Plant Breeding Institute

Submitted by
Master of Science
Rodrigo José Galán
from Córdoba, Argentina

2021

This thesis was accepted as a doctoral dissertation in fulfillment of the requirements for the degree “Doktor der Agrarwissenschaften (Dr. sc. agr)” by the Faculty of Agricultural Sciences at the University of Hohenheim, on June 28, 2021.

Day of oral examination: July 21, 2021.

Examination Committee:

Dean	Prof. Dr. Ralf Vögele
Head of Committee:	Prof. Dr. Martin Hasselmann
1st examiner and reviewer:	apl. Prof. Dr. Thomas Miedaner
2nd examiner and reviewer:	Prof. Dr. Klaus Pillen (Martin-Luther-Universtität Halle-Wittenberg)
3rd examiner:	Prof. Dr. Hans-Peter Piepho

This Ph.D. thesis was funded by the German Federal Ministry of Food and Agriculture (BMEL) through the German Agency for Renewable Resources (FNR), grant number FKZ 22019716.

With support from

by decision of the
German Bundestag

To my parents and my lovely wife

In memory of my highly esteemed professor,
Agric. Eng. (M.Sc.) Eduardo Ruiz Posse
National University of Córdoba, Argentina

TABLE OF CONTENTS

1. General introduction 1

1.1. Rye – a versatile dual-purpose crop 1

1.2. Breeding rye for enhanced biomass yield 2

1.3. Incorporation of molecular data into plant breeding programs 5

1.3.1. Marker-assisted selection 5

1.3.2. Genomic selection 5

1.4. High-throughput phenotyping as a valuable breeding tool 7

2. Objectives 10

3. Publication I: Hyperspectral reflectance data and agronomic traits can predict biomass yield in winter rye hybrids 11

4. Publication II: Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye 13

5. Publication III: Early prediction of biomass in hybrid rye based on hyperspectral data surpasses genomic predictability in less-related breeding material 41

6. General Discussion 64

6.1. Hyperspectral imaging for indirect biomass estimation 64

6.1.1. Data acquisition and management 64

6.1.2. On the predictability of models based on vegetation indices and full-spectrum data 69

6.2. Factors influencing the prediction ability of GBLUP and HBLUP 71

6.2.1. The training set size 72

6.2.2. The genetic and environmental connectivity between training and validation data 72

6.2.3. Characteristics of the trait under study 74

6.3. Conclusions for rye biomass breeding in the "omics" era 75

7. Summary 81

8. Zusammenfassung 83

9. References 86

Acknowledgments 102

Curriculum Vitae 104

Declaration 105

Publications:

Publication I: Galán RJ, Bernal-Vasquez A-M, Jebsen C, Piepho H-P, Thorwarth P, Steffan P, Gordillo A, Miedaner T (2020) Hyperspectral reflectance data and agronomic traits can predict biomass yield in winter rye hybrids. BioEnergy Res 13:168–182. doi: 10.1007/s12155-019-10080-z

Publication II: Galán RJ, Bernal-Vasquez A-M, Jebsen C, Piepho H-P, Thorwarth P, Steffan P, Gordillo A, Miedaner T (2020) Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye. Theor Appl Genet 133: 3001–3015. doi: 10.1007/s00122-020-03651-8

Publication III: Galán RJ, Bernal-Vasquez A-M, Jebsen C, Piepho H-P, Thorwarth P, Steffan P, Gordillo A, Miedaner T (2021) Early prediction of biomass in hybrid rye based on hyperspectral data surpasses genomic predictability in less-related breeding material. Theor Appl Genet 134: 1409–1422 (2021). doi: 10.1007/s00122-021-03779-1

ABBREVIATIONS

BLUP	best linear unbiased prediction
CMS	cytoplasmic male sterility
CV	cross-validation
DMC	dry matter content
DMY	dry matter yield
EN	elastic net
EU	European Union
G×E	genotype-by-environment
GBLUP	genomic best linear unbiased prediction
GCA	general combining ability
GEBV	genomic estimated breeding value
GS	genomic selection
GY	grain yield
HBLUP	hyperspectral best linear unbiased prediction
HTPP	high-throughput phenotyping platform
IR	infrared radiation
LAI	leaf area index
Lasso	least absolute shrinkage and selection operator
MAS	marker-assisted selection
mSR	modified simple ratio
NDVI	normalized difference vegetation index
PH	plant height
PLS	partial least squares
QTL	quantitative trait locus
RR	ridge regression
RR-BLUP	ridge regression best linear unbiased prediction
TRN	training set
UAV	uncrewed aerial vehicle
VAL	validation set
VI	vegetation index
VS	visible spectrum

1. General introduction

Renewable sources of energy are increasingly used worldwide, among which bioenergy, defined as the energy produced from biomass, is the largest (World Bioenergy Association 2019). In the European Union (EU), bioenergy should represent at least 20% of the energy demanded by the end of 2020 (European Union 2009). The biogas sector in the EU can contribute towards a low carbon economy and has undergone, consequently, an exponential growth mostly driven by the anaerobic fermentation of agricultural feedstocks (Calderón et al. 2019). In Germany, the largest European biogas producer, silage maize is by far the most important fermentation substrate (Fachagentur Nachwachsende Rohstoffe e.V. 2019). Improved agricultural practices (e.g. crop rotations) are required to enhance the savings of greenhouse gas emissions from soil carbon accumulation. Maize as the predominant crop should be particularly reduced in acreage and, for 2021, a maximum of 44% of maize is acceptable in the fermentation substrate by the German Renewable Energy Sources Act "Erneuerbare-Energien-Gesetz" (EEG 2012, 2017). Therefore, the rising needs for substitute sources of biomass represent an exciting opportunity for alternative dual-purpose crops adapted to the European agroclimatic conditions, like rye or triticale.

1.1. Rye – a versatile dual-purpose crop

Winter rye (*Secale cereale* L.) is a prominent cereal in Europe, with about 70% of the total global area (4.1 million hectares) found in Russia, Poland, Germany, Belarus, Fenno-Scandinavia, and Ukraine (FAO 2020). Unlike many other small-grain cereals, rye is an allogamous species owning an effective gametophytic self-incompatibility system and, therefore, a strictly out-crossing crop (Lundqvist 1956). This unique characteristic was traditionally used to release open-pollinated and synthetic cultivars; however, the performance of these population varieties was surpassed already by the first developed hybrids in the early 1980s (Geiger and Miedaner 1999, 2009, Miedaner and Laidig 2019). Hybrid breeding in rye started at the University of Hohenheim (Geiger and Miedaner 2009) after the discovery of both, a source of cytoplasmic-male sterility (CMS) called "Pampa" coming from an Argentinian landrace (Geiger and Schnell 1970) and dominant genes

allowing the restoration of pollen fertility (Geiger 1972). Additionally, the presence of well-established genetically distinct heterotic groups (Petkus and Carsten pools) allows to exploit heterosis for relevant economic traits; thus, hybrids emerged as the cultivars of choice by European farmers (Geiger and Miedaner 2009). For instance, 76% of the German rye acreage is covered by hybrid varieties (Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz 2019).

Rye outstands for its vigorous growth, high yield potential, superior adaptation to infertile, light, or acid soils, and high tolerance to abiotic and biotic stress factors, resulting in an attractive alternative not only under good agroclimatic conditions but also in vast regions less suitable for the cultivation of other cereals (Geiger and Miedaner 2009). Rye was traditionally used for bread making, however, by 2018 in Germany, animal feeding (52%) is the most important rye market followed by food (23%) and bioenergy production (18%; Bundesanstalt für Landwirtschaft und Ernährung 2019). Thus, rye has recently attracted increased attention as a source of renewables with reduced food-bioenergy trade-off, emerging biomass yield as a new target trait in rye breeding (Miedaner et al. 2010, Roux et al. 2010, Hübner et al. 2011, Haffke et al. 2014, Igos et al. 2016). Nevertheless, the incorporation of biomass as a key trait within breeding programs is challenging. The traditional and costly destructive sampling techniques commonly used (Catchpole and Wheeler 1992) prevent grain yield (GY) from being recorded in those same plots. Considering the major advancement in digital tools with breeding applications, the integration of state-of-the-art technologies in optimized breeding programs may contribute to enhance biomass yields affordably (Furbank et al. 2019).

1.2. Breeding rye for enhanced biomass yield

The response to selection, also known as "the breeder's equation", is a useful guide to estimate the achievable response to selection for a given trait, and therefore, represents a suitable framework to evaluate the prospects of breeding programs (Cobb et al. 2019). As stated in Falconer and Mackay (1996, p.189), this model can be written as

$$R = ih^2\sigma_p \quad (1)$$

where R is the response to selection, i is the intensity of selection (selection differential in phenotypic standard deviations), h^2 the heritability of the target trait, and σ_p is the standard deviation of the phenotypic values of individuals. The efficiency of R over a certain period can be estimated by dividing R by the length (in years) of the selection cycle (Eberhart 1970).

The enhancement of dry matter yield (DMY), which is highly correlated (r =0.95, P < 0.01) to methane yield (Hübner et al. 2011), is the most important trait for breeding hybrid rye as biogas substrate (Miedaner et al. 2010). Assessed among different elite rye populations, DMY showed large phenotypic as well as genotypic variations (Miedaner et al. 2010, Hübner et al. 2011, Haffke et al. 2014) and its heritability (h^2=0.49) was slightly lower than that observed for a mid-heritable trait such as GY (h^2=0.52; Haffke et al. 2014). Thus, the amount of heritable variability for DMY shows that an essential requirement for successfully breeding rye as a bioenergy crop is fulfilled (eq. 1). However, the favorable prospect for breeding biomass rye seems not to be in line with traditional breeding programs.

As described in Miedaner and Laidig (2019), a selection cycle in hybrid rye breeding begins with the crossing of parental lines followed by several selfings and the evaluation of their *per se* performance in one-row observation plots at two to three locations for highly heritable traits (e.g., flowering time, plant height (PH), disease and lodging resistances, and thousand-kernel weight). The selected parental components (S_2-lines) are then topcrossed with testers from the opposite heterotic pool under strictly isolated field plots to avoid contamination by foreign pollen. The obtained testcrosses are then subject to a two-stage evaluation of their general combining ability (GCA) on larger plots (5–6 m^2) at several locations. Selection at the first stage (GCA-1) is primarily driven by GY, although some other secondary traits such as PH, lodging resistance, and quality traits are also scored. In the following year, a selected fraction of the lines is reevaluated for GY at a second stage (GCA-2) with more testers at a higher number of locations. At this stage, DMY is usually incorporated as a breeding target and is traditionally assessed by destructive methods at the late milk stage (BBCH 77, Meier 1997), making it necessary to plant GCA-2 trials twice for the estimation of both GY and DMY. A new selection cycle is started each year with new parental components derived from multiple genetic backgrounds.

Consequently, the genetic variation present in the population, a crucial factor determining the expected response to selection (eq. 1), can be entirely exploited for the enhancement of

GY. In contrast, the selection for improved DMY is carried out in a later stage in a strongly reduced population, mainly due to the high costs associated with this traditional harvest technique for large populations (Haffke et al. 2014). However, the exceptionally high cost associated with duplicating early-stage field experiments (e.g., large-scale GCA-1 trials) would not be feasible due to constraints of market share and budget (Miedaner et al. 2012). Thus, the enhancement of DMY in rye relies during the first selection stages on an adequate utilization of indirect selection (Falconer and Mackay 1996). PH at the heading stage, whose correlation with DMY is almost double that of GY (r = 0.33 and 0.64, respectively, $P < 0.01$; Haffke et al. 2014), was suggested as a superior secondary trait than GY to indirectly select DMY. However, selection for lodging resistance is then highly important (Roux et al. 2010, Haffke et al. 2014). To establish an effective and affordable dual-purpose breeding program in rye emerges, therefore, as crucial to introduce superior indirect criteria to take advantage of the full genotypic variance present at early stages for DMY in a cost-effective manner.

Fig. 1 Biomass estimation in rye carried out by destructive means

1.3. Incorporation of molecular data into plant breeding programs

1.3.1. Marker-assisted selection

From the early 1980s until the turn of the millennium, the use of molecular data for accelerating the selection gain in plant breeding was fundamentally focused on identifying molecular markers associated with major-effect quantitative trait locus (QTL) for their further use in marker-assisted breeding (Xu and Crouch 2008, Ben-Ari and Lavi 2013). After more than 20 years of intensive research, marker-assisted selection (MAS; Beckmann and Soller 1986) emerged as an appropriate genetic tool for traits controlled by few, large effect QTLs but ineffective for polygenic traits, which are influenced by numerous small-effect QTLs (Bernardo 2008). The low effectiveness of MAS for complex traits is fundamentally explained by the potentially overestimated marker effects resulted from the required QTL identification and validation and the small amount of the genetic variation that may be explained by these preselected QTLs (Beavis 1998, Meuwissen et al. 2001, Schön et al. 2004). Since a complex genetic architecture determines most of the traits with agronomic relevance in plant breeding, the practical applicability of MAS has been greatly limited (Tsai et al. 2020). However, the combination of significant advancements in the fields of marker technologies and statistics has allowed to affordably and precisely score many thousands of markers evenly distributed throughout the genome and the estimation of their effects altogether, without prior bias-prone marker selection (Whittaker et al. 2000), opening new avenues to reduce the cost and cycle length of breeding for quantitatively inherited traits (Heffner et al. 2009, Heffner et al. 2010, Jannink et al. 2010).

1.3.2. Genomic selection

The genomic information available for rye has been extraordinarily increased by the advent of medium-density marker assays with multiple potential benefits for practical breeding, including the direct incorporation of genomic selection (GS; Meuwissen et al. 2001) into

practical breeding routines (Miedaner et al. 2019). GS exploits genomewide molecular markers underlying a quantitative trait of interest instead of targeting large-effect QTLs. Several studies have shown the superiority of GS over MAS for enhancing quantitatively inherited traits in plant breeding (Bernardo and Yu 2007, Wong and Bernardo 2008, Lorenzana and Bernardo 2009, Heffner et al. 2010, Lorenz et al. 2011). Similarly, GS surpassed the prediction ability of MAS for polygenic traits such as GY and GY-related traits also in rye when analyzed within two bi-parental populations (Wang et al. 2014, Wang et al. 2015).

In GS, a group of genotyped and phenotyped individuals (training set, TRN) is used to calibrate a prediction model for selecting among genotyped individuals lacking phenotypic values (validation set, VAL) based on genomic estimated breeding values (GEBVs; Heffner et al. 2009, Jannink et al. 2010, Lorenz et al. 2011). The prediction ability of GS, defined as the correlation between GEBVs and the observed phenotypic values, is then commonly assessed by cross-validation (CV) procedures (Hastie et al. 2009). In this context, abundant marker data (*p*) are available for estimating the performance of by far less numerous genotypes (*n*), an impossible situation for standard multiple linear models commonly referred to as "large *p*, small *n*" problem (Jannink et al. 2010).

Several statistical methods have been developed to overcome this limitation, among which the genomic best linear unbiased prediction (GBLUP; Habier et al. 2013) is one of the most commonly used (Isidro et al. 2015, Vieira et al. 2017). In GBLUP, GEBVs are directly estimated by using a so-called genomic relationship matrix that explains the genetic relationship among individuals based on dense marker data (e.g., single nucleotide polymorphisms) in place of the traditional pedigree-based selection models proposed in animal breeding (Henderson 1975). The main advantage of replacing the pedigree-based by a marker-estimated matrix is that the latter allows a more accurate estimation of the random segregation that constitute the Mendelian sampling effect (i.e., the unequal transmission of the parental genome) reducing the co-selection of sibs and the inbreeding rate per generation (Daetwyler et al. 2007, Heffner et al. 2009). The Mendelian segregation is a crucial factor explaining the genetic variation in additive models under the absence of inbreeding (Pérez et al. 2010, Burgueño et al. 2012). Consequently, for traits of economic interest in plant breeding, marker-based models are preferred over models based only on

pedigree data (de los Campos et al. 2009, de los Campos et al. 2010, Crossa et al. 2010, Crossa et al. 2011, Burgueño et al. 2012). As shown by Habier et al. (2007), GBLUP is equivalent to the ridge regression best linear unbiased prediction (RR-BLUP; Whittaker et al. 2000, Meuwissen et al. 2001), a well-established penalized regression model that simultaneously and equally shrinks all marker effects towards zero as it assumes that all variables have a mean of zero and the same variance.

The use of GS for predicting breeding values across selection cycles, i.e., when genotypes used for model training and validation derived from different cycles, is the main application of GS towards more efficient breeding programs (Miedaner et al. 2019). However, the ability of GS for predicting complex traits is affected by multiple factors, which have received much interest in animal and plant breeding studies, such as the genetic relatedness of individuals included in TRN and VAL (Habier et al. 2007, Wientjes et al. 2013, Crossa et al. 2014, Marulanda et al. 2015). The low genetic relatedness typically observed among selection cycles in rye breeding suggests that across-cycles prediction of DMY emerges as highly challenging based only on marker data. Moreover, field trials are usually conducted among several contrasting environments, a situation that needs to be properly modeled since GS is also influenced by genotype-by-environment (G×E) interactions (Piepho 2009, Heslot et al. 2014, Crossa et al. 2014). Furthermore, GS also depends on the TRN size (VanRaden et al. 2009, Lorenz 2013, Marulanda et al. 2015), an aspect of great influence on the time and capital expenditures required for biomass breeding. The heritability of the trait under study also determines the prospects of GS (Heffner et al. 2009, Marulanda et al. 2015). The ability of GS for traits showing low heritability can be improved when correlated highly heritable traits are incorporated into multivariate models (Jia and Jannink 2012).

1.4. High-throughput phenotyping as a valuable breeding tool

The progress in plant phenotyping has been much slower than the fast development of genotyping technologies, remaining the accurate phenotyping of large-scale multi-location field trials as a significant challenge for the advancement of genetic research (Montes et al. 2007, Furbank 2009, White et al. 2012, Araus and Cairns 2014). However, the latest improvements in image data acquisition and modeling, data mining, aeronautics, as well as

robotics, have increased the interest in crop “phenomics”, a scientific discipline aiming at the description of phenotypes based on the gathering of high-dimensional phenotypic data (Houle et al. 2010). High-throughput phenotyping platforms (HTPPs), including uncrewed aerial vehicles (UAVs), have emerged, therefore, as a suitable option to alleviate the phenotyping bottleneck in a resources-effective and time-saving manner, potentially leveraging the genetic gain in breeding programs (Furbank and Tester 2011, Araus et al. 2018). The non-invasive measurement of the radiation reflected or emitted by the plants constitute the basis underlying the wide range of uses of remote sensing in agriculture, which includes the assessment of several agronomic and physiological crop parameters (Atzberger 2013, Mulla 2013), including the high throughput and repeatable estimation of biomass (Hansen and Schjoerring 2003, Mutanga and Skidmore 2004, Jong et al. 2010, Busemeyer et al. 2013, Bendig et al. 2014, Prabhakara et al. 2015, Cheng et al. 2017, Yue et al. 2017, Zhang et al. 2017, Li et al. 2018, Han et al. 2019, Walter et al. 2019, Jin et al. 2020).

UAVs are increasingly used in plant breeding for phenotyping large-scale trials due to their lower capital and time requirements, larger flexibility, higher working capacity, and superior spatio-temporal resolution compared not only to destructive sampling but also to other remote sensing approaches, such as proximal sensing conducted with ground-based devices and satellite-based imagery (Tattaris et al. 2016, Yang et al. 2017). Moreover, the replacement of conventional digital cameras by hyperspectral devices has expanded further the applications of UAVs in crop sciences (Araus and Cairns 2014). In contrast to digital cameras which collect only the information present at the red, green, and blue channels within the visible spectrum (400 - 700 nm; VS), hyperspectral cameras are high-resolution sensors (up to <1 nm) capable of exploring and recording also infrared radiation (IR) wavelengths (up to 2500 under special configurations) in a continuous mode (Mahlein et al. 2012, Araus and Cairns 2014). Thus, UAVs equipped with hyperspectral sensors stand as a powerful technology for biomass estimation and other agricultural applications (Adão et al. 2017).

The processing of the data collected by HTPPs has increased its complexity in proportion to the higher resolution of these devices mainly due to greater data dimensionality (Fahlgren et al. 2015, Yang et al. 2017), potentially delaying the adoption of this technology for field phenotyping (Araus and Cairns 2014). The development of vegetation indices (VIs)

represents a common approach for a straightforward extraction of meaningful information from vast reflectance datasets for measuring major vegetation characteristics, including biomass, vegetation cover, leaf area index (LAI) as well as chlorophyll and water content (Xue and Su 2017). However, VIs are based on a very small fraction of the total hyperspectral information available, hindering their ability to characterize complex traits in detail (Pauli et al. 2016). Consequently, a substantial information loss may occur (Aguate et al. 2017). Several statistical procedures have been proposed for modeling hyperspectral data beyond the calculation of VIs, allowing the use of whole-spectrum data for a better estimation of key plant traits (Thorp et al. 2017, Araus et al. 2018). With the progress of "omics" technologies, massive datasets are available for crop studies, whose integration into predictive modeling creates a unique opportunity for advancements in plant breeding (Langridge and Fleury 2011). Consequently, combining data obtained from high-throughput genotyping and phenotyping into prediction modeling opens new avenues for improving the estimation of the genetic merit of unphenotyped individuals with multiple benefits for affordably breed dual-purpose rye.

Fig. 2 Hyperspectral sensor-bearing UAV for biomass estimation in large-scale rye field trials.

2. Objectives

The aim of this research thesis was to investigate the prospects of combining hyperspectral, genomic, and phenotypic data for unlocking the potential of hybrid rye as a dual-purpose crop to meet the increasing demand for renewable sources of energy affordably.

In particular, the objectives were to:

i. Estimate relevant population parameters for vegetation indices (VIs) and agronomic traits (grain and biomass yields, plant height, and thousand-kernel weight) as well as the correlation among them (Publication I);
ii. Assess the prediction ability for dry matter yield (DMY) within and among environments by including VIs and agronomic data as secondary traits in multiple linear regression models (Publication I);
iii. Integrate hyperspectral and genomic information as well as plant height into multi-kernel and bivariate models and compare their predictive power over single-kernel models across different training set (TRN) sizes (Publication II);
iv. Perform variable selection to identify the most informative spectral regions to DMY prediction in rye (Publication II);
v. Investigate the influence of the genetic and environmental relationships between TRN and validation set (VAL) as well as trait heritability on the prediction ability of genomic- and hyperspectral-based models (Publication III).

3. Publication I: Hyperspectral reflectance data and agronomic traits can predict biomass yield in winter rye hybrids

Rodrigo José Galán[1], Angela-Maria Bernal-Vasquez[2], Christian Jebsen[2], Hans-Peter Piepho[3], Patrick Thorwarth[1], Philipp Steffan[2], Andres Gordillo[2], Thomas Miedaner[1].

[1] State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany.

[2] KWS LOCHOW GMBH, Ferdinand-von-Lochow Straße 5, 29303 Bergen, Germany.

[3] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, 70593 Stuttgart, Germany.

BioEnergy Research (2020) 13: 168–182

The original publication is available at
https://doi.org/10.1007/s12155-019-10080-z

Abstract

Winter rye (*Secale cereale* L.), a potential alternative substrate for biogas production, is generally bred for grain yield. Thus, we aimed to evaluate the prospects of dry matter yield prediction by integrating vegetation indices derived from visible to NIR spectral data. A total of 404 elite rye hybrids were evaluated for grain yield and a subset of this comprising 274 hybrids were also assessed for dry matter yield over two years and at four locations in Germany (i.e., eight environments). Spectral reflectance data (410 to 993 nm) were collected around solar noontime on mostly clear sky by an uncrewed aerial vehicle (UAV) on two dates. Observed variation among tested hybrids ranged between 3.64-10.53 Mg ha^{-1} for grain yield and 8.44-14.66 Mg ha^{-1} for dry matter yield across different environments. The 23 vegetation indices and the agronomic traits, such as dry matter yield, grain yield, and plant height, showed mostly moderate to high heritability estimates (h^2>0.50), and their genetic variances were significantly (P<0.001) different from zero. Plant height was preferred over grain yield for indirect selection of high dry matter yield. An index combining hyperspectral and agronomic data developed by a multiple regression procedure showed a cross-validated prediction ability of 0.75, resulting in an improvement of about 6% to a model based only on agronomic traits. In earlier selection stages, the proposed index could be a suitable tool for the cost-effective selection of improved candidates for biomass experiments based on grain yield trials.

4. Publication II: Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye

Rodrigo José Galán[1], Angela-Maria Bernal-Vasquez[2], Christian Jebsen[2], Hans-Peter Piepho[3], Patrick Thorwarth[1], Philipp Steffan[2], Andres Gordillo[2], Thomas Miedaner[1].

[1] State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany.

[2] KWS LOCHOW GMBH, Ferdinand-von-Lochow Straße 5, 29303 Bergen, Germany.

[3] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, 70593 Stuttgart, Germany.

Theoretical and Applied Genetics (2020) 133: 3001–3015

The original publication is available at
https://doi.org/10.1007/s00122-020-03651-8

Theoretical and Applied Genetics (2020) 133:3001–3015
https://doi.org/10.1007/s00122-020-03651-8

ORIGINAL ARTICLE

Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye

Rodrigo José Galán[1] · Angela-Maria Bernal-Vasquez[2] · Christian Jebsen[2] · Hans-Peter Piepho[3] · Patrick Thorwarth[1,2] · Philipp Steffan[4] · Andres Gordillo[4] · Thomas Miedaner[1]

Received: 5 February 2020 / Accepted: 3 July 2020 / Published online: 17 July 2020

Abstract

Key message **Hyperspectral and genomic data are effective predictors of biomass yield in winter rye. Variable selection procedures can improve the informativeness of reflectance data.**

Abstract Integrating cutting-edge technologies is imperative to sustainably breed crops for a growing global population. To predict dry matter yield (DMY) in winter rye (*Secale cereale* L.), we tested single-kernel models based on genomic (GBLUP) and hyperspectral reflectance-derived (HBLUP) relationship matrices, a multi-kernel model combining both matrices and a bivariate model fitted with plant height as a secondary trait. In total, 274 elite rye lines were genotyped using a 10 k-SNP array and phenotyped as testcrosses for DMY and plant height at four locations in Germany in two years (eight environments). Spectral data consisted of 400 discrete narrow bands ranging between 410 and 993 nm collected by an unmanned aerial vehicle (UAV) on two dates on each environment. To reduce data dimensionality, variable selection of bands was performed, resulting in the least absolute shrinkage and selection operator (Lasso) as the best method in terms of predictive abilities. The mean heritability of reflectance data was moderate ($h^2 = 0.72$) and highly variable across the spectrum. Correlations between DMY and single bands were generally significant ($p < 0.05$) but low (≤ 0.29). Across environments and training set (TRN) sizes, the bivariate model showed the highest prediction abilities (0.56–0.75), followed by the multi-kernel (0.45–0.71) and single-kernel (0.33–0.61) models. With reduced TRN, HBLUP performed better than GBLUP. The HBLUP model fitted with a set of selected bands was preferred. Within and across environments, prediction abilities increased with larger TRN. Our results suggest that in the era of digital breeding, the integration of high-throughput phenotyping and genomic selection is a promising strategy to achieve superior selection gains in hybrid rye.

Introduction

The European biogas sector has attracted increasing attention as a renewable source of heat, electricity, and transport suitable for climate change mitigation with additional socioeconomic advantages (Scarlat et al. 2018). Political directives (European Renewable Energy Directive 2009/28/EC) supporting the production of bioenergy have already been implemented in Europe (European Union 2009). This legislation stated that, by 2020, the energy demanded in the European Union (EU) should be supplied in at least 20% by renewable sources. Among the EU members, therefore, the role of energy crops as bioenergy feedstocks has undergone a considerable increase, represented mainly by silage maize (European Commission 2018). Maize monocropping is, however, discouraged by regulations toward enhanced sustainability of the biomass production (European Union 2010).

Communicated by Susanne Dreisigacker.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00122-020-03651-8) contains supplementary material, which is available to authorized users.

✉ Thomas Miedaner
thomas.miedaner@uni-hohenheim.de

[1] State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany

[2] KWS SAAT SE, Grimsehlstraße 31, 37574 Einbeck, Germany

[3] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, 70593 Stuttgart, Germany

[4] KWS LOCHOW GMBH, Ferdinand-von-Lochow Straße 5, 29303 Bergen, Germany

Additionally, in Germany, the principal biogas producer in Europe, a limit was placed on the amount of maize acceptable in the fermentation substrate. In 2012, this limit was set to 60%, while in 2021, it will be reduced further to 44% (Renewable Energy Sources Act "EEG"; EEG 2012, 2017). Consequently, the growing demand for bioenergy combined with the search for alternative sources of biomass opens a very attractive opportunity for diversifying crop rotations.

Winter rye (*Secale cereale* L.) is a small-grain cereal with vigorous growth and enhanced tolerance to abiotic (e.g., low temperatures, light or acid soils with low fertility) and biotic stress factors. It can, therefore, be cultivated in vast areas less suited for other cereal crops (Geiger and Miedaner 2009), representing a sustainable biomass source with reduced competition with food or feed (Miedaner et al. 2012; Geiger and Miedaner 2009). Although it is present worldwide, rye is mostly grown in Northeastern Europe, where Germany, Poland, Russia, and Fennoscandia concentrate about 60% of the total area of rye cultivation (FAO 2019). Considering its potential as a dual-purpose crop, enhanced dry matter yield (DMY) has emerged as a new target in rye breeding, which has been primarily driven by grain yield GY (Haffke et al. 2014). In contrast to GY, which in our breeding program is already tested at the first year of general combining ability testing (GCA-1), DMY is traditionally evaluated through destructive methods at later selection stages on a strongly reduced set of genotypes. Thus, lower selection gains can be expected due to the loss of important genetic variation during the breeding process.

Efficient indirect selection for dry matter yield (DMY) would, therefore, be needed to exploit the full genetic variation present at early selection stages. Plant height (PH) has been identified as an indirect selection target for enhanced DMY, but biomass-specific trials with a particular focus on lodging resistance were still recommended (Roux et al. 2010; Haffke et al. 2014). Genomic selection (GS) (Meuwissen et al. 2001) aims to indirectly select unphenotyped candidates based on a model trained in a reduced set of genotyped and phenotyped entries (training set, TRN). Genomic tools have been proposed to increase the efficiency of selection in hybrid rye breeding (Miedaner et al. 2019). For instance, GS has been recommended for enhanced prediction of grain yield in rye across breeding cycles (Auinger et al. 2016; Bernal-Vasquez et al. 2017). Another study in rye showed that, in terms of prediction accuracy, GS was preferred to marker-assisted selection (MAS) in intra-pool crosses not only for GY but also for PH and quality traits (i.e., starch and pentosan content, Wang et al. 2014).

The development of molecular techniques has increased the needs of reliable and cost-effective phenotypic information, representing a great challenge for the progress of plant-genetic studies (Araus and Cairns 2014; Montes et al. 2007). High-throughput phenotyping (HTP) has emerged as a suitable strategy for phenotyping thousands of new genotypes effectively and affordably based on reflectance information (Furbank and Tester 2011; White et al. 2012). Unmanned aerial vehicles (UAVs) such as polycopters outperform ground-based HTP platforms regarding working capacity while deriving high-resolution image data (Araus and Cairns 2014). So, they may represent a suitable approach for screening multi-environment field trials, exponentially increasing the amount of data available. In this context, a positive impact on practical plant breeding may be expected if reflectance data are associated with the target trait (Rutkoski et al. 2016). This would be of great interest, for instance, to enhance indirect estimation of DMY within a breeding population at first stage of GY trials, when a direct assessment of the trait by destructive measures would not be feasible, but aiming for a dual-purpose program with genotypes being superior for both DMY and GY.

Hyperspectral sensors deliver information of hundreds of wavelengths (hereafter referred as "bands") at a nanometer-level resolution covering a broad spectral range (from 350 up to 2500 nm) that includes the visible spectrum (VS) and the infrared (IR) regions (Mahlein et al. 2012). This imaging technique is a promising tool for field phenotyping but presents additional computation efforts due to the increased data dimensionality (Fahlgren et al. 2015). To address this issue, several strategies have been proposed for integrating reflectance data into practical plant breeding. One approach is to summarize a few individual bands into vegetation indices (VIs; Xue and Su 2017; Galán et al. 2020). However, prediction accuracy of VIs was found to be lower than equations incorporating whole-spectrum data by ordinary least squares (OLS), partial least squares (PLS), and Bayesian shrinkage for GY prediction in maize (Aguate et al. 2017) and by Bayesian functional models in wheat (Montesinos-López et al. 2017).

In both studies, models were tested under $p < n$ scenarios, where the number of predictors (p) was smaller than the population size (n). On the contrary, when $p >> n$ as in GS, regularization (penalized) models have shown to be suitable for incorporating thousands of predictors, including several unrelated to the trait of interest, or highly intercorrelated (Ogutu et al. 2012). A similar situation may be expected when analyzing hyperspectral data collected in several environments and on several dates. To reduce multicollinearity, increase prediction accuracy, minimize calculation time, and extract the most informative features, regularization methods such as the elastic net (Zou and Hastie 2005) or the least absolute shrinkage and selection operator (Lasso; Tibshirani 1996) are also preferred for facing high-dimensional spectral data (Liu and Li 2017).

Alternatively, Krause et al. (2019) found that deriving relationship matrices from hyperspectral data was a suitable approach to integrate whole-spectrum reflectance

information into multi-kernel GS for predicting GY in wheat within multi-environment field trials. Multivariate models integrating correlated traits have demonstrated to be more precise than univariate models in GS (Jia and Jannink 2012). In wheat, for instance, GS prediction ability of GY was significantly enhanced by fitting traits derived from hyperspectral data (Sun et al. 2019; Rutkoski et al. 2016; Crain et al. 2018).

Similar to GS, models seeking the estimation of breeding values utilizing hyperspectral information also need phenotypic data (e.g., DMY) for model training. In our study case, the TRN size is economically highly relevant, since the acquisition of the phenotypic data requires to evaluate the candidates in GY-plots and DMY-plots separately under the conditions of a dual-purpose breeding program. The positive relationship between GS accuracy and TRN size is widely known (VanRaden et al. 2009). However, a broader TRN represents an increase in breeding costs. Thus, efficient breeding programs would benefit from reduced TRN while maintaining, or at least minimizing the loss of prediction accuracy. Approaches to enable the highest accuracy for a reduced TRN by integrating phenotypic and hyperspectral information to GS are, therefore, highly relevant for delivering high-yielding DMY varieties.

The aim of the present study was to test these approaches within the same breeding population by evaluating a set of 274 elite rye lines as a testcross series in multi-environment field trials on a phenotypic, genotypic, and hyperspectral level. In particular, the objectives were (1) to identify the most relevant spectral regions to DMY prediction in rye, (2) to integrate the different sources of information into multi-kernel and bivariate models for leveraging selection gain of DMY in rye, and (3) to compare prediction ability of models across different TRN sizes.

Materials and methods

Plant materials and field experiments

The plant materials and field experiments analyzed in the present study are described in detail in Galán et al. (2020). In short, a total of 264 recombinant inbred lines (RILs) of generation S_4 (i.e., lines after continued self-fertilization of single plants for four consecutive years) were derived from ten diverse parental lines of the Petkus (seed parent) gene pool following a single round-robin design (Verhoeven et al. 2006). In practical plant breeding, these parental lines represent elite breeding material, since in contrast to a diverse panel of genetic resources, they were obtained after several selection cycles for line per se performance and general combining ability (GCA). Testcross seed was produced from the cross of these 264 RILs and their ten parental components with a single-cross tester from the opposite (pollinator) gene pool. The obtained 274 genotypes, thus, correspond to three-way hybrids, $(A\cdot B)\times C$. They were analyzed for their dry matter yield (DMY) and plant height (PH) in two trials with a size of 130 and 134 entries, respectively, laid out as resolvable incomplete block designs (α-lattice design) with two replicates. These field trials were grown adjacent to each other and conducted in 2017 and 2018 at each of four environmentally contrasting locations in Northern Germany (Suppl. Table 1), thus comprising eight environments (location–year combinations). Plots were harvested by a commercial plot chopper at late milk stage (BBCH 77; Meier 1997) to get the respective yield per plot as fresh matter yield (FMY, dt ha $^{-1}$). For DMY (dt ha $^{-1}$) determination, representative samples of about 1000 g were weighted from each plot and oven-dried at 110 °C till a constant weight was reached. Dry matter content (DMC) in percentage was determined from weight differences of the samples. DMY per plot was estimated as $DMY = FMY \times DMC/100$. Also, PH (cm) was recorded at each plot.

Hyperspectral data

Hyperspectral data consisting of 400 bands ranging from 410 to 993 nm were obtained in all environments and for all genotypes by an unmanned aerial vehicle (UAV; Camflight FX8HL, Sandnes, Norway) that was fitted with a hyperspectral camera (HySpex Mjolnir V-1240, Skedsmokorset, Norway) as described previously in detail Galán et al. (2020). Reflectance data were recorded after flowering (i.e., during the grain filling stage) at two flight dates in each environment, except for location Bernburg in 2017 (BBG 2017) where only one flight was conducted (Suppl. Table 1). On each flight date, the UAV quadcopter flew at about 25 m above plots, around solar noontime. Each plot was demarcated on the obtained images by a polygon, provided by digital geographic information system (GIS) field plans. Raw data were radiometrically calibrated (HySpex PostProcessor Version 1.2). This is a hyperspectral standard procedure (Adão et al. 2017) to convert the arbitrary digital numbers to values, which are proportional to the International System of Units (SI) unit W/sr nm m^2 (HySpex Mjolnir-1024 User's Manual). Coefficients of incident sunlight were captured by placing a 70-by-150 cm wooden board painted gray in the center of the field and using it as a reference to account for different irradiance conditions at each data collection time. The chosen gray panel reflects 60% of incident sun light, minimizing the risk of oversaturation of the hyperspectral sensor under varying sunlight conditions. The spectrum from the gray reflection target was assumed to represent the maximal reflection for each wavelength derived from sunlight. Normalized hyperspectral data (NormHyp) were then estimated based on this spectrum according to the formula

Normhyp = Hyperspectral reflectance/Gray panel spectrum. Further, hyperspectral imaging data were orthorectified and georeferenced via the PARGE Software (ReSe Applications LLC, Wil, Switzerland).

Finally, all data points per each wavelength within each polygon were spatially averaged, resulting in one spectrum per plot. Consequently, each plot contains a single value for each wavelength in the studied spectrum. A tabular data frame was constructed, including the computed reflectance values of all bands.

Genotypic data

All 274 genotypes (264 RILs and their ten parental components) were genotyped with an Illumina INFINIUM chip with 9,963 single-nucleotide polymorphisms (SNPs) assays (KWS SAAT SE & Co. KG, Einbeck, Germany). The SNPs of this assay are partially overlapping with the 5 k-SNP assay of Martis et al. (2013) and the 600 k-SNP assay of Bauer et al. (2017), whereof 3017 markers have been previously mapped by Bauer et al. (2017). SNPs showing more than 10% of missing values or a minor allele frequency < 0.05 were excluded. Imputation of the missing values in the remaining set of SNPs was performed with Linkimpute (Money et al. 2015). After imputation, data were filtered again for low minor allele frequency (< 0.05). Thus, 6420 markers were retained for subsequent analyses.

Phenotypic data analysis

Within and across environments, phenotypic data (i.e., DMY and PH) were analyzed by different mixed models to obtain variance components and BLUEs (best linear unbiased estimators) of genotypes for later use in prediction modeling.

A combined analysis across locations and years was conducted by applying the following mixed model:

$$\begin{aligned}\gamma = {} & G : L + Y \\ & + L \cdot G + Y \cdot G + Y \cdot L + L \cdot Y \cdot G \\ & + \mathrm{ENV} \cdot T + \mathrm{ENV} \cdot T \cdot R + \mathrm{ENV} \cdot T \cdot R \cdot B + e\end{aligned} \quad (1)$$

where γ denotes the observed genotype performance, G the genotypes,L the locations, Y the years, T the trials within environments ENV (equivalent to year–location combinations), R the replicates within trials, B the blocks within replicates, and e the error associated with the observation γ. Error, trial, block, and replicate variances were assumed heterogeneous among environments. In model (1), the dot operator (·) specifies crossed effects (A·B) and fixed and random terms are separated by a colon (:), with fixed terms appearing first (Piepho et al. 2003). Variance components and pairwise variances of genotype mean (BLUEs) differences (needed for heritability estimation) were estimated by restricted maximum likelihood (REML) for all random effects in model (1). This also holds for estimation of the genotypic variance (σ_g^2), which required an additional analysis fitting the above model with random genotypic effects. Significance of variance component estimates was tested by model comparisons using likelihood ratio tests (Stram and Lee 1994).

BLUEs of genotypes were also analyzed within environments by the following mixed model:

$$\gamma = G : T + T \cdot R + T \cdot R \cdot B + e \quad (2)$$

This model (2) differs from the first model (1) only in dropping the year and location main effects and corresponding interactions with genotypes. Variance components for single environments were estimated as described previously for model (1). Phenotypic outliers were tested for DMY and PH based on the Bonferroni–Holm test (method "M4r"; Bernal-Vasquez et al. 2016). Plots flagged as outliers were excluded from the analysis. Hyperspectral information was excluded from plots flagged as an outlier for DMY.

Three-stage analysis for DMY prediction

To reduce computing cost, prediction ability of DMY based on different information sources was conducted by a three-stage procedure (Piepho et al. 2012), where in the first two stages, hyperspectral data were analyzed across dates and environments to obtain BLUEs per genotype, which were then incorporated into DMY prediction models in the last stage.

First-stage models

In the first stage, hyperspectral bands were adjusted across dates per environment according to the model

$$\begin{aligned}\gamma = {} & G : D + D \cdot G \\ & + T + T \cdot R + T \cdot R \cdot B \\ & + D \cdot T + D \cdot T \cdot R + D \cdot T \cdot R \cdot B + e\end{aligned} \quad (3)$$

where γ is the observed band value, G the genotypes, D the measurement dates, T the trials, R the replicates within trials, B the blocks within replicates, and e the error associated with the observation γ. Errors of different measurement dates on the same plot are correlated; therefore, a correlation structure ("Compound Symmetry") was assumed for e as described in Piepho et al. (2004). This model was used here because there were only two measurement dates per environment. The random effects for trials, replicates, and blocks also imply a compound symmetry variance–covariance structure for repeated observations on these units. For BLUEs estimation, all factors included in model (1) except G were considered as random. For single bands in each flight date ("first" and "second"), the random effects

of the date, including the corresponding interaction terms, were excluded from model (3). To allow a fair comparison between across and within flight dates, data collected in BBG (2017), where only one flight was conducted, were included in both single-date and across-dates analyses.

Second-stage models

In the second stage, variance components and BLUEs per genotypes were estimated across environments following the model

$$\gamma = G : \mathrm{ENV} + G \cdot \mathrm{ENV} + e \quad (4)$$

where γ is the adjusted genotype mean (BLUEs) from the first stage for the band value, G and ENV denote genotypes and environments, respectively, and e is the error associated with the observation γ. When adjusted means from the first stage are forwarded to second-stage models, the incorporation of a weighting method is preferable (Möhring and Piepho 2009). Means were therefore weighted by the diagonal elements of the inverse of their variance–covariance matrix calculated in the first stage as proposed by Smith et al. (2001). For hyperspectral data, estimates of variance components, pairwise variances of genotype mean differences (BLUEs) as well as significance tests of variance components were computed as for the phenotypic data. The syntax of models (1), (2), (3), and (4) is also compatible.

At this stage, heritability (h^2) was estimated for DMY, PH, and each band for single and for combined flight dates across environments as (Piepho and Möhring 2007)

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\bar{v}}{2}} \quad (5)$$

where $\bar{v}$ is the mean variance of a difference between two adjusted genotype means (BLUEs) derived from model (1) or from model (4) for phenotypic and hyperspectral data, respectively. All statistical analyses were performed within the R-environment v. 3.4.4 (R Core Team 2018). BLUEs of genotypes were calculated with the software package ASReml-R v. 3.0 (Gilmour et al. 2009).

Third-stage models

In the third stage, the obtained phenotypic and hyperspectral BLUEs from model (1 or 2) and (3 or 4), respectively, were used for fitting several models (described in Table 1) for predicting DMY, including genetic, hyperspectral, and phenotypic data. A weighting method was applied also on this stage as described before, with weights derived from models (1) or (2).

Table 1 Overview over the models used

Model	Integrated variables
Single-kernel models	
GBLUP	Genotypic data
HBLUP	Hyperspectral data
Multi-kernel model	
G + H	Genotypic + hyperspectral data
Bivariate models	
Bivariate_G	Genotypic data + plant height
Bivariate_H	Hyperspectral data + plant height
Bivariate_G + H	Genotypic + hyperspectral data + plant height

The predictive power of these models was assessed in two different scenarios: (S1) across the series of eight environments by cross-validation (CV) and (S2) by fitting prediction models with data collected on a variable number of environments ($E = 1,2,\ldots,7$), while one environment not included in E was used for model validation. Coefficients of phenotypic correlation r (Pearson's coefficients of correlation) between DMY and all other traits were calculated from the BLUEs of genotypes from model (1) or (2) for prediction scenarios S1 or S2, respectively.

Third-stage models were single-kernel and multi-kernel prediction models, providing best linear unbiased predictions (BLUP) of genotypic effects of DMY, which differ in the information used to model the random genotypic effect. Single-kernel prediction models were fitted with genetic (genomic BLUP, GBLUP) or hyperspectral (hyperspectral BLUP, HBLUP) information with $n = 274$ individuals, based on m SNP markers or b bands, respectively. Thus, genomic estimated breeding values (GEBVs) were derived from the GBLUP model, whereas hyperspectral estimated breeding values (HEBVs) were obtained from the HBLUP model.

The two models were defined as

$$\mathrm{GBLUP} : \boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{g}_K + \boldsymbol{e}, \quad (6)$$

$$\mathrm{HBLUP} : \boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{g}_H + \boldsymbol{e}, \quad (7)$$

where $\boldsymbol{y}$ is the n-dimensional vector of BLUEs of DMY obtained from model (1) or model (2) for prediction scenarios S1 or S2, respectively, μ is the overall mean, 1_n an n-dimensional vector of ones, $\boldsymbol{g}_K$ and $\boldsymbol{g}_H$ are n-dimensional vectors of random genotypic effects, and $\boldsymbol{e}$ is the n-dimensional vector of residuals. The vector of residuals e associated with $\boldsymbol{y}$ was assumed as normally distributed with zero mean and variance $\mathbf{R}$ [$\boldsymbol{e} \sim \mathrm{N}(0, \mathbf{R})$]. $\mathbf{R}$ is defined as a diagonal matrix with diagonal elements equivalent to the inverses of the diagonal elements of inverse of the original variance–covariance matrix of the means adjusted on

the second stage of this analysis (Smith et al. 2001). When means adjusted in the second stage are forwarded to third-stage models, the incorporation of a weighting method was performed as described before.

For GBLUP, the random genetic values were estimated as $g_K \sim N(0, \mathbf{G}\,\sigma_g^2)$ where σ_g^2 is the genetic variance and **G** the genomic additive relationship matrix (Habier et al. 2013). For estimating genotypic values based on hyperspectral data, the random genetic values in model 7 were calculated as $g_H \sim N(0, \mathbf{H}\,\sigma_b^2)$ where σ_b^2 is the hyperspectral band variance and **H** a hyperspectral reflectance-based relationship matrix.

G was estimated with the synbreed package (Wimmer et al. 2012) in R following the first method of VanRaden (VanRaden 2008) as $G = \frac{ZZ'}{2\sum p_i(1-p_i)}$, where Z = M-P, M is the $n \times m$ marker matrix of alleles coded as 0 (A_1A_1), 1 (A_1A_2), or 2 (A_2A_2) for the nth individual at the mth SNP position, P contains a $n \times m$ matrix of allele frequencies multiplied by 2, p_i is the allele frequency of the ith allele.

H was also calculated for the $n = 274$ genotypes by incorporating the BLUEs for each band derived from model (4) or (3) for prediction scenarios S1 and S2, respectively. These matrices were of the form $H = DD'$, where D is a $n \times b$ hyperspectral matrix of the standardized BLUEs of the bands. Standardization was done by subtracting the arithmetic mean and dividing by the standard deviation of all BLUEs. For **H** estimation, different numbers of bands were considered: $\mathbf{H_{all}}$ is derived from the total number of bands available ($b = 400$), whereas $\mathbf{H_{vsel}}$ ($b = 32$) and $\mathbf{H_{h2}}$ ($b = 216$) are based on a reduced set of bands. Bands included in $\mathbf{H_{vsel}}$ were selected as described in the next sections, while $\mathbf{H_{h2}}$ is based only on bands with h^2 larger than the mean value observed for all bands ($h^2 > 0.72$).

Finally, a multi-kernel prediction model combining genetic and hyperspectral information was fitted:

$$y = \mu 1_n + g_K + g_H + e, \tag{8}$$

where all factors listed are defined as above in models (6) and (7). The random vectors g_K and g_H in (8) are considered as independent of each other and normally distributed. Here, the **H** matrix assumes the form of $\mathbf{H_{vsel}}$. For exploring the benefits of incorporating PH as a predictor, model (9) was extended to a bivariate model (Bivariate_G + H) as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1_n 0_n \\ 0_n 1_n \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} g_{K1} \\ g_{K2} \end{bmatrix} + \begin{bmatrix} g_{H1} \\ g_{H2} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{9}$$

where y_1 is a vector of BLUEs for DMY, y_2 is a vector of BLUEs for PH, with y_1 and y_2 incorporating BLUEs derived from model (1) for prediction scenario S1, μ_1 is the overall mean for DMY, μ_2 is the overall mean for PH, g_{k1} and g_{H1} are n-dimensional vectors of random effects for DMY, g_{K2} and g_{H2} are n-dimensional vectors of random effects for PH, e_1 is the n-dimensional vector of residuals for DMY, and e_2 is the n-dimensional vector of residuals for PH. The random vectors are considered as independent of each other and normally distributed according to $\begin{bmatrix} g_{K1} \\ g_{K2} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{C_K} \otimes \mathbf{G}), \begin{bmatrix} g_{H1} \\ g_{H2} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{C_H} \otimes \mathbf{H})$, and $\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \sim N(0, \mathbf{R} \otimes \mathbf{I})$, where **G** is defined as in model (4), $\otimes$ is the Kronecker product (direct product) operator, $\mathbf{C_K}$ and $\mathbf{C_H}$ are the 2×2 variance–covariance matrices for the breeding values of the two traits**,** **H** is defined as in model (7) and adopts the form of $\mathbf{H_{vsel}}$, **I** is an identity matrix, and **R** is the residual variance–covariance matrix for DMY and PH. The covariance matrices $\mathbf{C_K}$**,** $\mathbf{C_H}$**,** and **R** were considered unstructured. At this stage, model (9) was fitted without a weighting method to reduce computing costs. Bivariate_G + H aims to predict DMY based on PH as well as hyperspectral and genetic data. For addressing the impact of PH on the predictive power of bivariate models based only on hyperspectral (Bivariate_H) or genetic (Bivariate_G) data, two additional bivariate models were analyzed. These two models are a reduced version of model (9). For models Bivariate_H and Bivariate_G, the terms.

$\begin{bmatrix} g_{K1} \\ g_{K2} \end{bmatrix}$ or $\begin{bmatrix} g_{H1} \\ g_{H2} \end{bmatrix}$ were dropped, respectively. All three-stage prediction models were fit using the R package "sommer" (Covarrubias-Pazaran 2016).

Feature selection for the hyperspectral data

Multicollinearity in regression equations is expected when numerous highly intercorrelated hyperspectral variables are incorporated (Dunagan et al. 2007). To overcome this, two variable selection methods were used and implemented in the GlmNet R package (Friedman et al. 2010). Since weighted and unweighted variable selection procedures yielded similar results, we performed the following methods without the incorporation of a weighting factor.

The least absolute shrinkage and selection operator (Lasso; Tibshirani 1996) is a well-known and powerful regression method for regularization and variable selection for minimizing the prediction error. Applying the $l1$ penalty sets some of the regression coefficients to zero, while others are shrunk toward zero yielding a sparse solution. The Lasso should, however, be used with care in the case of sets of highly correlated variables since it tends to arbitrarily select one variable and overlook the rest (Friedman et al. 2010).

The elastic net (EN; Zou and Hastie 2005) was developed to overcome the restrictions of Lasso. It combines both $l1$ (Lasso) and $l2$ (Ridge Regression, Hoerl and Kennard 1970) penalization terms to obtain a more stable solution to highly correlated predictors.

The estimators ($\hat{\beta}$) for Lasso and EN can be calculated from the following penalized equation (Wimmer et al. 2013):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \; \| \boldsymbol{\gamma} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 + \text{Pen}(\beta) \quad (10)$$

where $\boldsymbol{\gamma}$ is defined as in model (4), $\boldsymbol{X}$ is a $n \times b$ matrix of bands; $\boldsymbol{\beta}$ is the vector of the regression coefficients of the bands; Pen($\boldsymbol{\beta}$) is the penalization term, which is defined by the quadratic $l2$ norm for RR as $\text{Pen}(\boldsymbol{\beta}) = \lambda \| \boldsymbol{\beta} \|_2^2 = \lambda \sum_{j=1}^{p} \boldsymbol{\beta}_j^2$, by the $l1$ norm for Lasso with $\text{Pen}(\boldsymbol{\beta}) = \lambda \| \boldsymbol{\beta} \|_1 = \lambda \sum_{j=1}^{p} \left| \boldsymbol{\beta}_j \right|$, and for EN by combining both as $\text{Pen}(\boldsymbol{\beta}) = \lambda_1 \boldsymbol{\beta}_1 + \lambda_2 \boldsymbol{\beta}_2^2$. For EN, the procedure can be described as a penalized least square method with $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$; thus, Eq. (10) is equivalent to the optimization problem $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \; \| \boldsymbol{\gamma} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2$, subject to $P_\alpha(\boldsymbol{\beta}) = (1-\alpha)\boldsymbol{\beta}_1 + \alpha \boldsymbol{\beta}_2^2 \leq s$ for some s.

For fitting and comparing the Lasso and EN models, the optimal values for the tuning parameter ($\lambda \geq 0$), which control the degree of shrinkage of the estimator, were obtained by tenfold cross-validation with the function cv.glmnet of the GlmNet R package (Friedman et al. 2010) with default settings. In addition, for the defined optimal λ, the best value for α for the EN was estimated outside the GlmNet package by a tenfold cross-validation.

Validation of variable selection procedures and proposed prediction models

In the present study, two prediction scenarios were considered, namely S1 and S2. A fivefold cross-validation (CV) was used to assess the predictive ability of models in S1, where models were fitted to fourfold (~ 219 genotypes), and model error was estimated when predicting the remaining validation fold (~ 55 genotypes). This was conducted for all five possible validation folds, and the obtained estimates of prediction error were combined. This procedure was repeated 100 times (i.e., 500 cross-validations), each repetition with a random composition of folds to assess CV error.

To investigate the effect of the TRN size on the prediction ability of all models in scenario S1, TRN was sampled according to a defined size (i.e., 55, 110, 165, or 220 individuals) and the validation set (VAL) consisted on the remaining genotypes. As described before, models were fitted to the TRN and model error was determined when predicting the VAL. This process, including the random sampling of the TRN, was repeated 500 times. For the larger TRN size, the prediction models were further evaluated. This procedure consisted of extracting, at each CV iteration, the predicted best yielding genotypes ranked above certain thresholds (10, 20, 30, and 40%). Then, the performance of the selected fraction was assessed in terms of its observed DMY and PH according to the BLUEs derived from model (1). Finally, the prediction ability of each model for each selected fraction was estimated as described below (see suppl. Table S4).

In scenario S2, HBLUP fitted with $\mathbf{H}_{\mathbf{all}}$ was tested across all possible combinations between E and validation environments. Also, the environmental distinctiveness was assessed by the discriminant analysis of principal components DAPC (Jombart et al. 2010) using the R package adegenet (Jombart 2008) based on hyperspectral BLUEs derived from model (3).

For all validation approaches, prediction ability for DMY was assessed as the correlation r between estimated breeding values and the observed BLUEs derived from model (1) for S1 and from model (2) for S2. Predictive abilities of bivariate models were estimated based on PH, hyperspectral, and genetic data (for Bivariate_H and Bivariate_G, only the corresponding data were included), whereas DMY was additionally used only for model training. Mean prediction abilities were compared according to Tukey's honestly significant difference (HSD) test ($p < 0.01$) with the R package multcomp (Hothorn et al. 2016). For Lasso and EN, each predictor (band), whose regression coefficient was not set to zero ($\hat{\beta} \neq 0$), was extracted and saved in a tabular form. Across variable selection runs, bands retaining > 40% of the time were considered as selected (recovery rate). The regularization method with the highest prediction ability based on the smallest number of selected bands was considered as the best procedure for reducing multicollinearity in the hyperspectral data. For $\mathbf{H}_{\mathbf{vsel}}$ estimation, selected bands derived from the best regularization scheme were used.

Results

In the present study, hyperspectral data were collected by two different flights performed after the heading stage, which were analyzed both individually and jointly. For all the issues under analysis, similar trends with no major contradictions could be observed, regardless of the number of flights considered. Therefore, the following sections are based on the joint analysis of both flight dates. The main results of the adjustment of individual flights can be found in the supplementary files (Suppl. Fig. S2, Suppl. Fig. S5, Suppl. Table S3).

Heritability and correlation estimates

Across eight environments and two flight dates, the mean heritability of the reflectance data was moderate ($h^2 = 0.72$, Fig. 1). VS bands had mostly higher estimates than those from the IR. Generally, h^2 decreased in the VS with higher

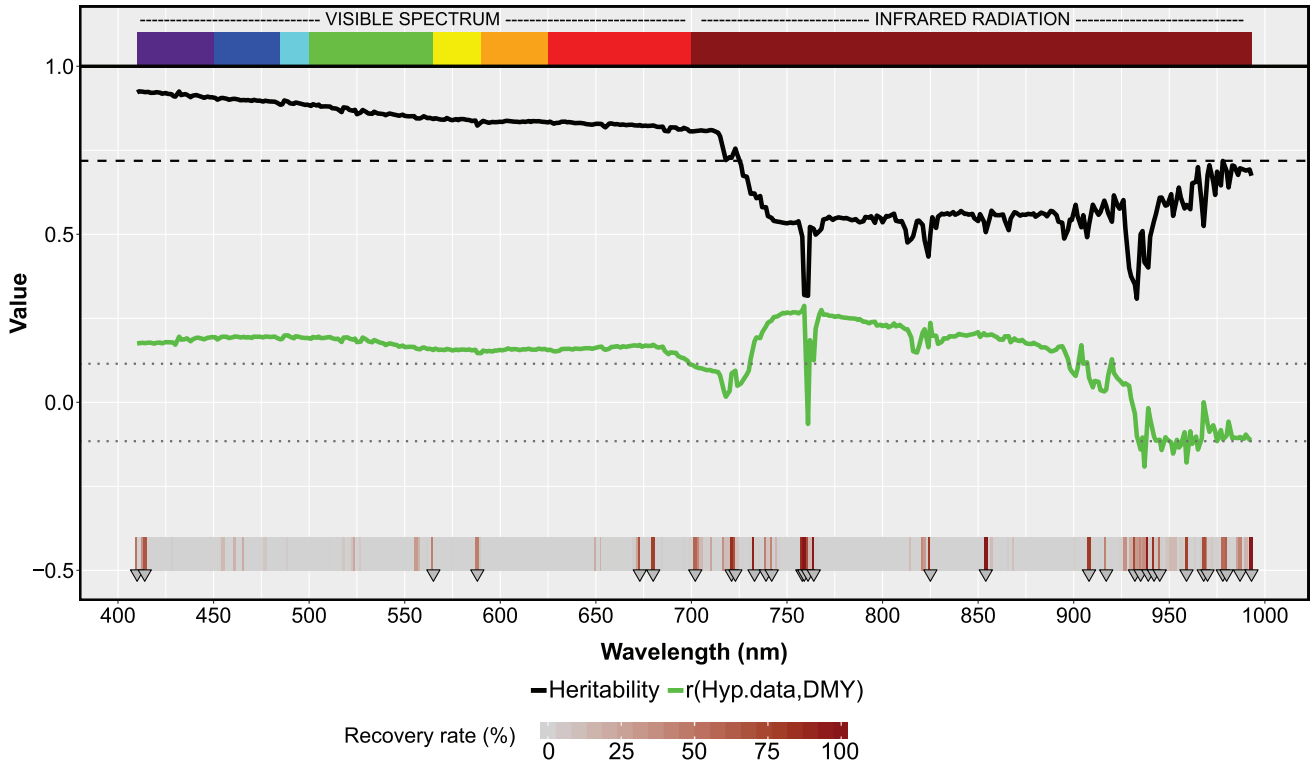


Fig. 1 Heritability estimates (black line) for the hyperspectral bands, phenotypic correlations (r, green line) between hyperspectral bands and dry matter yield, and recovery rate (%) of hyperspectral bands after the least absolute shrinkage and selection operator (Lasso, gray-red heatmap) for 274 winter rye hybrids assessed in eight environments and two flight dates. The mean heritability across all wavelengths is denoted by the dashed black line. Correlation values $\geq |0.12|$ are significant ($p < 0.05$) as shown by the gray dotted lines. Selected hyperspectral bands (recovery rate > 40%) are indicated by the gray triangles (Lasso variable selection)

wavelength, while the opposite was observed for the IR. Estimates were highly variable among the whole spectrum (from 0.31 to 0.92), especially in the red edge region (~ 720–750 nm), wherein about 30 nm, h^2 dropped from 0.73 (720 nm) to 0.32 (761 nm). Also, DMY and PH were analyzed in the present study and showed moderate ($h^2 = 0.50$ for DMY) to high ($h^2 = 0.82$ for PH) estimates (Fig. 2).

The magnitudes of the correlations involving DMY were higher for PH ($r = 0.57$, $p < 0.001$, Fig. 2) than for each of the 400 bands. Between DMY and the hyperspectral data, r ranged from −0.19 for bands around 930 to 0.29 nm for bands around 750 nm. Estimates $\geq |0.12|$ were significant at the 5% probability level. The mean correlation among bands in the VS was slightly higher than the observed for the IR (0.17 and 0.11, respectively). On the other hand, bands were highly intercorrelated. Bands within the VS as well as within the IR were highly positively intercorrelated (Suppl. Fig. S1). In contrast, correlations between both regions were highly negative. Interestingly, r was very low between a small group of bands from the red edge region and the rest of the spectrum.

Feature selection for the hyperspectral reflectance data

The two regularization methods (Lasso and EN) applied to the hyperspectral data performed similarly when predicting DMY ($r = 0.54$, Suppl. Fig. S3). However, they were based on a different number of selected variables (Suppl. Fig. S4). From the total 400 available bands, only 32 (~ 8%, Suppl. Table S2) and 54 (~ 13%) bands were selected by Lasso and EN, respectively. EN selected more bands than Lasso; however, all chosen bands by Lasso were also included in the EN selection (Suppl. Fig. S4).Thus, Lasso emerges as the procedure of choice for the present study because it yielded the same predictive power as EN but is based on a simpler prediction model. From the 32 selected bands by Lasso, 26 corresponded to the IR and only six to the VS (Fig. 1). These 26 bands were mostly located at both ends of IR (700–780 nm and 925–993 nm). Selected bands for the individual flight dates can be also found in Suppl. Fig. S4.

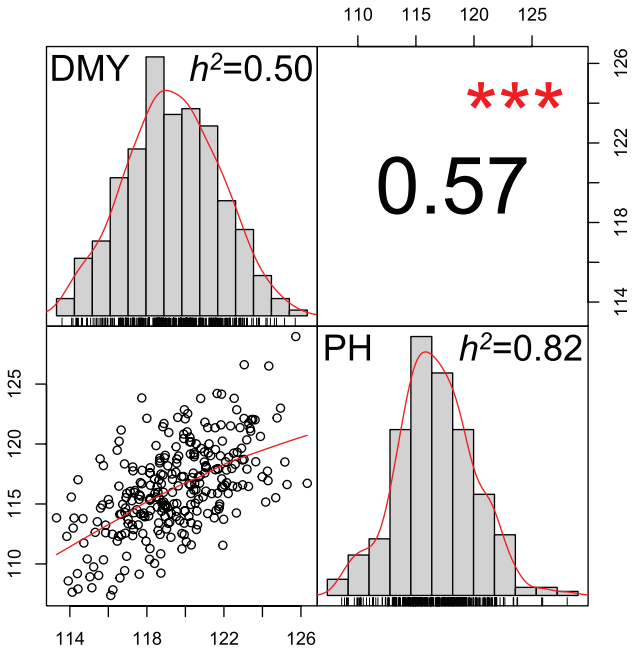


Fig. 2 Histograms of dry matter yield (DMY) and plant height (PH) as well as the phenotypic correlation between both traits, determined for 274 winter rye hybrids assessed in eight environments. h^2 shows the heritability estimates of both traits. ***Significant at the 0.001 probability level

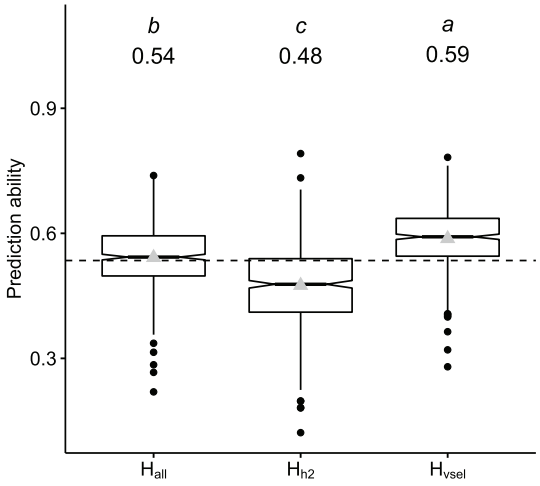


Fig. 3 Prediction ability for dry matter yield of hyperspectral best linear unbiased predictor model (HBLUP) based on different **H** relationship matrices, including all available 400 bands ($\mathbf{H_{all}}$), bands with heritability > 0.72, ($\mathbf{H_{h2}}$), and only selected bands by Lasso ($\mathbf{H_{vsel}}$) for 274 winter rye hybrids. Mean values are shown above each box plot and by gray triangles and are significantly different when headed by no letter in common (Tukey's honestly significant difference test; $\alpha = 0.01\%$). The dashed line shows the mean value across models

Prediction abilities of models

Two key factors largely affecting the accuracy of prediction models based on reflectance data were investigated, namely the composition of the **H** relationship matrix and the TRN size. For addressing the first factor affecting HBLUP predictive power, three HBLUP models based on dissimilar **H** relationship matrices ($\mathbf{H_{all}}$, $\mathbf{H_{h2}}$, and $\mathbf{H_{vsel}}$) were evaluated across the series of environments (Fig. 3, Suppl. Fig. S5). Thus, models differed in their number and composition of incorporated bands. In terms of prediction ability, the composition of **H** was highly relevant. Across environments, models incorporating all available bands ($r = 0.54$) or only bands selected by Lasso ($r = 0.59$) were considerably more accurate than models based only on bands with heritabilities > 0.72 ($r = 0.48$). For scenario S1, HBLUP models based on $\mathbf{H_{all}}$ and $\mathbf{H_{h2}}$ were therefore discarded and hereafter HBLUP models are all based on $\mathbf{H_{vsel}}$.

For addressing the second factor (i.e., the TRN size), the performance of genotypes in scenarios S1 and S2 was predicted based on TRN of increased size. In S1, the prediction ability of proposed single-kernel, multi-kernel, and bivariate models (Table 1) was assessed with variable TRN sizes across environments (Table 2). The TRN sizes evaluated ranged from 55 (~ 20%) to 220 individuals (~ 80%). In general, the larger the TRN size, the higher the prediction ability of all models, and the lower their variance. The Bivariate_G + H model showed the highest prediction ability across TRN sizes, followed by the Bivariate_G model, the multi-kernel prediction model, the Bivariate_H model, and the single-kernel models (HBLUP and GBLUP). On the other hand, Bivariate_G + H was associated with the highest variability in reduced TRN sizes. This is in particular observable in Suppl. Fig. S6, where this model was compared with single-kernel and multi-kernel models. The model Bivariate_G + H estimates breeding values of genotypes based on PH, genotypic, and hyperspectral data, while the multi-kernel model does not include PH. For the smaller TRN size, the former showed a predictive ability of 0.56, while the latter yielded a predictive ability of 0.46. For the largest TRN size, their prediction ability was 0.75 and 0.71, respectively. Interestingly, across different selection intensities, the three bivariate models consistently selected the taller genotypes, which were not always associated with the highest DMY. In contrast, the multi-kernel model selected relatively shorter genotypes with an acceptable yield (Supp. Table S4). Both single-kernel models performed similarly with larger TRN sizes. For example, for TRN size of 80%, r was close to 0.60. On the other hand, HBLUP was more

Table 2 Mean prediction abilities and standard errors for dry matter yield of six models across different training set sizes for 274 winter rye hybrids assessed in eight environments across two flight dates

Model[a]	Training set size[b]			
	20 (%)	40 (%)	60 (%)	80 (%)
GBLUP	$0.32^{a} \pm 0.002$	$0.44^{a} \pm 0.002$	$0.54^{a} \pm 0.002$	$0.60^{a} \pm 0.003$
HBLUP	$0.42^{b} \pm 0.004$	$0.51^{b} \pm 0.002$	$0.56^{b} \pm 0.002$	$0.59^{a} \pm 0.003$
G+H	$0.46^{c} \pm 0.003$	$0.59^{d} \pm 0.002$	$0.66^{d} \pm 0.002$	$0.71^{d} \pm 0.003$
Bivariate_G	$0.54^{e} \pm 0.004$	$0.61^{e} \pm 0.002$	$0.66^{d} \pm 0.002$	$0.69^{c} \pm 0.003$
Bivariate_H	$0.50^{d} \pm 0.005$	$0.55^{c} \pm 0.004$	$0.60^{c} \pm 0.002$	$0.62^{b} \pm 0.003$
Bivariate_G+H	$0.56^{f} \pm 0.007$	$0.65^{f} \pm 0.004$	$0.71^{e} \pm 0.003$	$0.75^{e} \pm 0.002$

[a]See Table 1 for more information about the listed models

[b]Within a column, means with no letter in common are significantly different (Tukey's honestly significant difference test; $\alpha = 0.01\%$)

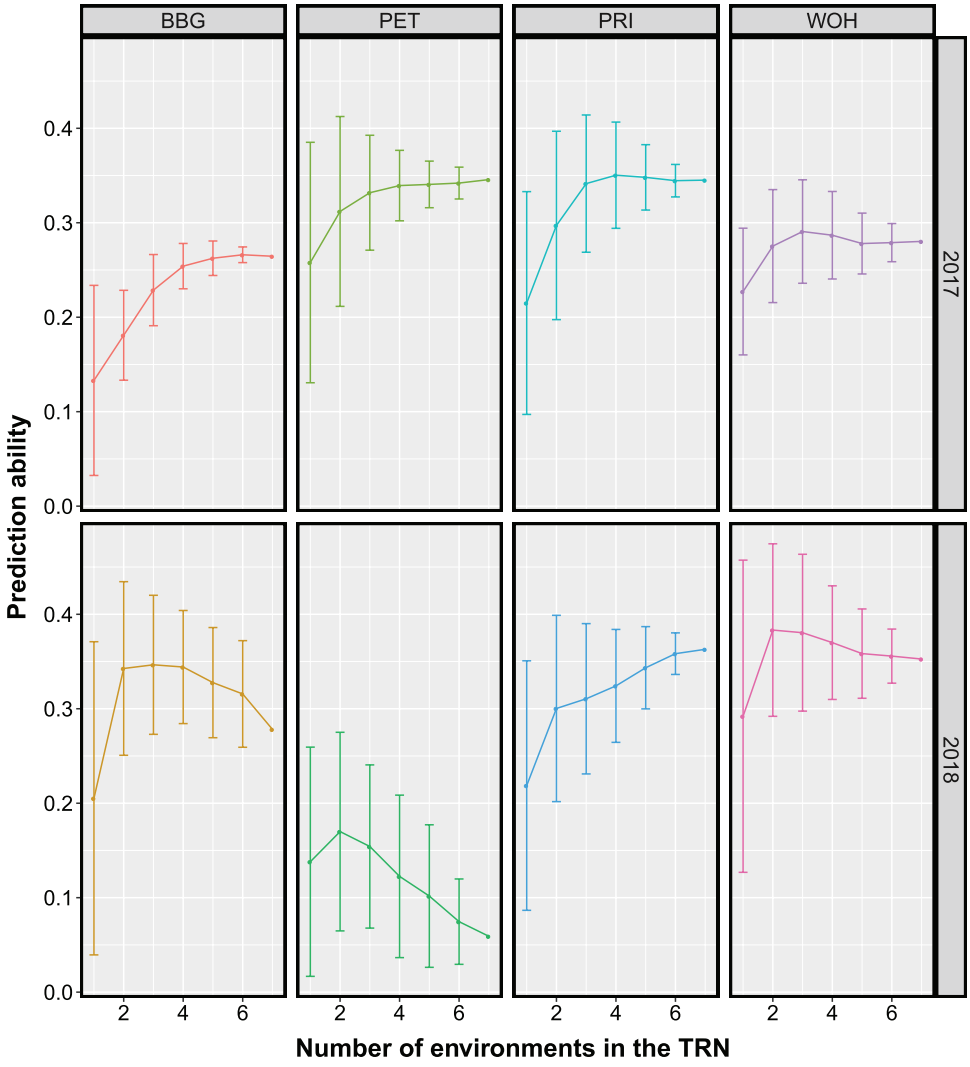


Fig. 4 Prediction ability for dry matter yield of the hyperspectral best linear unbiased predictor model (HBLUP) on each environment with increased number of environments included in the training set (TRN). Models were tested under validation scenario S2

predictively accurate than GBLUP based on smaller TRN sizes. For a TRN size of 55, HBLUP ($r = 0.42$) surpassed GBLUP ($r = 0.32$) by about 25%. A comparison among prediction models based on single flight data under validation scenario S1 is shown in Suppl. Table S3.

In S2, predictions were based on HBLUP models fitted with all bands ($\mathbf{H_{all}}$) collected in a variable number of environments. These environments were highly diverse according to a discriminant analysis (DAPC) based on reflectance data (Suppl. Fig. S7). Locations from the same year were mostly clustered together. The PET (2018) environment, on the other hand, was distinct to both clusters. Overall, the prediction of DMY in individual environments was improved by increased TRN size (Fig. 4). Thus, the higher the number of environments included in the TRN, the more accurate the prediction. However, the prediction ability was highly variable across environments. For TRN including only one environment, WOH (2018) showed the highest prediction ability ($r = 0.29$), while BBG (2017) showed the lowest ($r = 0.13$). When seven environments were considered as TRN, DMY had the highest prediction ability in PRI (2018) ($r = 0.36$), while it had the smallest in PET (2018) ($r = 0.06$).

Discussion

The high versatility of rye as a dual-use crop (Miedaner and Laidig 2019) contrasts with traditional breeding programs, which are mainly driven by GY (Geiger and Miedaner 2009). Thus, the improvement of DMY is often pushed into later selection stages. To overcome this situation, an effective indirect estimation of DMY based on data collected on GY plots would be needed. Thus, in the present study, single-kernel, multi-kernel, and bivariate models based on different information sources collected within the same breeding population were compared regarding their DMY prediction ability across different validation approaches.

Impact of heritability estimates of bands on HBLUP models

Across the spectrum, the magnitude and variability of h^2 estimates were higher than those of the correlation (r) between bands and DMY for combined (Fig. 1) and single flight dates (Suppl. Fig. S2). Highly variable h^2 for bands were also reported in wheat (Krause et al. 2019; Montesinos-López et al. 2017). We observed that h^2 and r showed the greatest variability and the lower values within the IR. HBLUP models exploiting all available bands were substantially more precise than those fitted only with highly heritable bands ($h^2 > 0.72$). This seems counterintuitive from a breeding perspective since, according to quantitative-genetic theory (Falconer and Mackay 1996), highly heritable secondary traits correlated with the feature of interest are preferred for indirect selection of the target trait. A possible explanation of the low performance observed by models based on bands with $h^2 > 0.72$ is that the proposed threshold excluded almost all the bands belonging to IR. Despite their relatively lower mean r with DMY, based on our results, this spectral region still captures information closely related to DMY, since bands around 750 nm had the highest correlation with DMY (Fig. 1). The magnitudes of these correlations were rather low but significant (< 0.29; $p < 0.05$) and are comparable to those stated for biomass in wheat by Hansen and Schjoerring (2003).Thus, the exclusion of bands from the IR because of their relatively lower h^2 deteriorated the predictive power of HBLUP models. This is in agreement with Montesinos-López et al. (2017), who found that GY prediction in wheat was not improved by removing bands with lower h^2.

Reduction in the dimensionality of hyperspectral data

High-throughput phenotyping is a promising tool for overcoming the phenotyping bottleneck in modern plant breeding (Araus and Cairns 2014). On the one hand, the use of hyperspectral sensors can substantially increase the amount of data available for dissecting the genetics behind the trait of interest. On the other hand, the application of this technology on multi-environmental trials is computationally and economically challenging.

The exploitation of a vast amount of hyperspectral data should be performed with caution, since the combination of a large number of predictors, each with small effects, can negatively influence the accuracy of regression models (Ogutu et al. 2012). The high multicollinearity found among contiguous bands (Suppl. Fig S1) suggests that performing variable selection could be beneficial. In this context, Lasso was a valuable tool for reducing the number of predictors incorporated into the HBLUP model. Also, with the constant development of high-resolution HTP sensors, the utility of feature selection procedures may be increased in proportion to the incorporation of broader spectral regions.

Informativeness of the VS and IR spectral regions

Use of HTP based on hyperspectral sensors can be time-consuming and resource-intensive although recent substantial improvements have occurred. Considerable overlaps were observed among specific bands highlighted by Lasso and EN in single and combined flight dates (Suppl. Fig. S4). Therefore, the reflectance data from these specific wavelengths may be of great interest to practical plant breeders. Redirecting computational costs toward these selected regions could reduce the efforts in data management. By

this, the superiority of hyperspectral sensors in terms of data collection and calibration compared to cheaper devices covering fewer reflectance regions (e.g., RGB cameras, Araus and Cairns 2014) may be fully exploited in a less resource-demanding manner.

In the present study, bands across the whole spectrum showed a significant correlation with DMY, with the IR displaying the highest correlation estimates (Fig. 1). Also, when the IR was excluded, the prediction ability of HBLUP substantially dropped as discussed above. The variable selection procedures applied have highlighted single bands located in the VS and the IR as highly informative for DMY prediction (Fig. 1, Suppl. Table S2, Suppl. Fig. S4). Nevertheless, the majority of the selected bands were located within the IR. These findings suggest that all spectral regions contain information potentially useful for DMY prediction; however, IR may be more informative than of VS.

These findings also indicate that a reduction in predictive power is expected if spectral fingerprints of genotypes are based on a reduced number of spectral regions. This is consistent with literature highlighting the importance of the VS and the IR in assessing essential plant parameters. The behavior of plants exposed to visible light has been widely investigated since a large proportion of this radiation is absorbed by the pigments present in green tissues (Lichtenthaler 1996). For instance, bands within the blue (450–520 nm) and green (520–600 nm) channels were found to be sensitive to aboveground biomass in wheat (Wang et al. 2017). In the transition from VS to IR, the so-called red edge, not only the highest correlation between bands and DMY was detected but also a relatively increased density of selected bands. The singularity of this region was also observed in the fact that it was correlated neither with VS nor with IR (Suppl. Fig. S1). Between 680–750 nm, the reflectivity of chlorophyll is sharply increased, a phenomenon that can be used to remotely assess plant health and growth (Seager et al. 2005) as well as chlorophyll concentrations (Filella and Penuelas 1994) and biomass at high canopy densities (Mutanga and Skidmore 2004). Similarly, the IR contains important information about physiological processes affecting biomass including chlorophylls and photosynthesis activity, as well as plant water status (Tucker 1979). The present work included IR data up to ~ 1000 nm, which have been revealed as highly relevant for DMY prediction. Considering that currently there are configurations that allow sensors to collect a broader IR spectrum, further research should focus on the benefits of deploying hyperspectral sensors capable of collecting additional reflectance data up to 2500 nm.

Improved prediction abilities by combining different sources of information

Under both validation procedures (S1 and S2), models were calibrated in a TRN of increased size. Overall, a positive correlation between the prediction abilities of models and TRN size was observed (Table 2, Fig. 4, Suppl. Fig. S6). The positive influence of TRN size in GS accuracy is well acknowledged in animal (VanRaden et al. 2009) and plant (Marulanda et al. 2015) breeding. Based on our results from the validation scenario S1, this trend also applies to HBLUP, multi-kernel, and bivariate models. Interestingly, the negative impact of reduced TRN was dissimilar across single-kernel models. While in larger TRN, GBLUP was more accurate than HBLUP, the opposite was observed in smaller TRN (Table 2, Suppl. Fig. S6). The reduction in the TRN size to a quarter (from 80 to 20%) represented a decay of about one-half and one-third in the prediction abilities of GBLUP and HBLUP, respectively. The predictive power of HBLUP was substantially higher than linear models fitted with VIs reported in a previous study (Galán et al. 2020). This is in complete agreement with Aguate et al. (2017) and Montesinos-López et al. (2017), who also found the superiority of models based on whole-spectrum data instead of on VIs.

In the validation scenario S2, prediction abilities were lower than in S1, indicating that predicting the yield of genotypes in a new environment is challenging (Fig. 4). In the DAPC (Suppl. Fig. S7), environments within the same year were grouped, reflecting the strong influence of the year effect, not only on agronomic traits (Galán et al. 2020) but also in the hyperspectral data collected at each site. The environmental conditions were very contrasting between 2017 and 2018. In Germany, 2018 was a very dry year, especially on the light sandy soils where rye is usually grown, and our experiments were conducted, e.g., Petkus, which has a very light soil (Suppl. Table S1). In this context, the inclusion of the maximum number of environments in the TRN, leaving only one as a validation environment was beneficial. Under CV accounting for environmental sampling, Utz et al. (2000) also observed that the proportion of the genotypic variance explained by models was enhanced by the inclusion of more environments in the TRN, especially for moderate inherited traits such as GY and GY components. In our study, HBLUP performance was even smaller if the VAL was composed of an environment poorly correlated with the sites within the TRN. Since models in S2 borrow information from closely related environments, prediction of these low correlated environments following this scheme is not recommended.

These findings suggest that the incorporation of hyperspectral data to enhance DMY prediction in rye could improve breeding efficiency. First, if due to budget constraints, a larger TRN size is not affordable, HBLUP could

be a valid strategy to precisely predict DMY. Second, the higher prediction ability of multi-kernel and bivariate models indicates that the incorporation of reflectance data and agronomic traits like PH into GS routines has a synergetic effect, when these data are correlated with the target trait. These findings are consistent with Krause et al. (2019), who found that for predicting GY in wheat, single-kernel models fitted with genomic- or hyperspectral-derived relationship matrices yielded similar results but multi-kernel models integrating both matrices surpassed both.

However, our results suggested that the use of bivariate models should be used with caution. On the one hand, they had the highest variability in small TRN. Under these circumstances, the sampling variability is substantially increased. Therefore, the advantage of bivariate over univariate models is reduced. Thus, multivariate regression analysis is not recommended for small sample sizes. On the other hand, the positive correlation between PH and DMY (Fig. 2) suggests that these prediction models should be used with care because taller genotypes would tend to be favored in the selection as indeed observed in Suppl. Table S4. So, breeding for increased lodging resistance would be highly advisable since even small differences in PH of the selected genotypes will multiply when subsequent breeding cycles are contemplated.

It has to be considered that we estimated our prediction abilities within one larger population by fivefold cross-validation. Validation scenario S1 was fitted with environmentally and genetically related data, representing a possible source of bias on the estimation of the predictive power of the models. Further research is needed to predict DMY across genetically different plant materials and/or different selection cycles, i.e., after recombination of selected entries. This would also include results from untested environments that might have a high impact on prediction ability, as shown in Fig. 4.

Conclusions

While the needs of sustainable renewable energy sources increase, the interest for high-yielding varieties to diversify maize-based cropping systems is boosted in proportion. To meet this demand, novel breeding strategies are needed to fully exploit the potential of rye as a biomass substrate. This study provided strong evidence that hyperspectral data can substantially improve the indirect selection of DMY within the same breeding population, thus enabling a cost-effective dual-purpose program using both DMY and GY as target traits. The reduction in data dimensionality could further enhance the prediction ability of models based on reflectance data. Relationship matrices derived from HTP data could be utilized as an alternative to GS when molecular data are not available, especially under reduced TRN sizes. Additionally, they are a suitable complementary source of information to leverage the accuracy of genomic tools. The superiority of the bivariate model over the multi-kernel model indicates that agronomic traits correlated with DMY can further enhance the efficiency of selection. Similar to the comparison of model performances across different TRN sizes, it would be relevant for practical breeding to investigate prediction ability across a varying degree of relatedness between the TRN and the VAL. Such analysis could assist breeders facing challenging prediction scenarios, including predicting new environments or novel lines that are unrelated to the training population.

Acknowledgements Open Access funding provided by Projekt DEAL. This study was funded by the German Federal Ministry of Food and Agriculture (BMEL) through the German Agency for Renewable Resources (FNR), grant number FKZ 22019716 to TM. We gratefully acknowledge the excellent support of the technical staff at each experimental station. We are particularly grateful to Hans-Otto Wegener, Jörn-Claus Gudehus, Karsten Sell, KWS LOCHOW GmbH, Bergen, Germany, for seed production and conducting field trials. We are also grateful for helpful comments by two anonymous referees.

Author's contributions statement RG analyzed the data and the wrote the manuscript. AMBV, HPP, and PT supported with statistical advice. CJ conducted hyperspectral phenotyping at all environments. PS supervised data collection at Wohlde, Prislich, and Bernburg and provided scientific advice. TM and AG designed the research project. TM further edited the manuscript. All the authors read and approved the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards The authors declare that the experiments comply with the current laws of Germany.

References

Adão T, Hruška J, Pádua L, Bessa J, Peres E, Morais R, Sousa J (2017) Hyperspectral imaging: a review on UAV-based sensors, data

processing and applications for agriculture and forestry. Remote Sens 9(11):1110
Aguate FM, Trachsel S, Pérez LG, Burgueño J, Crossa J, Balzarini M, Gouache D, Bogard M, Gdl C (2017) Use of hyperspectral image data outperforms vegetation indices in prediction of maize yield. Crop Sci 57(5):2517–2524
Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19(1):52–61. https://doi.org/10.1016/j.tplants.2013.09.008
Auinger H-J, Schönleben M, Lehermeier C, Schmidt M, Korzun V, Geiger HH, Piepho HP, Gordillo A, Wilde P, Bauer E (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). Theor Appl Genet 129(11):2043–2053
Bauer E, Schmutzer T, Barilar I, Mascher M, Gundlach H, Martis MM, Twardziok SO, Hackauf B, Gordillo A, Wilde P (2017) Towards a whole-genome sequence for rye (*Secale cereale* L.). Plant J 89(5):853–869
Bernal-Vasquez A-M, Utz H-F, Piepho HP (2016) Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. Theor Appl Genet 129(4):787–804. https://doi.org/10.1007/s00122-016-2666-6
Bernal-Vasquez A-M, Gordillo A, Schmidt M, Piepho HP (2017) Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC Genet 18(1):51. https://doi.org/10.1186/s12863-017-0512-8
European Commission (2018) EU agricultural outlook for markets and income, 2018–2030., European Commission, DG Agriculture and Rural. Brussels
R Core Team (2018) R. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria
Covarrubias-Pazaran G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. PLoS ONE 11(6):e0156744
Crain J, Mondal S, Rutkoski J, Singh RP, Poland J (2018) Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. Plant Genome. https://doi.org/10.3835/plantgenome2017.05.0043
Dunagan SC, Gilmore MS, Varekamp JC (2007) Effects of mercury on visible/near-infrared reflectance spectra of mustard spinach plants (*Brassica rapa* P). Environ Pollut 148(1):301–311
EEG (2012) Gesetz für den Ausbau erneuerbarer Energien (Erneuerbare-Energien-Gesetz - EEG). https://www.erneuerbare-energien.de/EE/Redaktion/DE/Gesetze-Verordnungen/eeg_2012_bf.html. Accessed 02 Nov 2019
EEG (2017) Gesetz für den Ausbau erneuerbarer Energien (Erneuerbare-Energien-Gesetz - EEG). https://www.gesetze-im-internet.de/eeg_2014/EEG_2017.pdf. Accessed 02 Nov 2019
Fahlgren N, Gehan MA, Baxter I (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Curr Opin Plant Biol 24:93–99
Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman, Essex
FAO (2019) FAOSTAT database. Food and Agriculture Organization of the United Nations. https://www.fao.org/faostat/en/#data/QC. Accessed 05 Nov 2019
Filella I, Penuelas J (1994) The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status. Int J Remote Sens 15(7):1459–1470
Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22
Furbank RT, Tester M (2011) Phenomics–technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16(12):635–644
Galán RJ, Bernal-Vasquez A-M, Jebsen C, Piepho H-P, Thorwarth P, Steffan P, Gordillo A, Miedaner T (2020) Hyperspectral Reflectance Data and Agronomic Traits Can Predict Biomass Yield in Winter Rye Hybrids. Bioenerg Res 13(1):168–182. https://doi.org/10.1007/s12155-019-10080-z
Geiger HH, Miedaner T (2009) Rye breeding. In: Carena MJ (ed) Cereals. Springer Science & Business Media, Berlin, pp 157–181
Gilmour AR, Gogel BJ, Cullis BR, Thompson R, Butler D (2009) ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK
Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded. A look into the black box of genomic prediction. Genetics 194(3):597–607
Haffke S, Kusterer B, Fromme FJ, Roux S, Hackauf B, Miedaner T (2014) Analysis of covariation of grain yield and dry matter yield for breeding dual use hybrid rye. Bioenerg Res 7(1):424–429. https://doi.org/10.1007/s12155-013-9383-7
Hansen PM, Schjoerring JK (2003) Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. Remote Sens Environ 86(4):542–553
Hoerl AE, Kennard RW (1970) Ridge regression. Biased Estim Nonorthogonal Problems Technometrics 12(1):55–67
Hothorn T, Bretz F, Westfall P, Heiberger RM, Schuetzenmeister A, Scheibe S, Hothorn MT (2016) Package 'multcomp'. Simultaneous inference in general parametric models. Project for statistical computing, Vienna, Austria
Jia Y, Jannink J-L (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192(4):1513–1522
Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24(11):1403–1405
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(1):94
Krause MR, González-Pérez L, Crossa J, Pérez-Rodríguez P, Montesinos-López O, Singh RP, Dreisigacker S, Poland J, Rutkoski J, Sorrells M, Gore MA, Mondal S (2019) Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat. G3 (Bethesda, Md.) 9(4):1231–1247. https://doi.org/10.1534/g3.118.200856
Lichtenthaler HK (1996) Vegetation stress: an introduction to the stress concept in plants. J Plant Physiol 148(1–2):4–14
Liu W, Li Q (2017) An efficient elastic net with regression coefficients method for variable selection of spectrum data. PLoS ONE 12(2):e0171122
Mahlein A-K, Oerke E-C, Steiner U, Dehne H-W (2012) Recent advances in sensing plant diseases for precision crop protection. Eur J Plant Pathol 133(1):197–209
Martis MM, Zhou R, Haseneyer G, Schmutzer T, Vrána J, Kubaláková M, König S, Kugler KG, Scholz U, Hackauf B (2013) Reticulate evolution of the rye genome. Plant Cell 25(10):3685–3698
Marulanda JJ, Melchinger AE, Würschum T (2015) Genomic selection in biparental populations: assessment of parameters for optimum estimation set design. Plant Breed 134(6):623–630
Meier U (1997) Growth stages of mono- and dicotyledonous plants. Blackwell Wissenschafts-Verlag, Berlin
Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829
Miedaner T, Laidig F (2019) Hybrid breeding in rye (*Secale cereale* L.). In: Al-Khayri JM et al (eds) Advances in plant breeding strategies: cereals, vol 5. Springer, Cham, Switzerland, pp 343–372
Miedaner T, Koch S, Seggl A, Schmiedchen B, Wilde P (2012) Quantitative genetic parameters for selection of biomass yield in hybrid rye. Plant Breeding 131(1):100–103

Miedaner T, Korzun V, Bauer E (2019) Genomics-based hybrid rye breeding. In: Miedaner T, Korzun V (eds) Applications of genetic and genomic research in cereals, Elsevier, Amsterdam, Netherlands, pp 329–348

Möhring J, Piepho HP (2009) Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci 49(6):1977–1988

Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S (2015) LinkImpute. Fast and accurate genotype imputation for nonmodel organisms. G3 Genes, Genomes, Genet 5(11):2383–2390

Montes JM, Melchinger AE, Reif JC (2007) Novel throughput phenotyping platforms in plant genetic studies. Trends Plant Sci 12(10):433–436. https://doi.org/10.1016/j.tplants.2007.08.006

Montesinos-López OA, Montesinos-López A, Crossa J, de Los Campos G, Alvarado G, Suchismita M, Rutkoski J, González-Pérez L, Burgueño J (2017) Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. Plant Methods 13(1):4. https://doi.org/10.1186/s13007-016-0154-2

Mutanga O, Skidmore AK (2004) Narrow band vegetation indices overcome the saturation problem in biomass estimation. Int J Remote Sens 25(19):3999–4014

Ogutu JO, Schulz-Streeck T, Piepho HP (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings 6(2):S10. https://doi.org/10.1186/1753-6561-6-S2-S10

Piepho HP, Möhring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177(3):1881–1888. https://doi.org/10.1534/genetics.107.074229

Piepho HP, Büchse A, Emrich K (2003) A hitchhiker's guide to mixed models for randomized experiments. J Agron Crop Sci 189(5):310–322

Piepho HP, Büchse A, Richter C (2004) A mixed modelling approach for randomized experiments with repeated measures. J Agron Crop Sci 190(4):230–247

Piepho HP, Moehring J, Schulz-Streeck T, Ogutu JO (2012) A stage-wise approach for the analysis of multi-environment trials. Biometrical Journal 54(6):844–860

Roux SR, Wortmann H, Schlathölter M (2010) Rye (*Secale cereale* L.) for biogas production-breeding capability. J Für Kulturpflanzen 62(5):173–182

Rutkoski J, Poland J, Mondal S, Autrique E, Pérez LG, Crossa J, Reynolds M, Singh R (2016) Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3: Genes, Genomes, Genet 6(9):2799–2808. https://doi.org/10.1534/g3.116.032888

Scarlat N, Dallemand J-F, Fahl F (2018) Biogas. Developments and perspectives in Europe. Renewable Energy 129:457–472. https://doi.org/10.1016/j.renene.2018.03.006

Seager S, Turner EL, Schafer J, Ford EB (2005) Vegetation's red edge. A possible spectroscopic biosignature of extraterrestrial plants. Astrobiology 5(3):372–390

Smith A, Cullis B, Gilmour A (2001) Applications: the analysis of crop variety evaluation data in Australia. Aust N Z J Stat 43(2):129–145

Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. Biometrics 50:1171–1177

Sun J, Poland JA, Mondal S, Crossa J, Juliana P, Singh RP, Rutkoski JE, Jannink J-L, Crespo-Herrera L, Velu G, Huerta-Espino J, Sorrells ME (2019) High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. TAG Theor Appl Genet 132(6):1705–1720. https://doi.org/10.1007/s00122-019-03309-0

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc: Ser B (Methodol) 58(1):267–288

Tucker CJ (1979) Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens Environ 8(2):127–150

European Union (2009) Directive 2009/28/EC of the European Parliament and of the Council of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC. Off J Euro Union 5:8–16

European Union (2010) Communication from the Commission on the practical implementation of the EU biofuels and bioliquids sustainability scheme and on counting rules for biofuels (2010/C 160/02). Off J Euro Union

Utz HF, Melchinger AE, Schön CC (2000) Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. Genetics 154(4):1839–1849

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423

VanRaden PM, van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92(1):16–24

Verhoeven KJF, Jannink JL, McIntyre LM (2006) Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity 96(2):139–149

Wang Y, Mette MF, Miedaner T, Gottwald M, Wilde P, Reif JC, Zhao Y (2014) The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. BMC genomics 15(1):556

Wang C, Feng M, Yang W, Ding G, Xiao L, Li G, Liu T (2017) Extraction of sensitive bands for monitoring the winter wheat (Triticum aestivum) growth status and yields based on the spectral reflectance. PLoS ONE 12(1):e0167679

White JW, Andrade-Sanchez P, Gore MA, Bronson KF, Coffelt TA, Conley MM, Feldmann KA, French AN, Heun JT, Hunsaker DJ (2012) Field-based phenomics for plant genetics research. Field Crops Res 133:101–112

Wimmer V, Albrecht T, Auinger H-J, Schön C-C (2012) synbreed. A framework for the analysis of genomic prediction data using R. Bioinformatics 28(15):2086–2087

Wimmer V, Lehermeier C, Albrecht T, Auinger H-J, Wang Y, Schön C-C (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195(2):573–587

Xue J, Su B (2017) Significant remote sensing vegetation indices: a review of developments and applications. J Sens 2017(1):17

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Stat Soc: Series B (Stat Methodol) 67(2):301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Journal:

Theoretical and Applied Genetics

Title:

Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye

Authors:

Rodrigo José Galán[1], Angela-Maria Bernal-Vasquez[2], Christian Jebsen[2], Hans-Peter Piepho[3], Patrick Thorwarth[1,2], Philipp Steffan[4], Andres Gordillo[4], Thomas Miedaner[1].

Affiliation:

[1] State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany.

[2] KWS SAAT SE, Grimsehlstraße 31, 37574 Einbeck, Germany.

[3] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, 70593 Stuttgart, Germany.

[4] KWS LOCHOW GMBH, Ferdinand-von-Lochow Straße 5, 29303 Bergen, Germany.

Corresponding author:

Thomas Miedaner
e-mail: thomas.miedaner@uni-hohenheim.de
Telephone No.: +49 711 459-22690

Available ORCID of the authors:

Angela-Maria Bernal-Vasquez: https://orcid.org/0000-0003-0415-831

Patrick Thorwarth: https://orcid.org/0000-0003-4456-2358

Thomas Miedaner: https://orcid.org/0000-0002-9541-3726

Online resource 1

Supplementary Table S1 Characterization of the four locations in Germany.

Location	Coordinates	Altitude (m a.s.l.)	Soil type	pH	P_2O_5 [mg/100g soil] [a]	K_2O [mg/100g soil] [a]	Mg [mg/100g soil] [a]	OM [%][a]	Nmin [kg/ha] [a]	Growing-Season Precipitation (mm) [b]	Growing-Season Average temperature (°C) [b]	Sowing dates	Harvest dates	Flight dates
				---------- 2017 ----------										
Bernburg	51°49'37.66N 11°43'37.15"E	85	Black soil	7.4	26	26	8	2	60	203	8.6	Sept. 30 2016	June 15 2017	May 23 2017
				---------- 2018 ----------										
				7.4	25	33	8	2	44	300	7.2	Sept. 29 2017	June 6 2018	May 28 & June 06 2018
				---------- 2017 ----------										
Petkus	51°59'2.61"N 13°21'29.38"E	137	Sandy loam	6.48	13.2	5.22	7.21	-	36	319	6.2	Oct. 10 2016	June 20 2017	June 08 & June 18 2017
				---------- 2018 ----------										
				6.48	1.2	5.71	4.43	1.51	33	258	8.2	Oct 17 2017	June 8 2018	May 31 & June 07 2018
				---------- 2017 ----------										
Wohlde	52°48'36.01"N 10°0'46.73"E	77	Loamy sand	5.7	7	10.8	8	-	45	440	7.1	Sept. 29 2016	June 21 2017	May 19 & June 19 2017
				---------- 2018 ----------										
				5.8	7	14.8	4.7	-	25	410	8.05	Oct. 17 2017	June 13 2018	May 25 and June 05 2018
				---------- 2017 ----------										
Prislich	53°15'58.73"N 11°39'36.11"E	38	sandy loam	5.8	4.2	16	6	-	32	180	13.8	Oct. 10 2016	June 21 2017	June 09 and June 20 2017
				---------- 2018 ----------										
				5.9	4.5	19	3	-	4	137	7.8	Sept. 20 2017	June 14 2018	May 29 & June 12 2018

[a] Soil condition: Phosphorus Pentoxide P_2O_5 [mg/100g soil]; Potassium oxide K_2O [mg/100g soil]; Magnesium Mg [mg/100g soil]; Organic matter OM [%]; Nitrogen concentration in the 0-60 cm soil layer Nmin [kg/ha].

[b] Weather data is incomplete due to technical problems in the weather stations in Bernburg (Oct.-Dec. 2016) and Prislich (Oct 2016-March 2017).

Supplementary Table S2 Hyperspectral bands (WL) selected (recovery rate > 40%) by the least absolute shrinkage and selection operator (Lasso) for 274 winter rye hybrids assessed in eight environments and two flight dates.

Visible Spectrum [a]	Infrared radiation [a]
WL410, WL414, WL565, WL588, WL673, WL680	WL702, WL721, WL723, WL733, WL739, WL742, WL758, WL759, WL761, WL764, WL825, WL854, WL908, WL917, WL932, WL935, WL939, WL942, WL945, WL959, WL968, WL970, WL978, WL980, WL987, WL993

[a] WLλ is the reflectance at a specific wavelength λ (±1 nm at the center of the band)

Supplementary Table S3 Mean prediction abilities and standard errors for dry matter yield of six models across different training set sizes determined for 274 winter rye hybrid assessed in eight environments. Hyperspectral data was collected on two flight dates ("first" and "second"), which were separately analyzed.

Model [a]	Flight date	Training set size [b]			
		20%	40%	60%	80%
HBLUP	First	$0.41^{g} \pm 0.004$	$0.49^{e} \pm 0.002$	$0.53^{f} \pm 0.002$	$0.54^{de} \pm 0.004$
	Second	$0.44^{ef} \pm 0.003$	$0.49^{e} \pm 0.002$	$0.51^{g} \pm 0.003$	$0.52^{ef} \pm 0.004$
G+H	First	$0.42^{fg} \pm 0.003$	$0.56^{d} \pm 0.002$	$0.64^{d} \pm 0.002$	$0.68^{c} \pm 0.003$
	Second	$0.48^{d} \pm 0.003$	$0.59^{c} \pm 0.002$	$0.65^{c} \pm 0.002$	$0.70^{b} \pm 0.003$
Bivariate_G	First	$0.53^{bc} \pm 0.004$	$0.61^{b} \pm 0.002$	$0.67^{bc} \pm 0.003$	$0.68^{c} \pm 0.003$
	Second	$0.54^{ab} \pm 0.003$	$0.65^{a} \pm 0.003$	$0.67^{bc} \pm 0.003$	$0.68^{c} \pm 0.003$
Bivariate_H	First	$0.45^{e} \pm 0.005$	$0.50^{e} \pm 0.004$	$0.55^{e} \pm 0.003$	$0.55^{d} \pm 0.004$
	Second	$0.44^{ef} \pm 0.006$	$0.48^{e} \pm 0.006$	$0.49^{g} \pm 0.005$	$0.50^{f} \pm 0.005$
Bivariate_G+H	First	$0.52^{c} \pm 0.007$	$0.61^{b} \pm 0.005$	$0.68^{b} \pm 0.003$	$0.72^{b} \pm 0.003$
	Second	$0.56^{a} \pm 0.007$	$0.65^{a} \pm 0.004$	$0.70^{a} \pm 0.004$	$0.75^{a} \pm 0.002$

[a] See Table 1 for more information about the listed models.

[b] Within a column, mean values followed by no letter in common are significantly different (Tukey's honestly significant difference test; $\alpha = 0.01\%$).

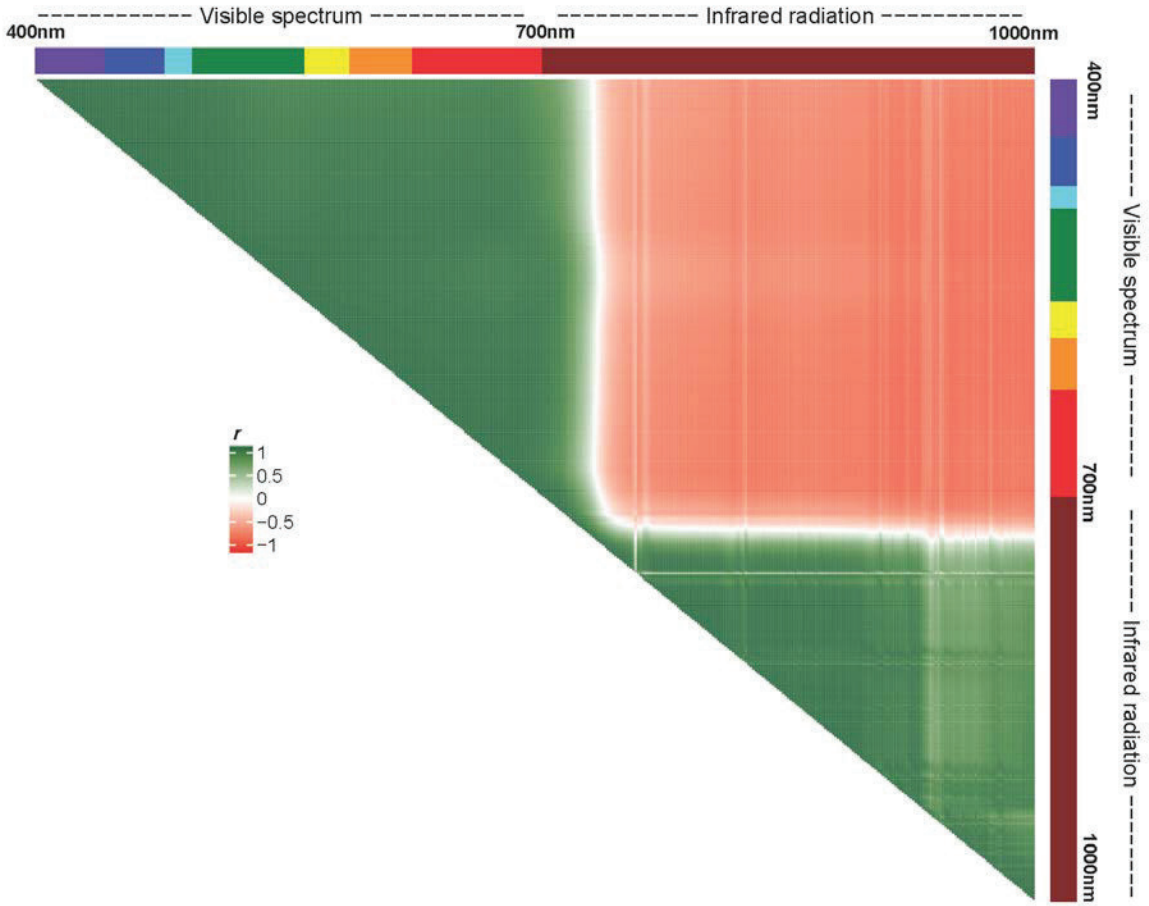
Supplementary Table S4 Prediction ability (r), dry matter yield (DMY, dt ha^{-1}), and plant height (PH, cm) obtained at different selected fractions by prediction models (TRN=80%) fitted with 274 winter rye hybrids assessed in eight environments and two flight dates.

Model [a]	Criterium	Selected fraction								$\bar{x}_{overall}$ [b]
		10%		20%		30%		40%		
		$\bar{x}$	SE	$\bar{x}$	SE	$\bar{x}$	SE	$\bar{x}$	SE	
GBLUP	r	0.33	0.02	0.35	0.01	0.43	0.01	0.46	0.01	0.39 b
HBLUP	r	0.25	0.02	0.31	0.01	0.38	0.01	0.42	0.01	0.34 c
G+H	r	0.30	0.02	0.39	0.01	0.43	0.01	0.47	0.01	0.40 b
Bivariate_G	r	0.42	0.01	0.44	0.01	0.49	0.01	0.51	0.01	0.47 a
Bivariate_H	r	0.25	0.02	0.36	0.01	0.41	0.01	0.40	0.01	0.36 c
Bivariate_G+H	r	0.33	0.02	0.45	0.01	0.48	0.01	0.50	0.01	0.44 a
GBLUP	DMY	122.30	0.04	121.76	0.03	121.21	0.02	120.88	0.02	121.54 (+1,7%) d
HBLUP	DMY	122.23	0.04	121.77	0.03	121.31	0.02	120.98	0.02	121.57 (+1,8%) d
G+H	DMY	122.64	0.04	122.04	0.03	121.59	0.02	121.26	0.02	121.89 (+2,0%) b
Bivariate_G	DMY	122.67	0.04	121.99	0.03	121.41	0.02	121.07	0.02	121.79 (+1,9%) c
Bivariate_H	DMY	122.31	0.04	121.73	0.03	121.30	0.02	121.05	0.02	121.60 (+1,8%) d
Bivariate_G+H	DMY	122.84	0.03	122.13	0.03	121.65	0.02	121.33	0.02	121.99 (+2,1%) a
GBLUP	PH	119.24	0.06	118.57	0.04	117.95	0.04	117.64	0.03	118.35 (+1,7%) e
HBLUP	PH	119.10	0.05	118.74	0.04	118.36	0.03	118.06	0.03	118.57 (+1,9%) d
G+H	PH	119.63	0.05	119.00	0.04	118.56	0.03	118.26	0.03	118.86 (+2,2%) c
Bivariate_G	PH	120.34	0.06	119.33	0.05	118.72	0.03	118.36	0.03	119.19 (+2,5%) b
Bivariate_H	PH	119.92	0.06	119.37	0.04	118.83	0.03	118.41	0.03	119.13 (+2,4%) b
Bivariate_G+H	PH	120.42	0.06	119.63	0.04	119.08	0.03	118.74	0.03	119.47 (+2,7%) a

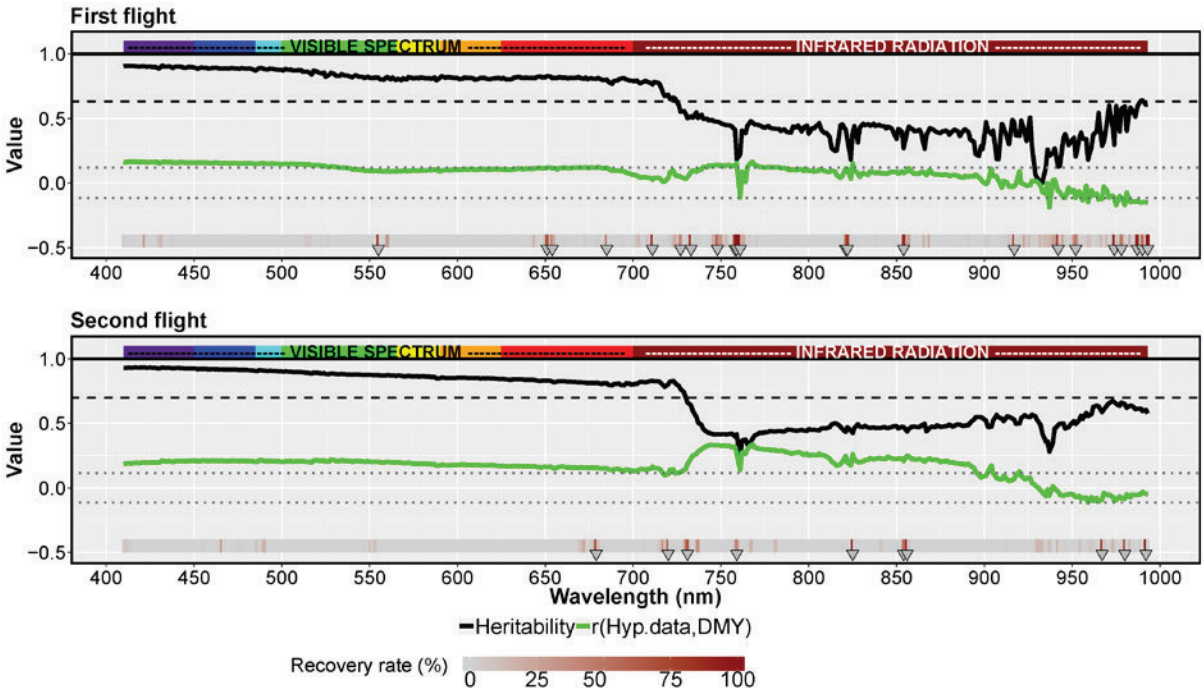
[a] See Table 1 for more information about the listed models.

$\bar{x}$, mean; SE, standard error of the mean; $\bar{x}_{overall}$, mean across all selected fractions.

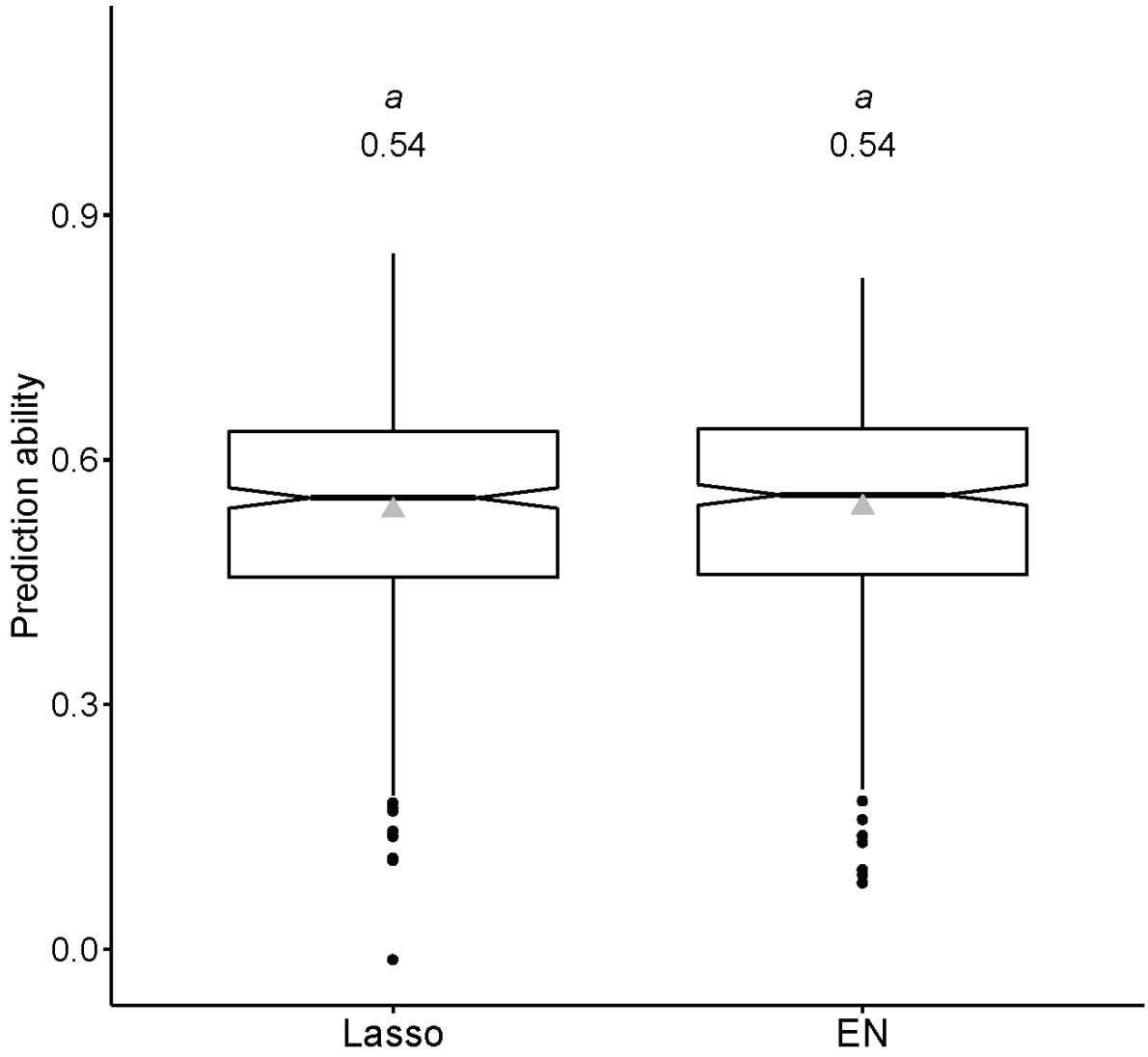
[b] The percent difference between the mean across all selected fractions and the population mean (DMY=119.48 dt ha^{-1}; PH=116.34 cm) is listed in brackets. Mean values followed by no letter in common are significantly different (Tukey's honestly significant difference test; α =0.01%).



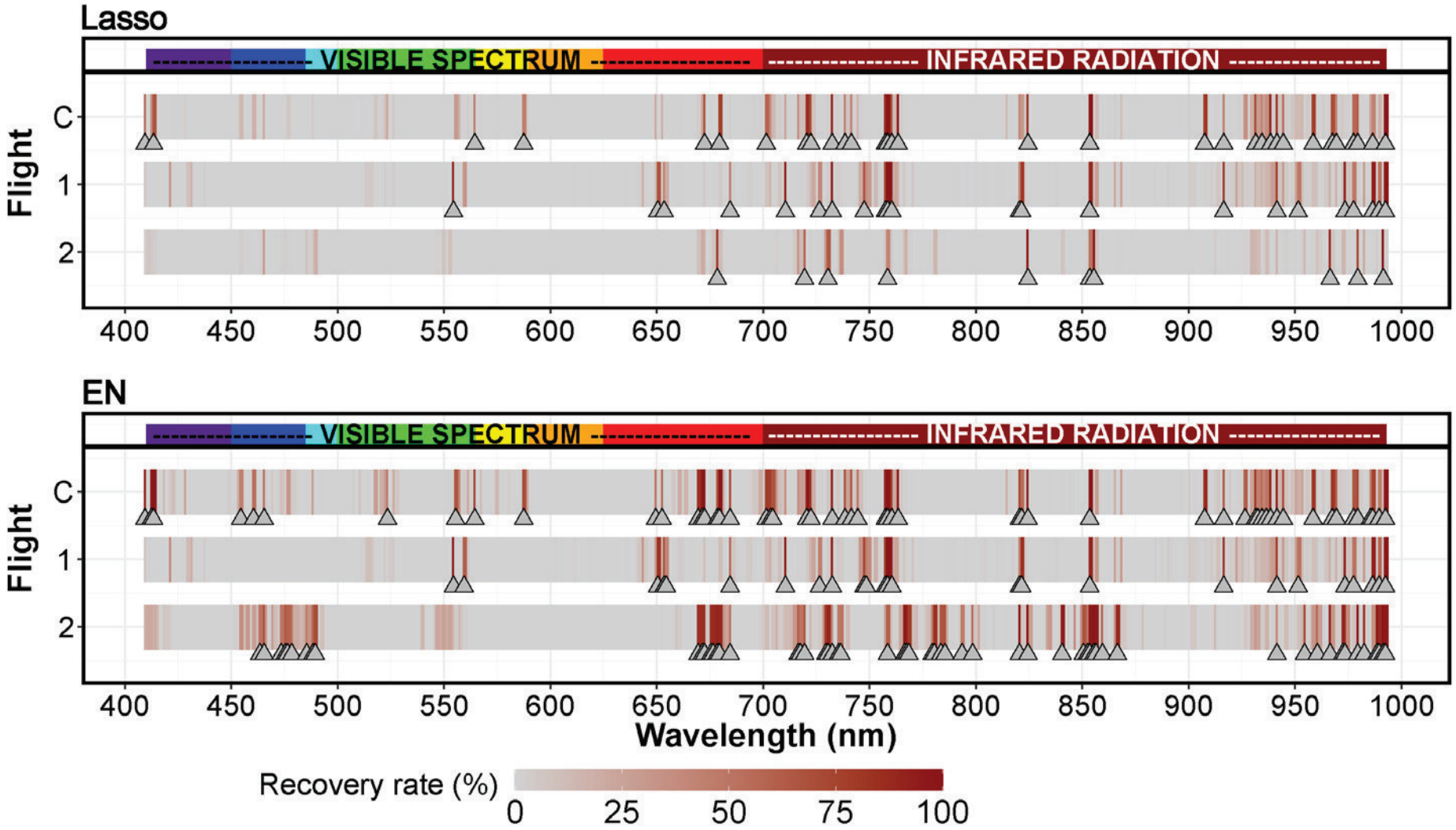
Supplementary Fig. S1 Pearson's coefficients of correlation (r) across hyperspectral bands based on 274 rye hybrids across eight environments.



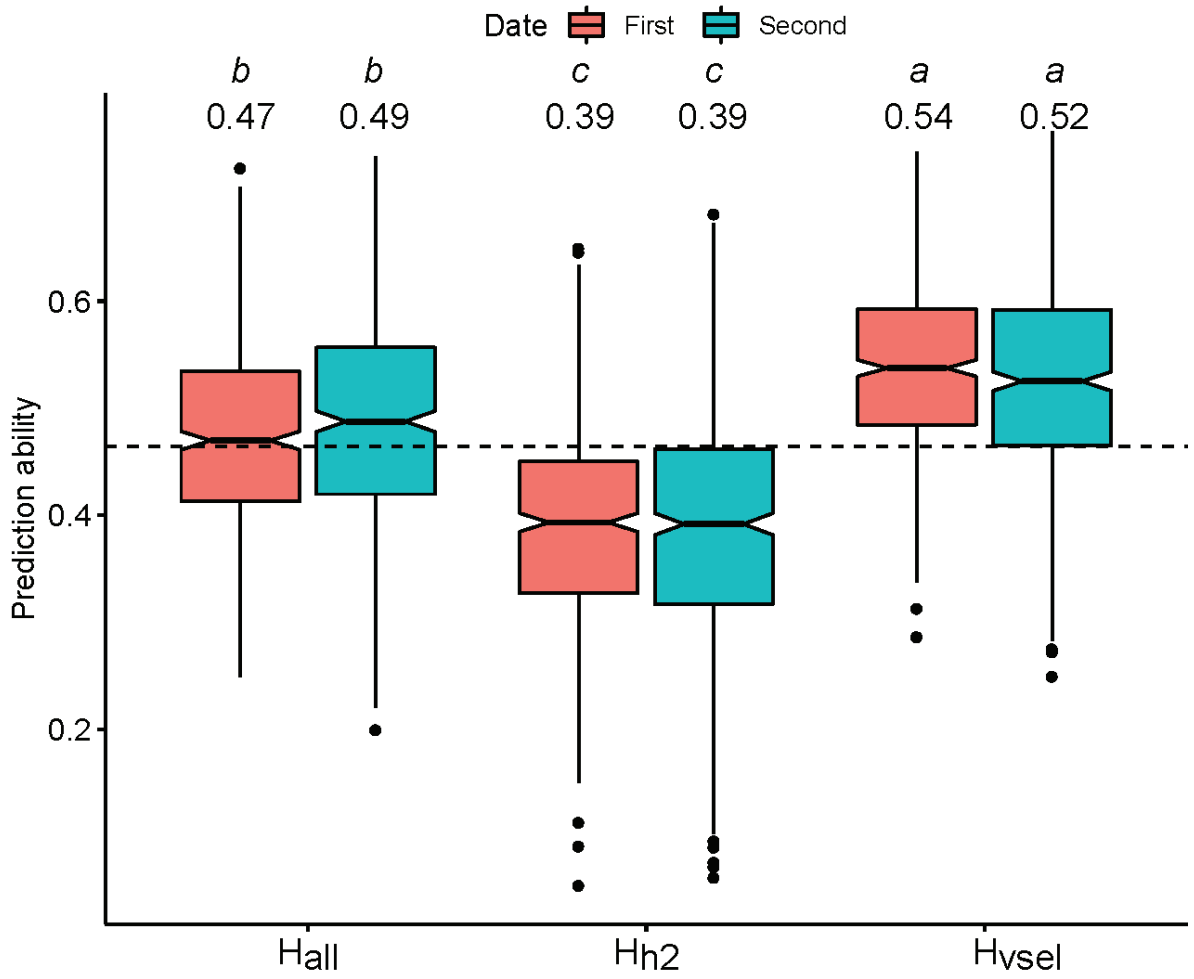
Supplementary Fig. S2 Heritability estimates (black line) for the hyperspectral bands, phenotypic correlations (r, green line) between hyperspectral bands and dry matter yield, and recovery rate (%) of hyperspectral bands after the least absolute shrinkage and selection operator (Lasso, gray-red heatmap) for 274 winter rye hybrids assessed in eight environments shown by flight date. The mean heritability across all wavelengths is denoted by the dashed black line. Correlation values $\geq |0.12|$ are significant ($p<0.05$) as showed by the gray dotted lines. Selected hyperspectral bands (recovery rate > 40%) are indicated by the gray triangles.



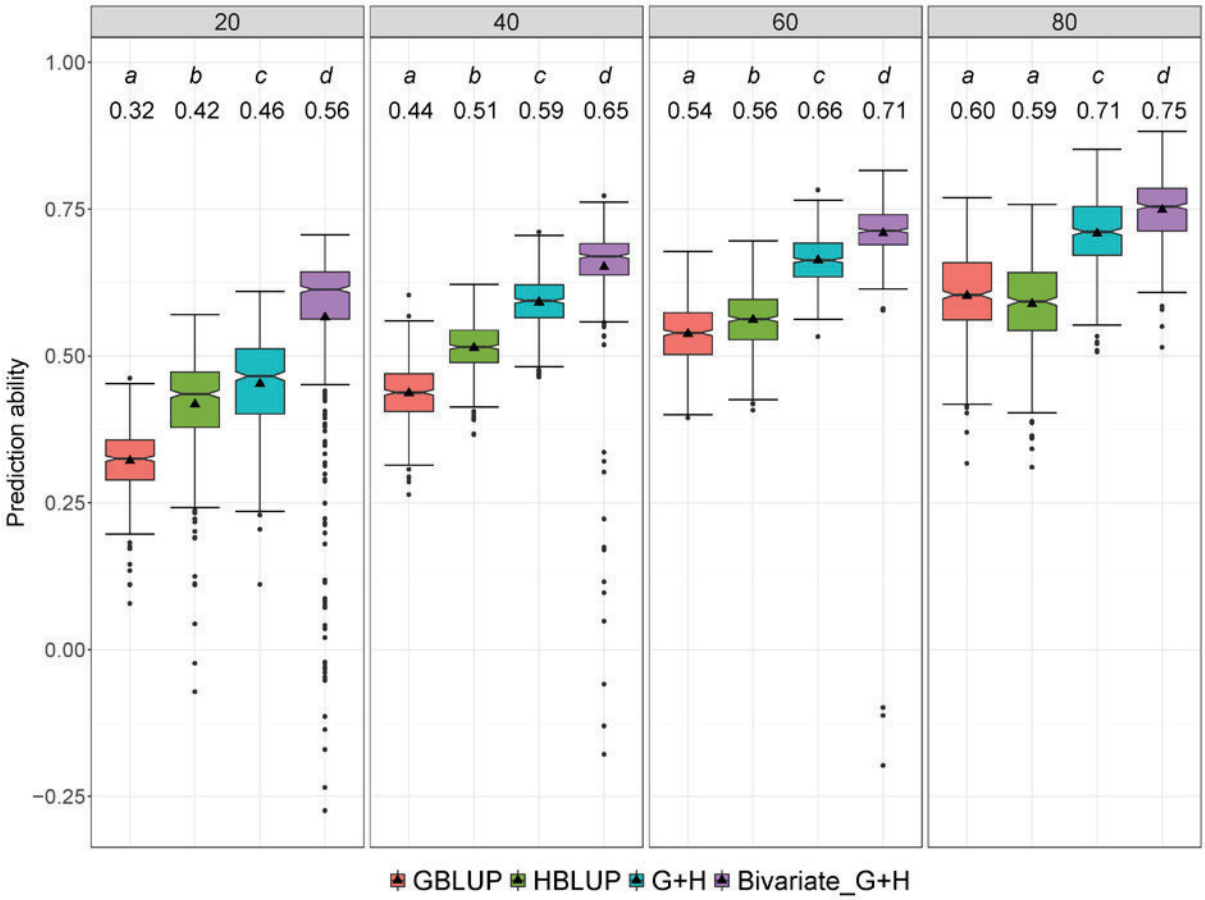
Supplementary Fig. S3 Prediction ability for dry matter yield based the least absolute shrinkage and selection operator (Lasso) and elastic net (EN) fitted with hyperspectral data collected in eight environments and two flight dates for 274 winter rye hybrids. Mean values are shown above each box plot and by gray triangles and are significantly different when headed by no letter in common (Tukey's honestly significant difference test; α =0.01%).



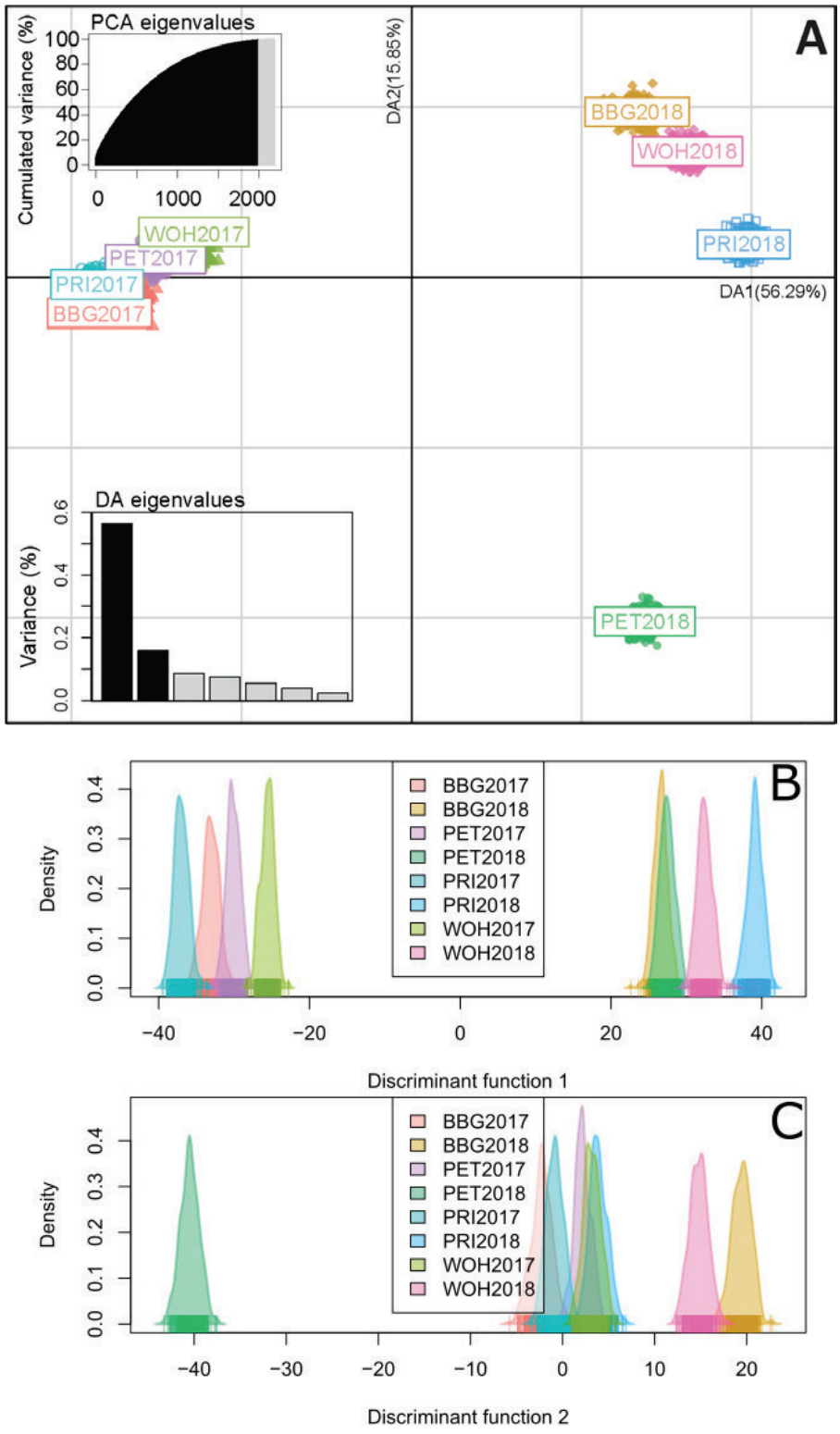
Supplementary Fig. S4 Hyperspectral bands selected (recovery rate > 40%) by the least absolute shrinkage and selection operator (Lasso) and elastic net (EN) fitted with hyperspectral data collected for 274 winter rye hybrids assessed in eight environments and two flight dates (1; First, 2; Second), which were individually and combined (C) analyzed.



Supplementary Fig. S5 Prediction ability for dry matter yield of hyperspectral best linear unbiased predictor model (HBLUP) collected on single flight dates based on different **H** relationship matrices, including all available 400 bands (**H_{all}**), bands with heritability > mean heritability across wavelengths **(H_{h2})**, and only selected bands (**H_{vsel}**) for 274 winter rye hybrids. Mean values are shown above each box plot and are significantly different when headed by no letter in common (Tukey's honestly significant difference test; α =0.01%). The mean prediction ability across all models and flight dates is denoted by the dashed black line.



Supplementary Fig. S6 Prediction ability for dry matter yield of single-kernel (Genomic best linear unbiased predictor, GBLUP and Hyperspectral best linear unbiased predictor, HBLUP), multi-kernel (G+H), and bivariate (Bivariate_G+H) models trained in four different training set (TRN) sizes for 274 winter rye hybrids. Models were tested under **validation scenario S1.** TRN sizes in percentage are shown at the top of each subplot. Mean values are shown above each box plot and by black triangles. Means headed within the same TRN size by no letter in common are significantly different (Tukey's honestly significant difference test; α =0.01%).



Supplementary Fig. S7 Discriminant analysis of principal components (DAPC) showing the clustering pattern of 274 winter rye hybrids across eight environments based on hyperspectral reflectance data. Scatter plot (A) for the first two discriminant functions (DA) including at the top left the PCA eigenvalues retained (in black) and at the bottom left the variation explained by each DA eigenvalues. Densities of individuals on the first (B) and the second (C) DA are also displayed.

5. Publication III: Early prediction of biomass in hybrid rye based on hyperspectral data surpasses genomic predictability in less-related breeding material

Rodrigo José Galán[1], Angela-Maria Bernal-Vasquez[2], Christian Jebsen[2], Hans-Peter Piepho[3], Patrick Thorwarth[1,2], Philipp Steffan[4], Andres Gordillo[4], Thomas Miedaner[1].

[1] State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany.

[2] KWS SAAT SE, Grimsehlstraße 31, 37574 Einbeck, Germany.

[3] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, 70593 Stuttgart, Germany.

[4] KWS LOCHOW GMBH, Ferdinand-von-Lochow Straße 5, 29303 Bergen, Germany.

Theoretical and Applied Genetics (2021) 134: 1409–1422

The original publication is available at
https://doi.org/10.1007/s00122-021-03779-1

Theoretical and Applied Genetics (2021) 134:1409–1422
https://doi.org/10.1007/s00122-021-03779-1

ORIGINAL ARTICLE

Early prediction of biomass in hybrid rye based on hyperspectral data surpasses genomic predictability in less-related breeding material

Rodrigo José Galán[1] · Angela-Maria Bernal-Vasquez[2] · Christian Jebsen[2] · Hans-Peter Piepho[3] · Patrick Thorwarth[1,2] · Philipp Steffan[4] · Andres Gordillo[4] · Thomas Miedaner[1]

Received: 21 August 2020 / Accepted: 19 January 2021 / Published online: 17 February 2021

Abstract

Key message **Hyperspectral data is a promising complement to genomic data to predict biomass under scenarios of low genetic relatedness. Sufficient environmental connectivity between data used for model training and validation is required.**

Abstract The demand for sustainable sources of biomass is increasing worldwide. The early prediction of biomass via indirect selection of dry matter yield (DMY) based on hyperspectral and/or genomic prediction is crucial to affordably untap the potential of winter rye (*Secale cereale* L.) as a dual-purpose crop. However, this estimation involves multiple genetic backgrounds and genetic relatedness is a crucial factor in genomic selection (GS). To assess the prospect of prediction using reflectance data as a suitable complement to GS for biomass breeding, the influence of trait heritability (H^2) and genetic relatedness were compared. Models were based on genomic (GBLUP) and hyperspectral reflectance-derived (HBLUP) relationship matrices to predict DMY and other biomass-related traits such as dry matter content (DMC) and fresh matter yield (FMY). For this, 270 elite rye lines from nine interconnected bi-parental families were genotyped using a 10 k-SNP array and phenotyped as testcrosses at four locations in two years (eight environments). From 400 discrete narrow bands (410 nm–993 nm) collected by an uncrewed aerial vehicle (UAV) on two dates in each environment, 32 hyperspectral bands previously selected by Lasso were incorporated into a prediction model. HBLUP showed higher prediction abilities (0.41 – 0.61) than GBLUP (0.14 – 0.28) under a decreased genetic relationship, especially for mid-heritable traits (FMY and DMY), suggesting that HBLUP is much less affected by relatedness and H^2. However, the predictive power of both models was largely affected by environmental variances. Prediction abilities for DMY were further enhanced (up to 20%) by integrating both matrices and plant height into a bivariate model. Thus, data derived from high-throughput phenotyping emerges as a suitable strategy to efficiently leverage selection gains in biomass rye breeding; however, sufficient environmental connectivity is needed.

Keywords Biomass · Genetic relatedness · High-throughput phenotyping · Genomic prediction · Prediction ability · Rye

Introduction

Worldwide, the consumption of energy obtained from renewable origins, especially bio-based sources, is rising (World Bioenergy Association 2019). In the European Union (EU), for instance, the share of renewable energy is expected to be between 55 and 75% of the total energy consumption in 2050, increasing in proportion the needs for biomass (European Commission 2011). New policy directives have established sustainability guidelines for bioenergy production (European Union 2010). For example, in Germany, the principal European biogas producer, the permitted share of maize (*Zea mays* L.) as the most

Communicated by Thomas Lubberstedt.

✉ Thomas Miedaner
thomas.miedaner@uni-hohenheim.de

[1] State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany

[2] KWS SAAT SE, Grimsehlstraße 31, 37574 Einbeck, Germany

[3] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, 70593 Stuttgart, Germany

[4] KWS LOCHOW GMBH, Ferdinand-von-Lochow Straße 5, 29303 Bergen, Germany

common fermentation substrate has been limited to 44% by 2021 (Renewable Energy Sources Act "EEG", EEG 2017). Thus, suitable alternatives are welcome to diversify maize-based biomass production.

Among the small-grain cereals, winter rye (*Secale cereale* L.) stands out for its vigorous growth and enhanced tolerance to abiotic and biotic stress factors. Europe is the largest rye grower worldwide covering about 81% of the global area with Russia, Poland, and Germany being the main producers (FAO 2019). In a previous study, rye demonstrated its high dry matter yield (DMY) potential even on sandy soils and under drought stress (Galán et al. 2020a). Under these conditions, rye yielded 8.4 t dry matter ha^{-1}, and under better environmental conditions, yields were up to 14.7 t dry matter ha^{-1}. Rye can, therefore, represent a suitable alternative for biomass production in a variety of agroecological conditions, including areas where the cultivation of other cereal crops would not be competitive (Geiger and Miedaner 2009). Considering that three quarters of the rye harvest is used for non-food purposes, rye appears as a sustainable alternative source of biomass (Geiger and Miedaner 2009; Miedaner et al. 2012).

In Germany, only 4 rye varieties are currently registered for whole plant silage (Bundessortenamt 2019). Rye is, however, mainly bred for grain yield (GY; Haffke et al. 2014) which is, in the breeding scheme proposed here, already assessed in the first year of general combining ability testing (GCA-1), generally sharing only less-related genotypes over the years (Suppl. Fig. 1). Then, within each selection cycle, a selected fraction of GCA-1 is re-evaluated for GY and additionally for DMY by destructive methods in duplicated GCA-2 experiments the following year, mainly due to the high costs of assessing DMY in a large GCA-1 population. At these first selection stages, the enhancement of DMY is, therefore, heavily dependent on the adequate exploitation of indirect selection (Falconer and Mackay 1996).

Higher selection gains have been reported when plant height (PH) was used as a secondary trait instead of GY to indirectly estimate DMY in hybrid rye (Haffke et al. 2014; Galán et al. 2020a). Recently, multi-kernel models jointly using reflectance and genomic data as alternative sources of information and bivariate models including also the routinely assessed PH were suggested as superior strategies to leverage rye as a dual-purpose crop in an affordable manner for the breeder (Galán et al. 2020b). By this, the available genetic variation present in the GCA-1 population may be better exploited without the need to duplicate these large-scale trials and, therefore, the selection gain for DMY could be further enhanced. In consequence, fewer and superior DMY-genotypes being tested in GCA-1 trials could be identified, reducing the amount of capital, time, and labor needed to conduct the destructive sampling of DMY in GCA-2 trials. In this context, the non-destructive assessment of DMY at earlier stages arises as a crucial prerequisite.

Imaging-based phenotyping quantitatively measures the interaction (e.g., absorbance, reflectance, or transmittance of photons) between the incident light and plant tissues, which at specific regions of the electromagnetic spectrum is linked to a wide range of morphological and physio-chemical canopy properties (Li et al. 2014). As observed by Rincent et al. (2018), this interaction is mainly mediated by the chemical composition of the tissues, which is itself determined by endophenotypes, intermediate molecular phenotypes associated with a quantitative trait (Mackay et al. 2009), and genetics. Thus, based on reflectance data, high-throughput phenotyping (HTP) can acquire a considerable amount of detailed phenotypic information of key traits from a large number of genotypes, emerging as a valuable breeding tool (Montes et al. 2007; Cabrera-Bosquet et al. 2012; Würschum 2019).

Examples of the application of HTP in plant breeding are among others, the estimation of above-ground biomass (Babar et al. 2006; Montes et al. 2011; Busemeyer et al. 2013; Fu et al. 2014; Barmeier and Schmidhalter 2017; Yue et al. 2017, 2018) as well as GY, plant responses to biotic and abiotic stress, nitrogen use efficiency, nutrient status, early plant vigor, seeds quality traits, leaf physiology and biochesmistry, vegetation cover fraction, and leaf area index (reviewed by Fahlgren et al. 2015; Yang et al. 2017; Würschum 2019). Therefore, it has been proposed to remotely phenotype large breeding populations in a reliable and cost-effective manner (Furbank and Tester 2011; White et al. 2012). HTP platforms, including uncrewed aerial vehicles (UAVs) such as drones mounted with hyperspectral cameras, can simultaneously collect hundreds of high-resolution images, screening the electromagnetic spectrum (from 400 up to 2500 nm) in a continuous mode (Araus and Cairns 2014). Consequently, this noninvasive technology represents a valuable tool for the improvement of complex traits (Finkel 2009; Fiorani and Schurr 2013).

Genome-wide molecular markers integrated into genomic selection (Meuwissen et al. 2001) have been successfully applied in several study cases in hybrid rye breeding for relevant traits, e.g., GY and GY components (Auinger et al. 2016; Bernal-Vasquez et al. 2017; Miedaner et al. 2019). Moreover, in previous studies, reflectance fingerprints recorded by HTP platforms represented a valuable tool to improve the prediction ability of DMY in hybrid rye of models based on agronomic (Galán et al. 2020a) and genomic information (Galán et al. 2020b). These studies have shown the benefits of integrating hyperspectral and molecular information for predicting DMY of unphenotyped candidates within single or closely related populations. The proposed models were cross-validated, where rye lines derived from the same cross were randomly allocated to the training (TRN) or validation (VAL) sets. Considering

the breeding scheme at hand, where DMY is tested at later stages, predictions of candidates of subsequent selection cycles, where TRN and VAL correspond to different, largely independent genetic backgrounds, would be of utmost interest. This "across-cycles" prediction would allow, for instance, estimating the DMY performance of GCA-1 candidates (being tested only for GY at this stage) by training the model with GCA-2 phenotypic data from one or several previous selection cycles (Suppl. Fig. 1). It is under these scenarios where the largest contribution of predictive breeding towards an affordable dual-purpose rye breeding program is expected. If the data available consist of multiple connected cycles, breeders could consider to combine them to improve the predictive power of models (Auinger et al. 2016).

However, the predictive power of GS critically depends on a close relationship between TRN and VAL (Habier et al. 2007; Miedaner et al. 2019). Reduced or even negative prediction accuracies were reported for GS among less related bi- and multiparental families in several crops, including wheat (Herter et al. 2019), maize (Riedelsheimer et al. 2013; Lehermeier et al. 2014), sugar beet (Würschum et al. 2013), and barley (Thorwarth et al. 2017). Similarly, genomic prediction models showed modest prediction ability for complex traits in rye (e.g., GY) when applied between bi-parental families even though they were connected by a common parental line (Wang et al. 2014). Here, the question of whether alternative or complementary approaches to GS for leveraging prediction accuracies across less connected datasets emerges as highly relevant for biomass breeding in rye.

The aim of our study was, therefore, to answer this question by evaluating and comparing genomic- and hyperspectral-enabled predictions for three biomass-related traits (DMY, FMY, and DMC) in rye under a varying degree of relatedness between TRN and VAL. Additionally, the advantages of combining different sources of information in multi-kernel and bivariate models to leverage the prediction of DMY were evaluated. We employed 270 winter rye lines from nine interconnected bi-parental families, including their parental components tested as testcrosses in 8 environments (= location-year combination). While keeping the TRN size constant, our specific objectives were to perform (1) prediction of progenies from half-sib and unrelated parents, (2) prediction using only progenies from unrelated parents, and (3) prediction of new progenies in a new environment.

Materials and methods

Plant materials, field experiments, hyperspectral and molecular data

The plant materials, field experiments, molecular and hyperspectral data analyzed in the present study have been described before in detail by Galán et al. (2020b). In short, ten diverse parental lines of the Petkus (seed parent) gene pool were crossed following a single-round robin design (Verhoeven et al. 2006). F1 progenies were derived from each of the chain crosses, i.e., line 1 × line 2, line 2 × line 3, …, line 10 × line 1. After self-fertilization of single F1 plants for four consecutive years (S_4 generation), 264 recombinant inbred lines (RILs) were obtained. The ten bi-parental families ranged from 4 up to 32 RILs (Supp. Fig. S2) and were clearly distinct in a principal component analysis (PCA) based on molecular data with little overlap between unrelated crosses in the first two dimensions (Supp. Fig. S3). A total of 274 three-way hybrids [(A • B) × C] were produced from the cross of these 264 RILs and their ten parental components with a single-cross tester from the opposite (pollinator) gene pool. They were evaluated in two adjacent trials laid out as a resolvable incomplete block design (α-lattice design) with two replicates in 2017 and 2018 at each of four ecologically different locations (Bernburg, Petkus, Wohlde and Prislich) in Northern Germany (i.e., eight location-year combinations hereafter referred as "environments"). All 274 testcrosses were used for estimating means, variance components, and heritabilities (Table 2), whereas 4 genotypes were not considered for prediction modeling as described in later sections. Plots were harvested at the late milk stage (Meier 1997) to get the respective fresh biomass yield (FMY, dt ha^{-1}) per plot. During harvest, representative samples of about 1000 g were weighed from each plot and oven-dried to a constant weight at 110 °C. Dry matter content (DMC, %) was determined by weight differences. Then, DMY (dt ha^{-1}) per plot was estimated as DMY = FMY × DMC/100. Also, PH (cm) was recorded at each plot.

During the grain-filling stage, an UAV (Camflight FX8HL, Sandnes, Norway) fitted with a hyperspectral camera (HySpex Mjolnir V-1240, Skedsmokorset, Norway) collected reflectance fingerprints consisting of 400 bands (410 nm – 993 nm) for all genotypes in all environments. The UAV flew at about 25 m above plots, around solar noontime two times per environment (except in Bernburg 2017 where only one flight took place). Then each plot was identified on the obtained images by a polygon. Raw data were radiometrically calibrated (HySpex PostProcessor Version 1.2) and normalized based on the incident sunlight as well as orthorectified and georeferenced via the PARGE Software (ReSe Applications LLC, Wil, Switzerland). Lastly, all data points within each wavelength and polygon were averaged, resulting in one spectrum per plot. Then, these data were transferred to a tabular data frame, including the computed reflectance values of all bands for all genotypes for further analysis.

The 264 RILs and their ten parental components were also genotyped with an Illumina INFINIUM chip with 9,963 single nucleotide polymorphisms (SNPs) assays (KWS SAAT

SE & Co. KG, Einbeck, Germany). Data quality analysis consisted of the exclusion of SNPs showing more than 10% missing values or a minor allele frequency (MAF) < 0.05. Missing values in the remaining data were then imputed by the software Linkimpute (Money et al. 2015). Then, data were again screened for MAF < 0.05. After this procedure, 6,420 markers remained for further analyses.

Phenotypic data analysis

The analyses were based on adjusted entry means (best linear unbiased estimators, BLUEs) for all agronomic traits estimated within and across environments for subsequent incorporation into prediction models. The combined analysis across environments as well as the data adjustment within single environments were conducted following model (1) and model (2) from Galán et al. (2020b), respectively. The full model can also be found in the Supplementary File 1. For the analysis across environments within the same year, the year main effect and corresponding interactions with genotypes were dropped from the mixed model. Phenotypic data were filtered for outliers at the trial level using the Bonferroni-Holm test (Bernal-Vasquez et al. 2016). Bands were deleted from plots identified as an outlier for DMY.

Stage-wise procedure for biomass traits prediction

The incorporation of genomic and hyperspectral data for predicting DMY, FMY, and DMC was conducted by a three-stage procedure (Piepho et al. 2012). This analysis, together with the corresponding linear mixed and prediction models employed at each stage, was previously described in detail in Galán et al. (2020b). All statistical analyses were performed within the R-environment v. 3.4.4 (R Core Team 2018).

In the first stage of this analysis, bands were adjusted across flight dates per environment. Then, the obtained adjusted entry means (BLUEs) were used in the second stage for the estimation of BLUEs per genotypes across environments. At this second stage, heritability (H^2) was estimated for all agronomic traits and each band across environments as

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\bar{v}}{2}}$$

where $\bar{v}$ is the mean variance of a difference of two adjusted genotype means (BLUEs) estimated for phenotypic and hyperspectral data (Piepho and Möhring 2007). BLUEs of genotypes were calculated with the software package *ASReml-R* v. 3.0 (Gilmour et al. 2009).

In the third stage, the phenotypic and hyperspectral BLUEs were used for fitting prediction models to estimate best linear unbiased predictions (BLUP) of genotypic effects for each agronomic trait based on genetic and hyperspectral data. Two single-kernel prediction models were fitted with genetic (genomic BLUP, GBLUP) or hyperspectral (hyperspectral BLUP, HBLUP) data with $n = 270$ individuals, based on the $m = 6{,}420$ conserved SNP markers or $b = 32$ bands, respectively.

For GBLUP, the random genetic values (effects) were estimated based on genetic data incorporated into **G**, a genomic additive relationship matrix (Habier et al. 2013). **G** was calculated with the *synbreed* package (Wimmer et al. 2012) in R according to the "method I" of VanRaden (VanRaden 2008) as $G = \frac{ZZ'}{2\sum p_i(1-p_i)}$, where Z = M - P, M is the $n \times m$ marker matrix reflecting the SNP genotype of nth individual at the mth SNP position the of alleles coded as 0, 1, and 2 for A_1A_1, A_1A_2, and A_2A_2, respectively, P contains a $n \times m$ matrix of allele frequencies multiplied by 2, p_i is the allele frequency of the ith allele. For the prediction scenario S2 (described below), the GBLUP model was adapted from the model (7) in Bernal-Vasquez et al. (2017) as

$$y = Xb + Z_g u_g + Z_{ge} u_{ge} + e \quad (1)$$

where $\boldsymbol{y}$ is the vector of BLUEs of genotype trait values obtained from within-environments, $\boldsymbol{X}$ is the design matrix of the environments,$\boldsymbol{\beta}$ is the vector of environments effects, $\boldsymbol{Z_g}$ is the marker matrix for genotypes, and $\boldsymbol{u_g}$ the vector of marker effects. The genotype-by-environment effects is modelled by $\boldsymbol{w} = \boldsymbol{Z_{ge}}\boldsymbol{u_{ge}}$, with $\boldsymbol{Z_{ge}}$ standing for the marker matrix for genotypes-by-environment effects and $\boldsymbol{u_{ge}}$ the vector of marker-by-environment effects with variance $var(\boldsymbol{u_{ge}}) = \mathbf{I}\sigma_{ge}^2$, thus $var(\boldsymbol{w}) = \boldsymbol{Z_{ge}}\boldsymbol{Z_{ge}^T}\sigma_{ge}^2$. $\boldsymbol{Z_{ge}}$ is a block-diagonal matrix with blocks given by the marker coefficient matrices of genotypes in a given environment (Z_{ge_r}) and for the eight environments considered in the present study, it can be defined as $\begin{pmatrix} Z_{ge_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_{ge_8} \end{pmatrix}$. The variance of $\boldsymbol{w}$ stands for the linear structure of the genotype-by-environment variance–covariance matrix with the covariance of two genotypes within the same environment depending on the similarity in their marker profiles (Piepho 2009). Since the covariance among different environments is zero, any covariance between environments is captured by $\boldsymbol{Z_g}$.

As a measure of the genetic similarity among all n candidates, the Pearson's coefficients of correlation among rows of M were calculated. Based on their SNP alleles, this genomic correlation (r_{GC}) reflects the correlation pattern among individuals (Riedelsheimer et al. 2013). In contrast, for HBLUP the estimation of the random genetic values was based on reflectance data integrated into the hyperspectral reflectance-based relationship matrix **H** defined as $\boldsymbol{H = DD'}$, where $\boldsymbol{D}$ is a $n \times b$ hyperspectral matrix of the standardized

BLUEs of the bands, with $b = 32$. These 32 bands belong to the visible spectrum (VS) and the infrared radiation (IR), and they were selected in a previous study (Galán et al. 2020b) using the least absolute shrinkage and selection operator (Lasso; Tibshirani 1996) for reducing the multicollinearity observed among continuos bands and increasing, therefore, the predictive power of reflectance-based models. Following the same procedure as described before for r_{GC}, a second correlation (r_{HC}) among tested genotypes was developed based on hyperspectral data incorporated into **H´,** which was derived from $b = 400$ available bands. By this, the correlation pattern among lines was estimated based on their unique reflectance fingerprints along the whole spectrum.

For the prediction scenario S1B (described below), the advantages of integrating different information sources to improve the predictive ability of DMY were assessed following the procedures described in Galán et al. (2020b). For this, genetic and hyperspectral data were combined in a multi-kernel prediction model (G + H), which was further extended to a bivariate model (Bivariate_G + H) by incorporating PH a as predictor.

All third-stage prediction models were fitted using the *sommer* package in R (Covarrubias-Pazaran 2016), except model (1), which was fitted within the R package *ASReml-R* v. 3.0 (Gilmour et al. 2009).

Prediction schemes

To address the objectives of the present study, nine bi-parental families with a size of 24 to 32 individuals (Suppl. Fig. S2) and their parental components were divided into TRN and VAL following different schemes. The family 4 × 5 was not considered due to its reduced size ($n = 4$). The TRN composition varied in a controlled manner for testing the effect of the relatedness between this set and VAL on a genotypic level (S1) and both genotypic and environmental levels simultaneously (S2). An overview of the different prediction schemes is given in Table 1.

In S1, three different scenarios were analyzed, namely S1CV, S1A, and S1B, which have a decreasing genotypic relationship between TRN and VAL. Scenario S1CV consisted in ninefold cross-validation (CV) of the whole data set (the nine bi-parental families and their parental components), with eight folds were used for model training and the remaining fold for validation purposes. In contrast, in S1A and S1B, a leave-one-out (LOO) family validation scheme was followed. Here, TRN ranged from six to eight bi-parental families, and VAL consisted of single families with variable size. Whereas in S1A half-sibs (HS) and unrelated lines (UR) were sampled in TRN, in S1B, it included only UR. Parental lines of VAL were available for model training only in S1CV. In contrast, under S1A and S1B, the parents of the validation family were excluded from TRN. The remaining eight parents were considered as UR and could be incorporated accordingly. Genotypes were classified as "unrelated" to distinguish this cross type from FS and HS and, therefore, this term does not have the same meaning as in a population genetics.

To avoid the influence of the TRN size on the prediction ability, for all three scenarios, the TRN size was fixed to 174, which was the largest possible common size among scenarios. If TRN was initially larger than 174, a random sampling without replacement was conducted among possible candidates in order to achieve the targeted size of 174. This procedure was repeated 9,000 times, each repetition consisting of a random composition of TRN to assess model error. The phenotypic and hyperspectral data included in S1 validation scenarios were adjusted across the same four (within-year analysis) or eight (combined years analysis) environments.

In S2, the predictive power of models was assessed by a LOO family validation scheme, as described above for S1. An important difference between S1 and S2 scenarios, is that for S2, data not connected to TRN, either by environments nor by genotypes, was used as VAL. For this, data for model training was collected on UR from six to seven bi-parental families at three or seven environments, while validation data came from single families of variable size evaluated at a fourth or eighth disconnected environment, for the within-year or combined years predictions, respectively.

Table 1 Overview over the validation scenarios (TRN, training set; VAL, validation set; UR, Unrelated; HS, Half sibs; FS, Full sibs; P: Parental lines)

Name	TRN[a]	VAL	Relationship	No. environments sampled[b]	
				TRN	VAL
S1CV	8 random folds	1 Random fold	UR + HS + FS + P	8	8
S1A	8 families	1 Family	UR + HS	8	8
S1B	6 or 7 families	1 Family	UR	8	8
S2	6 or 7 families	1 Family	UR	7	1

[a]The TRN size remained constant across all S1-scenarios ($n = 174$)

[b]Corresponds to combined years predictions

For all prediction schemes, prediction ability was assessed as the Pearson's coefficients of correlation *r* between predicted breeding values and observed BLUEs derived from the combined analysis across environments for S1 and data adjustment within single environments for S2.

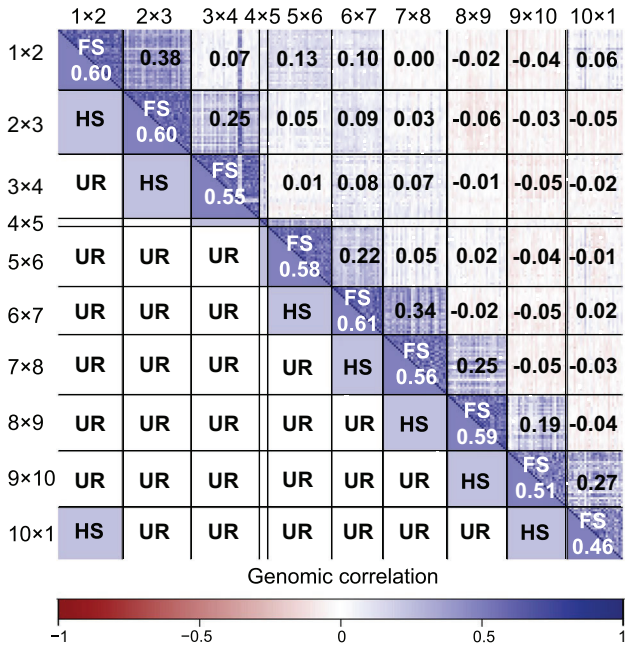


Fig. 1 Heatmap showing the relatedness based on prior pedigree information (below diagonal) and the genomic correlation (above diagonal) among 264 rye lines distributed among ten bi-parental families. The numbers in the blocks refer to average genomic correlations between all pairs of individuals. FS, full sibs; HS, half sibs; UR, unrelated (color figure online)

Results

Population structure, phenotypic and hyperspectral data analysis

The population showed a genomic correlation pattern (Fig. 1), which clearly reproduces the SRR mating design used in the present study (Suppl. Fig. S2). Thus, the mean genomic relationship among full sibs (FS), half sibs (HS), and unrelated lines (UR) followed the expected decay based on prior pedigree information (Fig. 2a). Nevertheless, a substantial overlap between the r_{GC} values from HS with FS and UR was observed. The mean r_{GC} among the nine FS families was 0.55, with a range from 0.61 to 0.46. For HS, the average r_{GC} was 0.27, almost the mean between FS and UR (0.01). The highest r_{GC} among HS was 0.38, while the smallest correlation coefficient was 0.06. Among UR, r_{GC} ranged from 0.04 to − 0.04. Interestingly, no clear distinction among lines could be drawn based on reflectance data (Fig. 2b). The r_{HC} for FS, HS, and UR was close to zero, with mean estimates equal to 0.07, zero, and − 0.09, respectively.

In 2017, FMY and DMY had higher mean estimates than in 2018 (Table 2). In the first year, these values were 355.96 dt ha^{-1} and 124.18 dt ha^{-1}, respectively, whereas in the second year, they dropped correspondingly to 304.82 dt ha^{-1} and 114.68 dt ha^{-1}. The contrary was observed for DMC, which showed a higher mean in 2018 (38.84%) than

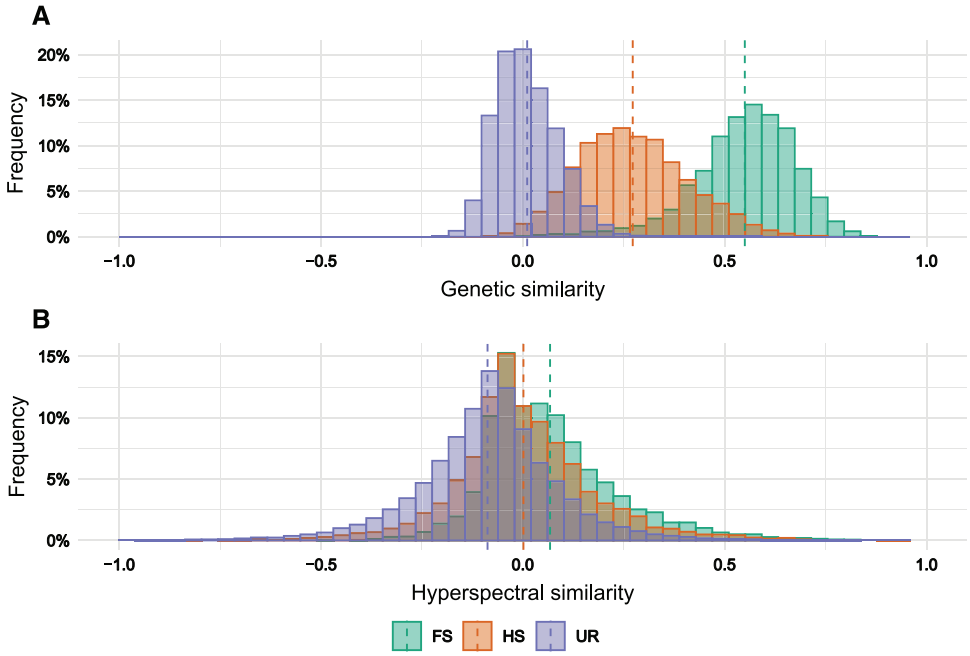


Fig. 2 Histograms of (A) genetic similarity and (B) hyperspectral similarity for full sibs (FS), half sibs (HS), and unrelated (color figure online)

Table 2 Means, ranges, estimates of variance components (genotypic, σ^2_g; genotype-by-location interaction, σ^2_{gl}; genotype-by-year-by-location interaction, σ^2_{gyl}; and residual error σ^2_ε), heritabilities H^2 determined from 274 winter rye hybrids assessed in two years, which were individually or combined analyzed

Trait[a]	Means and ranges			Variance components				H^2
	Mean	Min	Max	σ^2_g	σ^2_{gl}	σ^2_{gly}	σ^2_ε	
2017								
FMY (dt ha^{--1})	355.96	332.85	386.17	41.61***	43.76***	–	190.02	0.56
DMY (dt ha^{-1})	124.18	116.63	131.74	4.97***	6.04***	–	16.62	0.53
DMC (%)	35.24	34.02	37.06	0.23***	0,04***	–	0.36	0.80
2018								
FMY (dt ha^{-1})	304.82	284.92	323.87	25.80***	27.15***	–	213.14	0.46
DMY (dt ha^{-1})	114.68	105.85	122.31	5.85***	6.79***	–	26.22	0.54
DMC (%)	38.84	37.21	40.62	0.27***	0.13***	–	1.51	0.70
Combined								
FMY (dt ha^{-1})	330.68	312.29	351.91	21.31***	15.15***	19.04***	203.13	0.47
DMY (dt ha^{-1})	119.48	113.31	126.33	3.41***	2.54***	3.64***	21.49	0.50
DMC (%)	37.02	35.74	38.45	0.23***	0.02	0.07***	0.94	0.81

[a]Traits are fresh matter yield (FMY), dry matter yield (DMY), and dry matter content (DMC)

***Significant at the 0.001 probability level

in 2017 (35.24%). The estimated genotypic variance (σ^2_g) was significantly greater than zero ($p < 0.001$) for all traits. With one minor exception, the same holds for the genotype-by-location interaction (σ^2_{gl}) and genotype-by-location-by-year interaction (σ^2_{gly}) variances. The estimates of H^2 were in general higher in 2017 than in 2018. DMC displayed the higher H^2 estimates, which ranged from 0.70 to 0.81, whereas H^2 for FMY and DMY varied from 0.46 to 0.56. Across the analyzed hyperspectral spectrum, H^2 was highly heterogeneous. The mean value across the 32 selected hyperspectral bands (Suppl. F4) was higher when both years were analyzed together ($H^2 = 0.63$), followed by 2018 ($H^2 = 0.54$) and 2017 ($H^2 = 0.43$). Mean correlations with agronomic traits (considering absolute values) were rather low for all traits ($r \leq |0.16|$) with relatively broad ranges (up to $r \leq |0.41|$, Suppl. Fig. S4). FMY mostly displayed the highest correlation estimates, followed by DMY and DMC.

Prediction abilities under declining genotypic relationships (S1)

Overall, HBLUP was significantly more accurate than GBLUP for FMY and DMY, while the opposite was observed for DMC (Fig. 3). Combining the data across years was beneficial for HBLUP for all traits, and in the case of GBLUP only for DMC, while for FMY and DMY, GBLUP was mostly more accurate within single-year analysis. The highest prediction abilities for all models and traits were observed under validation scenario S1CV, which has the closest relationship between genotypes used for model training and validation, followed by scenarios S1A and S1B (Fig. 3). However, the impact of a reduced degree of genetic relatedness between TRN and VAL on the prediction ability was unequal between models. Interestingly, the reduction of the predictive power of GBLUP was considerably higher than for HBLUP when validated on less related sets. For instance, under S1CV for DMY adjusted across years (Fig. 3), GBLUP showed a prediction ability of 0.56, while it dropped to 0.37 and 0.20 under scenarios S1A and S1B, respectively. This represents a decay of about one third and two thirds, respectively. In contrast, the prediction abilities of HBLUP were more stable since they ranged between 0.58 (S1CV) and 0.51 (S1B). Thus, HBLUP retained about 90% of the predicted power shown under S1CV when predicting UR genotypes (S1B). A similar trend was observed for all other traits within as well as across years. The prediction of DMY in validation scenario S1B could be further enhanced by a bivariate model combining hyperspectral and genomic data as well as PH (up to 0.58, Fig. 4). In contrast, the predictive power of the multi-kernel model was very similar to the achieved by HBLUP, although a slight reduction in the variability of the predictions was observed in 2017 and combined-years analysis.

Predicting environmentally and genetically unconnected candidates (S2)

In the present study, the prediction models were additionally trained on unrelated data, either at a genotypic or environmental level, with the data used for model validation (scenario S2). Under S2, the prediction abilities of all models for all traits was significantly lower and displayed broader ranges when compared to their performance under S1 (Fig. 3). Under S2, HBLUP was also mostly more accurate than GBLUP for FMY and DMY, while the opposite was

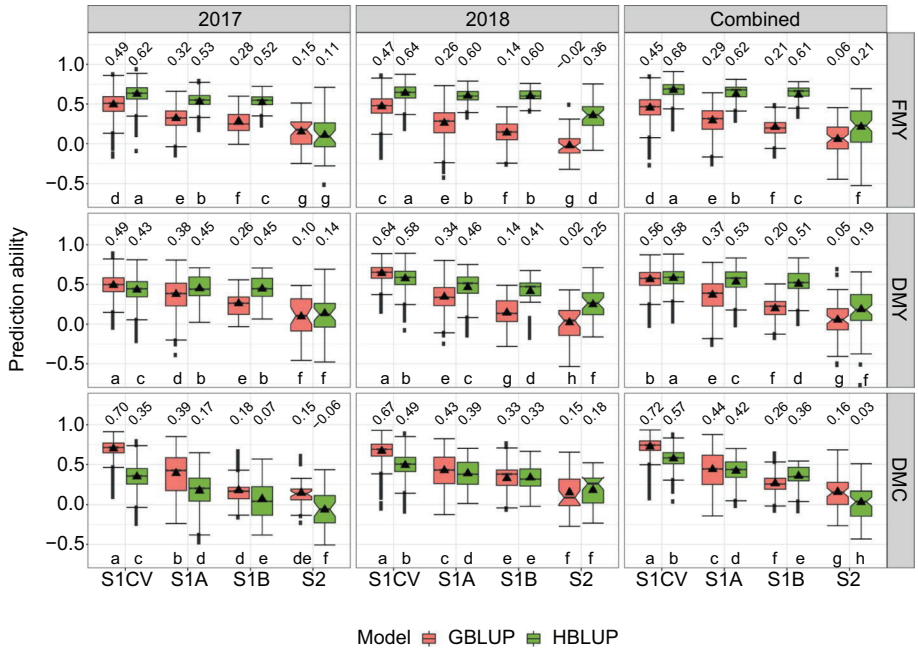


Fig. 3 Prediction abilities for fresh matter yield (FMY), dry matter yield (DMY), and dry matter content (DMC) of genomic (GBLUP) and hyperspectral (HBLUP) best linear unbiased predictions under four different validation schemes assessed across two years (2017 and 2018), which were individually and combined analyzed. Mean values are shown above each box plot and by black triangles and are significantly different, within each subplot, when no letter in common is shared (Tukey's honestly significant difference test; $\alpha = 0.01$) (color figure online)

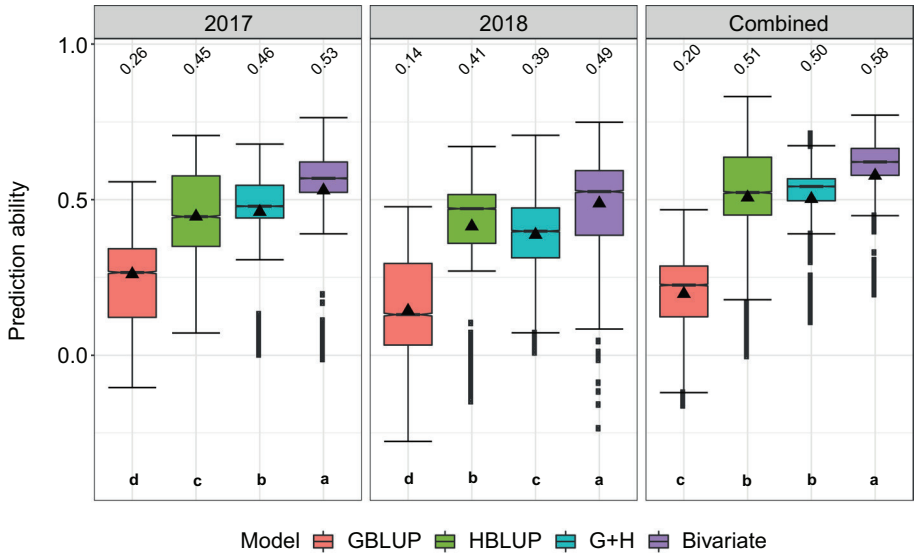


Fig. 4 Prediction abilities for dry matter yield of single-kernel (Genomic best linear unbiased predictor, GBLUP and Hyperspectral best linear unbiased predictor, HBLUP), multi-kernel (G+H), and bivariate (Bivariate_G+H) models assessed across two years (2017 and 2018), which were individually and combined analyzed. Models were tested under validation scenario S1B. Mean values are shown above each box plot and by black triangles and are significantly different, within each subplot, when no letter in common is shared (Tukey's honestly significant difference test; $\alpha = 0.01$) (color figure online)

observed for DMC. Ranges of mean prediction ability of HBLUP for FMY, DMY, and DMC were from 0.11 to 0.36, from 0.14 to 0.25, and from -0.06 to 0.18, respectively, while for GBLUP ranges lay between – 0.02 to 0.15, 0.02 to 0.10, and 0.15 to 0.16, respectively. In contrast to S1, no clear benefits of combining the data collected across years were observed under the S2 validation scheme.

Discussion

The accurate prediction of biomass at early stages via indirect selection for DMY based on GY trials is a fundamental requirement for the implementation of a resource-efficient dual-purpose breeding program in rye. In this way, the entire genetic variance could be exploited, leveraging

the expected selection gain. In our breeding program, each year represents a new selection cycle, where genotypes with different genetic backgrounds are evaluated in new GCA trials. The prediction of subsequent selection cycles implies an additional challenge since the data used for model training and validation are highly unconnected. Nevertheless, it is mainly under this scenario that breeding programs can benefit the most because the biomass improvement can be conducted at the first stage of testcross evaluation without an increase of the number of field plots. The objective of this study was, therefore, to assess and compare the prediction ability of genetic and hyperspectral data under varying genetic relationships between the training and validation sets.

Influence of the genetic composition of the TRN and traits characteristics

The degree of relatedness between individuals used for model training (TRN) and validation (VAL) directly influenced the prediction ability of all models; however, this impact was remarkably lower for HBLUP than for GBLUP (Fig. 3). The prediction abilities observed under scenario S1CV can be considered as an upper limit, where model training is performed across FS, HS, UR, and parental lines of genotypes used for model validation. Then, a systematic reduction in the predictive performance of all models accompanied the exclusion of genotypes genetically closest to VAL. The exclusion of FS and parental lines from TRN (S1A) represented, averaged across single and combined years analyzes and traits, a reduction of about 40% on the performance of GBLUP, while the further removal of HS signified an additional penalization of around 20%. The larger drop in the prediction abilities observed for S1A compared to those of S1B can be explained by the asymmetrical relevance of using closest relatives for genomic model training (Albrecht et al. 2011; Technow et al. 2014; Juliana et al. 2019). In contrast to GBLUP, the penalization observed for HBLUP in S1A was, on average, only nearly 15% and an additional 6% in S1B, allowing this model to show the highest prediction abilities between the single-kernel models in these scenarios. Model performance was also dissimilar across traits. GBLUP showed mostly the higher abilities for DMC, whereas HBLUP performed better for FMY and DMY. The differences in predictive abilities are most likely a consequence of both trait H^2 and the different information sources used by GS and reflectance-based models.

To adequately predict the performance of untested candidates, genomic models exploit the genetic relationships between them and individuals whose genotypic and phenotypic information is available, as previously shown in many empirical and simulation-based studies in animal and plant breeding (Habier et al. 2007 2010; Roos et al. 2009; Pszczola et al. 2012; Riedelsheimer et al. 2013; Würschum et al. 2013; Crossa et al. 2014; Lehermeier et al. 2014; Technow et al. 2014; Wang et al. 2014; Thorwarth et al. 2017; Herter et al. 2019). In line with these observations, our results also showed that the predictive power of GS dropped substantially when predictions are made among lowly related populations. For predictions across subsequent cycles in rye, GS could represent a suitable strategy when TRN is represented by aggregated multi-year data from several cycles (Auinger et al. 2016; Bernal-Vasquez et al. 2017). Nevertheless, the authors of these papers concluded that GS still relies heavily on a sufficient relationship between predicted candidates and those used for model training. Selection cycles need, for instance, to be connected by a sufficient number of common ancestors. This prerequisite may not be easily fulfilled in practical rye breeding since subsequent breeding cycles usually are largely unconnected. In addition, the success of GS depends, among others, on trait related factors, such as heritability (Jia and Jannink 2012; Marulanda et al. 2015). Thus, the better and less variable GBLUP performance observed for DMC (Fig. 3) is likely explained by the larger H^2 estimated for this trait in comparison to FMY and DMY (Table 2).

The reflectance fingerprints of the genotypes were more similar than their allelic status across relationship groups (Fig. 2), suggesting that the information imprinted among the spectrum is less sensitive to genetic distinctiveness among individuals than molecular data. These observations can explain why reflectance data allowed higher prediction abilities than marker data under decreased genetic relationships between TRN and VAL. In contrast to GBLUP, more highly heritable traits were not better predicted by HBLUP. In turn, for HBLUP to perform well, plant canopies should display specific absorption patterns related to some extent to the trait of interest as shown, for instance, by the correlations between the analyzed traits and bands. The most effectively predicted traits (FMY and DMY) showed higher correlations than the lowest predicted trait (DMC, Suppl. Fig. S4). Thus, the higher performance of HBLUP for FMY and DMY might be explained by the higher informativeness of the collected reflectance data for those traits than for DMC. Since the absorption of water and DMC is almost constant across the visible spectrum and the absorbance of these two features starts around 950 nm (Jacquemoud et al. 2000), where our spectrum was from 410 nm to 993 nm, further research could investigate the prospects of HBLUP based on reflectance data beyond 1000 nm to better predict DMC.

Several strategies have been investigated for taking advantage of reflectance data in predictive breeding.

Summarizing the reflectance characteristics of plants into simple vegetation indices (VIs) has been proposed to assess vegetation characteristics of interest like grain and biomass yields under different environmental conditions (Xue and Su 2017). However, prediction models benefited the most by the exploitation of whole-spectrum data (Aguate et al. 2017; Montesinos-López et al. 2017b; Krause et al. 2019; Galán et al. 2020b). Recently, highly heritable VIs genetically correlated with the trait of interest such as the Normalized Difference Vegetation Index (NDVI; Rouse et al. 1974; Tucker 1979) and the green NDVI (GNDVI; Gitelson et al. 1996), have been incorporated as secondary traits into multivariate pedigree and genomic prediction models to increase accuracy within the same wheat population and selection cycle (Rutkoski et al. 2016; Sun et al. 2017) as well as across selection cycles composed by closely related populations (Sun et al. 2019). Juliana et al. (2019) found that similar multivariate equations were superior to univariate genomic prediction models when predicting across populations and years. Still, the relationship between TRN and VAL was found crucial also for multivariate models, although the populations used for model training and validation were genetically related to some extent, and predictions were made within the same stressed environments. The results of the present study also showed that combining hyperspectral and genomic data in a multi-kernel model yielded only limited advantages over HBLUP for DMY prediction of less related progenies (Fig. 4). In this context, the prediction ability for DMY could be further increased up to 20% by a bivariate model including also PH. Nevertheless, the performance of G + H and the bivariate model in the present study were lower than when used for DMY prediction of highly related rye progenies, as reported in a previous research (Galán et al. 2020b). These findings reveal, on the one hand, the advantages of incorporating HTP data into prediction routines, and, on the other hand, the limits of GS in the context of across cycle predictions.

Prediction of new genotypes in untested environments

In validation scenario S1B, UR genotypes were assessed across the same environments (Table 1). In contrast, in S2, unrelated individuals were tested under new environmental conditions, allowing the simultaneous assessment of the genotypic and environmental sampling on the predictive power of marker- and hyperspectral-based models. Predictive abilities in S2 were significantly lower than in S1B, suggesting that predicting the performance of genetically and environmentally highly unconnected individuals is challenging. This is consistent with studies showing that the prediction of new candidates is less accurate when model training is performed without borrowing information of environments correlated to the one used for validation (Crossa et al. 2014; Krause et al. 2019). These poor predictions obtained in S2 might be explained by the substantial genotype-by-environment interactions (G × E) estimated for the predicted traits (Table 2) as well as by the high variability observed for hyperspectral bands among environments, resulting mainly from the extremely different conditions observed between growing seasons as reported in a previous study (Galán et al. 2020b). It seems, therefore, plausible that heterogeneous marker-to-trait and band-to-trait (Montesinos-López et al. 2017a) signals among environments adversely affected the prediction abilities from GBLUP and HBLUP. Therefore, to adequately predict untested genotypes under new environmental conditions, prediction equations need to be extended by environmental and genetic covariates for proper G × E modeling (Piepho 2009; Burgueño et al. 2012; Resende et al. 2020).

A forward-validation approach aims to predict the performance of new genotypes by exploiting the data from previous years (Bernal-Vasquez et al. 2017). Considering our breeding scheme (Suppl. Fig.1), data for model training could be obtained from split GCA-2 trials with biomass and grain yield plots, whereas model validation could be performed on GCA-1 data from a subsequent selection cycle. It should be kept in mind that we need large-drilled plots for biomass model training because this trait cannot be reliably measured on smaller observation plots. As different selection cycles involve new individuals from multiple genetic backgrounds, and usually hardly any common progeny is shared across cycles, the genetic relationship between the data used for model training and validation across cycles is expected to be substantially lower than in within-cycle predictions.

However, data used in across-cycles predictions is environmentally connected because GCA-2 genotypes are tested more intensively in a larger number of locations, within which the same environments as in GCA-1 are typically found. In practical plant breeding, large testing locations within the targeted environment are common, since they are more efficient in terms of logistics, trained personnel requirements, as well as field evaluation and management. In these testing sites, yield trials from different stages are planted next to each other, being reliable, large- scaled training data readily available for model calibration. Thus, scenario S1B mimics this practical situation much better than S2. Our results showed that, in this context, models incorporating hyperspectral data emerge as a promising strategy to achieve superior improvements in DMY in hybrid rye. Still, the relevance of S2 outcomes lies in a consistently unbiased estimation of the prediction abilities of the models (Utz et al. 2000), revealing the high impact of G × E not only on GBLUP but also on HBLUP.

Conclusions for biomass breeding in hybrid rye

Traditionally, biomass is estimated destructively at an earlier growth stage, preventing grain yield from being evaluated in those same plots. The effective indirect assessment of biomass at the early stages of the breeding program is crucial to entirely untap the potential of rye as a dual-purpose crop affordably. In this sense, prediction models accurately estimating the biomass yield of genotypes of diverse genetic backgrounds across selection cycles represent a valuable tool. In the present study, GBLUP achieved acceptable prediction abilities only for highly heritable traits across closely related individuals. In contrast, HBLUP was substantially less affected by genetic relatedness and trait heritability emerging as a suitable approach for predicting complex traits across highly distinct populations.

Considering that in modern plant breeding genomic information is usually already available before the candidate lines are evaluated as testcrosses in the expensive GCA trials, breeders usually perceive marker and HTP data as a complement, rather than an alternative. Here, HTP offers the possibility of screening large-scale field trials with reduced capital and time expenditures, than conventional methods (e.g. destructive sampling and visual scores). Moreover, combining hyperspectral, genomic, and PH in bivariate models allows more effective DMY predictions of genotypes showing low genetic connectivity to ones used for model training. The bivariate model here presented is flexible and allows the incorporation of GY and other correlated traits to DMY aimig superior predictive power. Nonetheless, by including several predictors, the complexity of the models increases in proportion.

Our results also show that not only GBLUP but also HBLUP was largely affected by G×E interactions, resulting in poor to negligible predictive power when the environments used for model training and validation were different. To fully exploit the advantages of hyperspectral-based models, it is, therefore, highly recommended to incorporate reflectance fingerprints of genotypes collected in the respective environment. Our study demonstrates the capability of hyperspectral-enabled predictions to leverage selection gains to meet the increasing demand for sustainable biomass sources worldwide. Lastly, the prospects of HTP as an economical alternative to traditional biomass sampling are expected to increase in proportion to future improvements in terms of image data acquisition and management.

Supplementary Information The online version contains supplementary material available at(https://doi.org/10.1007/s00122-021-03779-1) .

Acknowledgements This study was funded by the German Federal Ministry of Food and Agriculture (BMEL) through the German Agency for Renewable Resources (FNR), grant number FKZ 22019716 to TM. We gratefully acknowledge the excellent support of the technical staff at each experimental station. We are particularly grateful to Hans-Otto Wegener, Jörn-Claus Gudehus, Karsten Sell, KWS LOCHOW GmbH, Bergen, Germany, for seed production and conducting field trials. We also thank Dr. Peer Wilde, KWS LOCHOW GmbH, Bergen, for his valuable contributions to this project.

Author contributions statement RG analyzed the data and the wrote the manuscript, AMBV, HPP, and PT supported with statistical advice, CJ conducted hyperspectral phenotyping at all environments, PS supervised data collection at Wohlde, Prislich, and Bernburg and provided scientific advice, TM and AG designed the research project and edited the manuscript. All the authors read and approved the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards The authors declare that the experiments comply with the current laws of Germany.

References

Aguate FM, Trachsel S, Pérez LG, Burgueño J, Crossa J, Balzarini M, Gouache D, Bogard M, Gdl C (2017) Use of hyperspectral image data outperforms vegetation indices in prediction of maize yield. Crop Sci 57(5):2517–2524. https://doi.org/10.2135/cropsci2017.01.0007

Albrecht T, Wimmer V, Auinger H-J, Erbe M, Knaak C, Ouzunova M, Simianer H, Schön C-C (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123(2):339. https://doi.org/10.1007/s00122-011-1587-7

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19(1):52–61. https://doi.org/10.1016/j.tplants.2013.09.008

Auinger H-J, Schönleben M, Lehermeier C, Schmidt M, Korzun V, Geiger HH, Piepho H-P, Gordillo A, Wilde P, Bauer E (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.). Theor Appl Genet. 129(11):2043–2053. https://doi.org/10.1007/s00122-016-2756-5

Babar MA, Reynolds MP, van Ginkel M, Klatt AR, Raun WR, Stone ML (2006) Spectral reflectance to estimate genetic variation for in-season biomass, leaf chlorophyll, and canopy temperature in

wheat. Crop Sci 46(3):1046–1057. https://doi.org/10.2135/cropsci2005.0211
Barmeier G, Schmidhalter U (2017) High-throughput field phenotyping of leaves, leaf sheaths, culms and ears of spring barley cultivars at anthesis and dough ripeness. Front in Plant Sci 8:1920. https://doi.org/10.3389/fpls.2017.01920
Bernal-Vasquez A-M, Utz H-F, Piepho H-P (2016) Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. Theor Appl Genet 129(4):787–804. https://doi.org/10.1007/s00122-016-2666-6
Bernal-Vasquez A-M, Gordillo A, Schmidt M, Piepho H-P (2017) Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC Genet 18(1):51. https://doi.org/10.1186/s12863-017-0512-8
Bundessortenamt (2019) Beschreibende Sortenliste Getreide, Mais, Öl-und Faserpflanzen, Leguminosen, Rüben. Zwischen-früchte, Hannover, Bundessortenamt
Burgueño J, de Los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52(2):707–719. https://doi.org/10.2135/cropsci2011.06.0299
Busemeyer L, Ruckelshausen A, Möller K, Melchinger AE, Alheit KV, Maurer HP, Hahn V, Weissmann EA, Reif JC, Würschum T (2013) Precision phenotyping of biomass accumulation in triticale reveals temporal genetic patterns of regulation. Sci Rep 3:2442
Cabrera-Bosquet L, Crossa J, von Zitzewitz J, Serret MD, Luis Araus J (2012) High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. J Integr Plant Biol 54(5):312–320. https://doi.org/10.1111/j.1744-7909.2012.01116.x
Covarrubias-Pazaran G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. PLoS ONE. https://doi.org/10.1371/journal.pone.0156744
Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112(1):48–60. https://doi.org/10.1038/hdy.2013.16
de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. Genetics 183(4):1545–1553. https://doi.org/10.1534/genetics.109.104935
EEG (2017) Gesetz für den ausbau erneuerbarer energien (Erneuerbare-Energien-Gesetz - EEG). http://www.gesetze-im-internet.de/eeg_2014/EEG_2017.pdf. Accessed 02 Nov 2019
Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics,4th edn. Longman, Essex
Fahlgren N, Gehan MA, Baxter I (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Curr Opin Plant Biol 24:93–99
FAO (2019) FAOSTAT database. Food and agriculture organization of the united nations. http://www.fao.org/faostat/en/#data/QC. Accessed 05 Nov 2019
European Commission (2011) Energy roadmap 2050. communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions, com (2011) 885 final. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0885:FIN:EN:PDF. Accessed 02 Nov 2019
Finkel E (2009) With 'phenomics', plant scientists hope to shift breeding into overdrive. Science 325(5939):380–381. https://doi.org/10.1126/science.325_380
Fiorani F, Schurr U (2013) Future scenarios for plant phenotyping. Annu Rev Plant Biol 64:267–291. https://doi.org/10.1146/annurev-arplant-050312-120137
Fu Y, Yang G, Wang J, Song X, Feng H (2014) Winter wheat biomass estimation based on spectral indices, band depth analysis and partial least squares regression using hyperspectral measurements. Comput Electron Agric 100:51–59. https://doi.org/10.1016/j.compag.2013.10.010
Furbank RT, Tester M (2011) Phenomics–technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16(12):635–644. https://doi.org/10.1016/j.tplants.2011.09.005
Galán RJ, Bernal-Vasquez A-M, Jebsen C, Piepho H-P, Thorwarth P, Steffan P, Gordillo A, Miedaner T (2020a) Hyperspectral reflectance data and agronomic traits can predict biomass yield in winter rye hybrids. BioEnergy Res 13(1):168–182. https://doi.org/10.1007/s12155-019-10080-z
Galán RJ, Bernal-Vasquez A-M, Jebsen C, Piepho H-P, Thorwarth P, Steffan P, Gordillo A, Miedaner T (2020b) Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye. Theor Appl Genet. https://doi.org/10.1007/s00122-020-03651-8
Geiger HH, Miedaner T (2009) Rye breeding. In: Carena MJ (ed) Cereals, vol 3. Springer, New York, pp 157–181
Gilmour AR, Gogel BJ, Cullis BR, Thompson R, Butler D (2009) ASReml user guide release 3.0. VSN International Ltd, UK
Gitelson AA, Kaufman YJ, Merzlyak MN (1996) Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens Environ 58(3):289–298
Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397. https://doi.org/10.1534/genetics.107.081190
Habier D, Tetens J, Seefried F-R, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42(1):5. https://doi.org/10.1186/1297-9686-42-5
Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194(3):597–607. https://doi.org/10.1534/genetics.113.152207
Haffke S, Kusterer B, Fromme FJ, Roux S, Hackauf B, Miedaner T (2014) Analysis of covariation of grain yield and dry matter yield for breeding dual use hybrid rye. BioEnergy Res 7(1):424–429. https://doi.org/10.1007/s12155-013-9383-7
Herter CP, Ebmeyer E, Kollers S, Korzun V, Würschum T, Miedaner T (2019) Accuracy of within-and among-family genomic prediction for Fusarium head blight and Septoria tritici blotch in winter wheat. Theor Appl Genet 132(4):1121–1135. https://doi.org/10.1007/s00122-018-3264-6
Jacquemoud S, Bacour C, Poilve H, Frangi J-P (2000) Comparison of four radiative transfer models to simulate plant canopies reflectance: direct and inverse mode. Remote Sens Environ 74(3):471–481. https://doi.org/10.1016/S0034-4257(00)00139-5
Jia Y, Jannink J-L (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192(4):1513–1522. https://doi.org/10.1534/genetics.112.144246
Juliana P, Montesinos-López OA, Crossa J, Mondal S, González Pérez L, Poland J, Huerta-Espino J, Crespo-Herrera L, Govindan V, Dreisigacker S, Shrestha S, Pérez-Rodríguez P, Pinto Espinosa F, Singh RP (2019) Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. Theor Appl Genet 132(1):177–194. https://doi.org/10.1007/s00122-018-3206-3
Krause MR, González-Pérez L, Crossa J, Pérez-Rodríguez P, Montesinos-López O, Singh RP, Dreisigacker S, Poland J, Rutkoski J, Sorrells M, Gore MA, Mondal S (2019) Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat. G3 Genes Genomes Genet 9(4):1231–1247. https://doi.org/10.1534/g3.118.200856
Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, Flament P, Melchinger AE, Menz M, Meyer N, Moreau L, Moreno-González J, Ouzunova M, Pausch H, Ranc N, Schipprack

W, Schönleben M, Walter H, Charcosset A, Schön C-C (2014) Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. Genetics 198(1):3–16. https://doi.org/10.1534/genetics.114.161943

Li L, Zhang Q, Huang D (2014) A review of imaging techniques for plant phenotyping. Sensors 14(11):20078–20111. https://doi.org/10.3390/s141120078

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10(8):565–577. https://doi.org/10.1038/nrg2612

Marulanda JJ, Melchinger AE, Würschum T (2015) Genomic selection in biparental populations: assessment of parameters for optimum estimation set design. Plant Breed 134(6):623–630. https://doi.org/10.1111/pbr.12317

Meier U (1997) Growth stages of mono- and dicotyledonous plants. Blackwell Wissenschafts-Verlag, Berlin

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Miedaner T, Koch S, Seggl A, Schmiedchen B, Wilde P (2012) Quantitative genetic parameters for selection of biomass yield in hybrid rye. Plant Breed 131(1):100–103. https://doi.org/10.1111/j.1439-0523.2011.01909.x

Miedaner T, Korzun V, Bauer E (2019) Genomics-based hybrid rye breeding. In: Miedaner T, Korzun V (eds) Applications of genetic and genomic research in cereals. Elsevier, Amsterdam, Netherlands, pp 329–348

Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S (2015) LinkImpute fast and accurate genotype imputation for nonmodel organisms. G3 Genes Genomes Genet 5(11):2383–2390. https://doi.org/10.1534/g3.115.021667

Montes JM, Melchinger AE, Reif JC (2007) Novel throughput phenotyping platforms in plant genetic studies. Trends Plant Sci 12(10):433–436. https://doi.org/10.1016/j.tplants.2007.08.006

Montes JM, Technow F, Dhillon BS, Mauch F, Melchinger AE (2011) High-throughput non-destructive biomass determination during early plant development in maize under field conditions. Field Crops Res 121(2):268–273

Montesinos-López A, Montesinos-López OA, Cuevas J, Mata-López WA, Burgueño J, Mondal S, Huerta J, Singh R, Autrique E, González-Pérez L, Crossa J (2017a) Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper-spectral image data. Plant Methods 13:62. https://doi.org/10.1186/s13007-017-0212-4

Montesinos-López OA, Montesinos-López A, Crossa J, de Los Campos G, Alvarado G, Suchismita M, Rutkoski J, González-Pérez L, Burgueño J (2017b) Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. Plant Methods 13(1):4. https://doi.org/10.1186/s13007-016-0154-2

Piepho H-P (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49(4):1165–1176. https://doi.org/10.2135/cropsci2008.10.0595

Piepho H-P, Möhring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177(3):1881–1888. https://doi.org/10.1534/genetics.107.074229

Piepho H-P, Moehring J, Schulz-Streeck T, Ogutu JO (2012) A stage-wise approach for the analysis of multi-environment trials. Biom J 54(6):844–860. https://doi.org/10.1002/bimj.201100219

Pszczola M, Strabel T, Mulder HA, Calus MPL (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci 95(1):389–400. https://doi.org/10.3168/jds.2011-4338

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Resende RT, Piepho H-P, Silva-Junior OB, Silva FF, Resende MDV, Grattapaglia D (2020) Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. Theor Appl Genet

Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink J-L, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194(2):493–503. https://doi.org/10.1534/genetics.113.150227

Rincent R, Charpentier J-P, 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(12):3961–3972. https://doi.org/10.1534/g3.118.200760

Rouse JW, Haas RH, Schell JA, Deering DW (1974) Monitoring vegetation systems in the Great Plains with ERTS. Third ERTS Symposium, NASA SP-351:309–3017

Rutkoski J, Poland J, Mondal S, Autrique E, Pérez LG, Crossa J, Reynolds M, Singh R (2016) Canopy temperature and vegetation indices from highthroughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3 Genes Genomes Genet 6(9):2799–2808. https://doi.org/10.1534/g3.116.032888

Sun J, Rutkoski JE, Poland JA, Crossa J, Jannink J-L, Sorrells ME (2017) Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. Plant Genom. https://doi.org/10.3835/plantgenome2016.11.0111

Sun J, Poland JA, Mondal S, Crossa J, Juliana P, Singh RP, Rutkoski JE, Jannink J-L, Crespo-Herrera L, Velu G, Huerta-Espino J, Sorrells ME (2019) High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. Theor Appl Genet 132(6):1705–1720. https://doi.org/10.1007/s00122-019-03309-0

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197(4):1343–1355. https://doi.org/10.1534/genetics.114.165860

Thorwarth P, Ahlemeyer J, Bochard A-M, Krumnacker K, Blümel H, Laubach E, Knöchel N, Cselényi L, Ordon F, Schmid KJ (2017) Genomic prediction ability for yield-related traits in German winter barley elite material. Theor Appl Genet 130(8):1669–1683. https://doi.org/10.1007/s00122-017-2917-1

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc: Ser B (Methodol) 58(1):267–288

Tucker CJ (1979) Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens Environ 8(2):127–150

Union E (2010) Communication from the Commission on the practical implementation of the EU biofuels and bioliquids sustainability scheme and on counting rules for biofuels (2010/C 160/02). Off J Eur Un 2:8–16

Utz HF, Melchinger AE, Schön CC (2000) Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. Genetics 154(4):1839–1849

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423. https://doi.org/10.3168/jds.2007-0980

Verhoeven KJF, Jannink JL, McIntyre LM (2006) Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity 96(2):139–149. https://doi.org/10.1038/sj.hdy.6800763

Wang Y, Mette MF, Miedaner T, Gottwald M, Wilde P, Reif JC, Zhao Y (2014) The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of

marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. BMC Genom 15(1):556. https://doi.org/10.1186/1471-2164-15-556

White JW, Andrade-Sanchez P, Gore MA, Bronson KF, Coffelt TA, Conley MM, Feldmann KA, French AN, Heun JT, Hunsaker DJ (2012) Field-based phenomics for plant genetics research. Field Crops Res 133:101–112. https://doi.org/10.1016/j.fcr.2012.04.003

Wimmer V, Albrecht T, Auinger H-J, Schön C-C (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28(15):2086–2087. https://doi.org/10.1093/bioinformatics/bts335

World bioenergy association (2019) Global bioenergy statistics 2019. https://worldbioenergy.org/uploads/191129%20WBA%20GBS%202019_LQ.pdf. Accessed 17 Jul 2020

Würschum T (2019) Modern field phenotyping opens new avenues for selection. In: Miedaner T, Korzun V (eds) Applications of genetic and genomic research in cereals. Elsevier, Amsterdam, Netherlands, pp 233–250

Würschum T, Reif JC, Kraft T, Janssen G, Zhao Y (2013) Genomic selection in sugar beet breeding populations. BMC Genet 14(1):85. https://doi.org/10.1186/1471-2156-14-85

Xue J, Su B (2017) Significant remote sensing vegetation indices: a review of developments and applications. J Sens 2017:1–17. https://doi.org/10.1155/2017/1353691

Yang G, Liu J, Zhao C, Li Z, Huang Y, Yu H, Xu B, Yang X, Zhu D, Zhang X (2017) Unmanned aerial vehicle remote sensing for field-based crop phenotyping: current status and perspectives. Front Plant Sci 8:1111. https://doi.org/10.3389/fpls.2017.01111

Yue J, Yang G, Li C, Li Z, Wang Y, Feng H, Xu B (2017) Estimation of winter wheat above-ground biomass using unmanned aerial vehicle-based snapshot hyperspectral sensor and crop height improved models. Remote Sens 9(7):708. https://doi.org/10.3390/rs9070708

Yue J, Feng H, Yang G, Li Z (2018) A comparison of regression techniques for estimation of above-ground winter wheat biomass using near-surface spectroscopy. Remote Sens 10(1):66. https://doi.org/10.3390/rs10010066

Theoretical and Applied Genetics

Early prediction of biomass in hybrid rye based on hyperspectral data surpasses genomic predictability in less-related breeding material

Rodrigo José Galán[1], Angela-Maria Bernal-Vasquez[2], Christian Jebsen[2], Hans-Peter Piepho[3], Patrick Thorwarth[1,2], Philipp Steffan[4], Andres Gordillo[4], Thomas Miedaner[1].

[1] State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany.

[2] KWS SAAT SE, Grimsehlstraße 31, 37574 Einbeck, Germany.

[3] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, 70593 Stuttgart, Germany.

[4] KWS LOCHOW GMBH, Ferdinand-von-Lochow Straße 5, 29303 Bergen, Germany.

Corresponding author:

Thomas Miedaner
e-mail: thomas.miedaner@uni-hohenheim.de
Telephone No.: +49 711 459-22690

Available ORCID of the authors:

Angela-Maria Bernal-Vasquez: https://orcid.org/0000-0003-0415-831

Patrick Thorwarth: https://orcid.org/0000-0003-4456-2358

Thomas Miedaner: https://orcid.org/0000-0002-9541-3726

Online resource 1

Phenotypic data analysis

The combined analysis across environments was conducted following model (1) from Galán et al. (2020). For the combined analysis across locations, this mixed model is as follows:

$$\begin{aligned}\gamma \;=\; & G{:}\,L + Y \\ & + L \cdot G + Y \cdot G + Y \cdot L + \; L \cdot Y \cdot G \\ & + ENV \cdot T + ENV \cdot T \cdot R + ENV \cdot T \cdot R \cdot B \; + e \end{aligned} \qquad (1)$$

where γ is the observed genotype performance, G denotes the genotypes, L the locations, Y the years, T the trials within environments ENV (equivalent to year-location combinations), R the replicates within trials, B the blocks within replicates, and e the error associated with the observation γ. Error, trial, block, and replicate variances were assumed heterogeneous among environments. In model (1), the dot operator (·) specifies crossed effects (A·B) and fixed and random terms are separated by a colon (:), with fixed terms appearing first (Piepho et al. 2003). Variance components and pairwise variances of genotype mean (BLUEs) differences (needed for heritability estimation) were estimated by restricted maximum likelihood (REML) for all random effects in model (1). This also holds for estimation of the genotypic variance (σ_g^2), which required an additional analysis fitting the above model with random genotypic effects. Significance of variance component estimates was tested by model comparisons using likelihood ratio tests (Stram and Lee 1994).

Within environments, BLUEs of genotypes were analyzed following model (2) also from Galán et al. (2020). This mixed model is described as follows:

$$\gamma \;=\; G{:}\,T \; + T \cdot R \; + \; T \cdot R \cdot B \; + \; e \qquad (2)$$

This model (2) differs from the first model (1) only in dropping the year and location main effects and corresponding interactions with genotypes. Variance components for single environments were estimated as described previously for model (1).

References for online resource 1 only

Galán RJ, Bernal-Vasquez A-M, Jebsen C, Piepho H-P, Thorwarth P, Steffan P, Gordillo A, Miedaner T (2020) Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye. Theor Appl Genet. doi: 10.1007/s00122-020-03651-8

Piepho H-P, Büchse A, Emrich K (2003) A hitchhiker's guide to mixed models for randomized experiments. J Agron Crop Sci 189(5):310–322

Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. Biometrics:1171–1177

Theoretical and Applied Genetics

Early prediction of biomass in hybrid rye based on hyperspectral data surpasses genomic predictability in less-related breeding material

Rodrigo José Galán[1], Angela-Maria Bernal-Vasquez[2], Christian Jebsen[2], Hans-Peter Piepho[3], Patrick Thorwarth[1,2], Philipp Steffan[4], Andres Gordillo[4], Thomas Miedaner[1].

[1] State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany.

[2] KWS SAAT SE, Grimsehlstraße 31, 37574 Einbeck, Germany.

[3] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, 70593 Stuttgart, Germany.

[4] KWS LOCHOW GMBH, Ferdinand-von-Lochow Straße 5, 29303 Bergen, Germany.

Corresponding author:

Thomas Miedaner
e-mail: thomas.miedaner@uni-hohenheim.de
Telephone No.: +49 711 459-22690

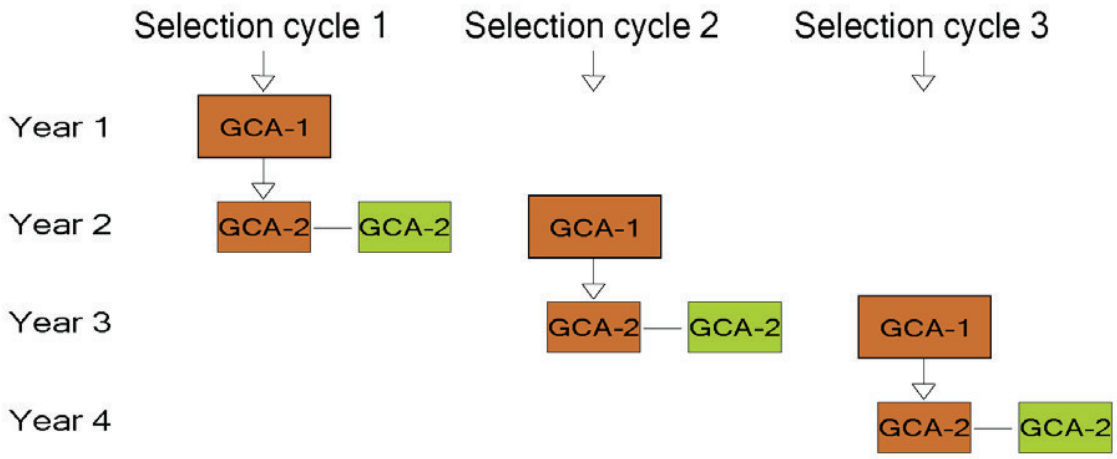
Available ORCID of the authors:

Angela-Maria Bernal-Vasquez: https://orcid.org/0000-0003-0415-831

Patrick Thorwarth: https://orcid.org/0000-0003-4456-2358

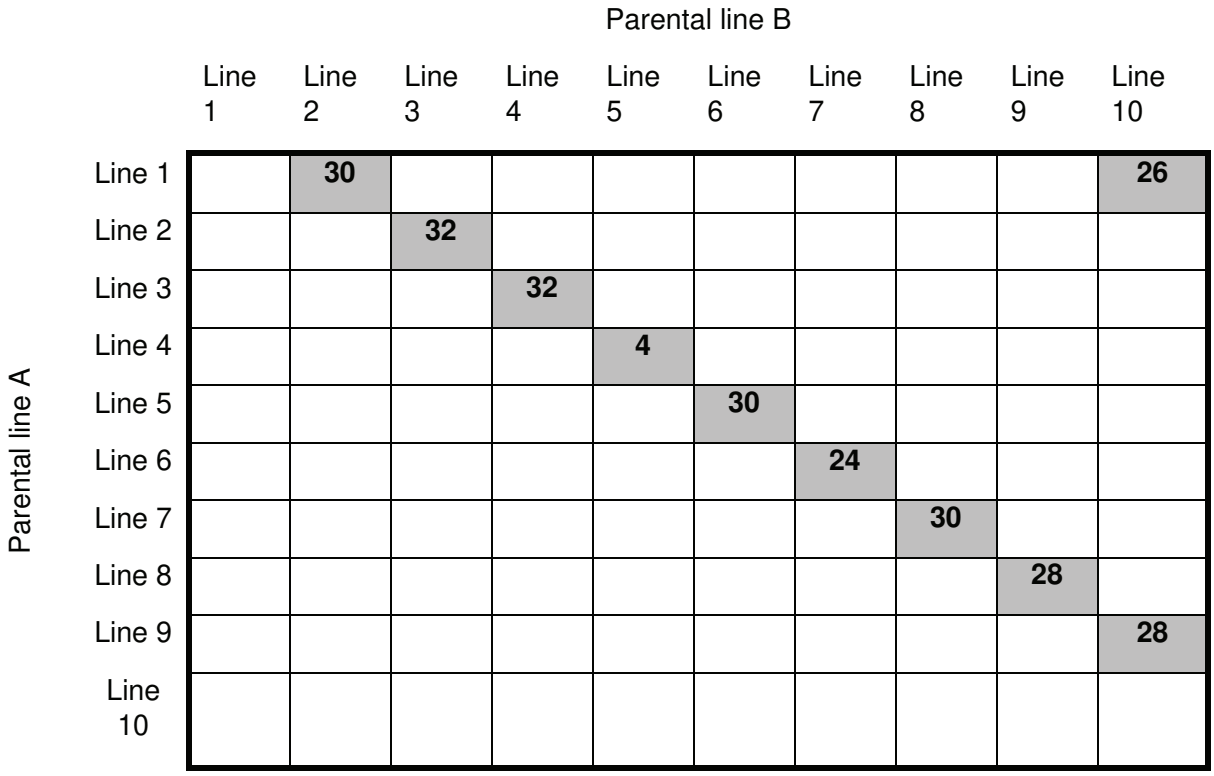
Thomas Miedaner: https://orcid.org/0000-0002-9541-3726

Online resource 2



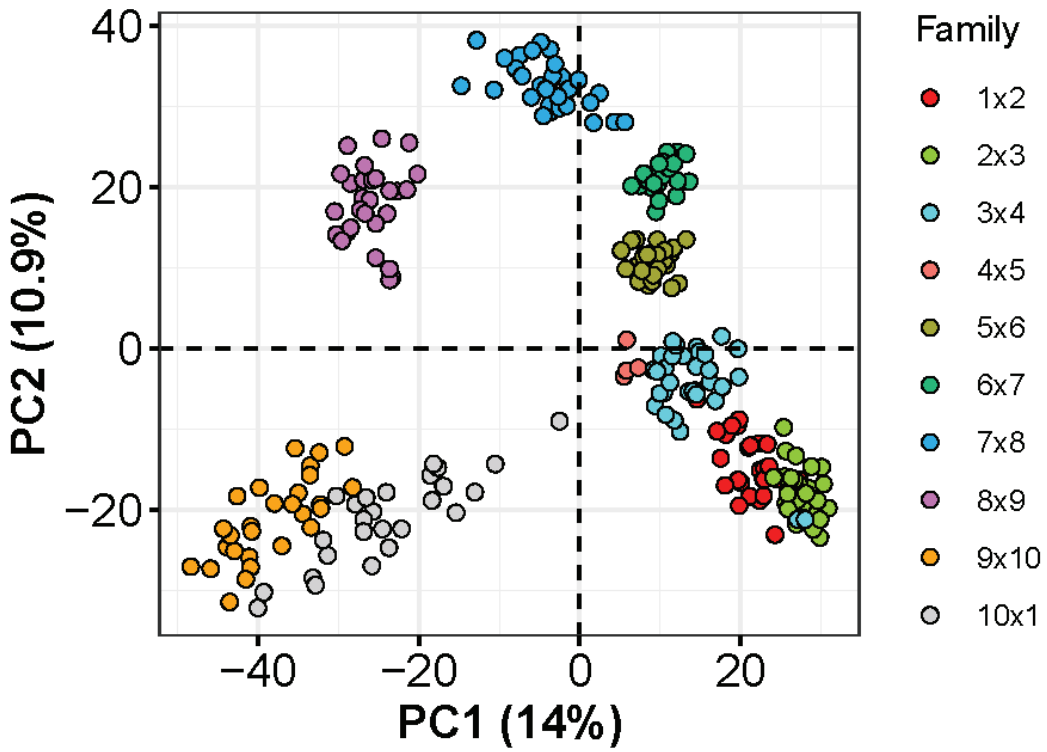
Supplementary Fig. S1 Schematic representation of the selection cycles in hybrid rye breeding program (adapted from Bernal-Vasquez et al. 2017). For across-cycles prediction, phenotypic, molecular, and hyperspectral data for model training could be collected in the second general combining ability trials (GCA-2), whereas prediction and validation are performed among the first GCA trial (GCA-1) and GCA-2 of a subsequent selection cycle, respectively. Brown boxes stand for grain yield trials, while green boxes for biomass trials.

Reference: Bernal-Vasquez A-M, Gordillo A, Schmidt M, Piepho H-P (2017) Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC Genet 18(1):51. doi: 10.1186/s12863-017-0512-8

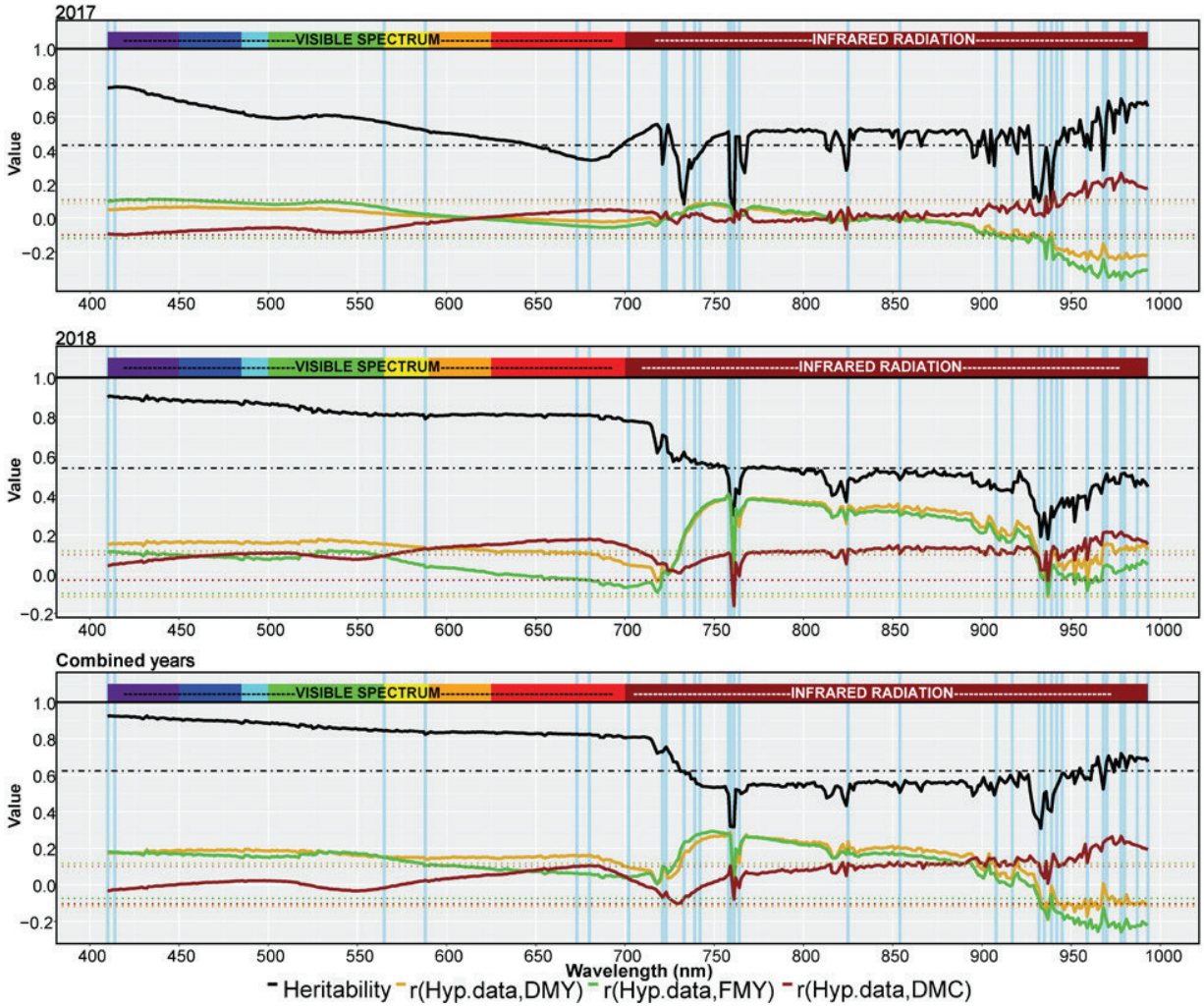


Supplementary Fig. S2 Schematic representation of the single-round robin design used in the present study. The F1 plants (*n*=264) were derived from each of the chain crosses (shaded cells showing the size of each bi-parental family). Adapted from Verhoeven *et al.* (2006)

Reference: Verhoeven KJF, Jannink JL, McIntyre LM (2006) Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity 96(2):139–149



Supplementary Fig. S3 Principal component analysis (PCA) of the ten bi-parental families based on SNP data showing the outcome for the first two principal components.



Supplementary Fig. S4 Heritability estimates (black line) for the hyperspectral bands, phenotypic correlations (r) between hyperspectral bands and dry matter yield (yellow line), fresh matter yield (green line), and dry matter content (brown line), and selected hyperspectral bands after the least absolute shrinkage and selection operator (Lasso, light blue vertical lines) for 274 winter rye hybrids assessed in two years, which were individually and combined analyzed. The mean heritability across selected wavelengths is denoted by the dot-dashed black line. Correlation estimates (in absolute values) ≥ to the respective dotted lines are significant ($p<0.05$).

6. General Discussion

Typically, hybrid rye breeding is driven by grain yield (GY), whereas biomass is destructively assessed in strongly reduced populations at later selection stages due to the prohibitive capital and time expenses that would arise from screening large field trials. Accurate and affordable indirect selection of biomass at early stages of the breeding program emerges, therefore, as crucial for leveraging the potential of rye as an alternative dual-purpose crop to meet the increasing bioenergy demands in the context of climate change mitigation.

The three main outcomes of this thesis are related to high-throughput phenotyping as a suitable tool for biomass breeding in hybrid rye, factors influencing the prediction ability of models based on reflectance and genomic data, and the incorporation of different data sources into predictive breeding.

6.1. Hyperspectral imaging for indirect biomass estimation

6.1.1. Data acquisition and management

For practical breeding programs to benefit from the breakthroughs in imaging and sensor technologies, challenges particularly related to data acquisition, management, and modeling must be addressed (Tardieu et al. 2017, Araus et al. 2018). Considering the repeatable nature of the phenomic approach for assessing plant phenotypes throughout the crop life cycle, the choice of the appropriate time point(s) to collect the reflectance data is crucial (Araus and Cairns 2014). Reflectance data measured at the heading and grain filling growth stages are preferred over earlier stages to maximize its correlation with GY (Monteiro et al. 2012, Rischbeck et al. 2016, Aguate et al. 2017, Krause et al. 2019, Prey et al. 2020) and DMY (Babar et al. 2006, Prasad et al. 2009, Prey et al. 2020). Thus, for practical purposes, yield prediction can be satisfactorily done with measurements after heading (Prasad et al. 2007b). In the present study, reflectance data was collected only after heading (i.e., during the grain filling stage) with small differences in terms of prediction abilities among both flight dates and no clear preference for one time-point across all analyzed models (Table 1). From a

breeding point of view, however, measurements conducted closely to harvest would be preferred assuming that they can carry information that more accurately reflects the yield potential of the plot (Rischbeck et al. 2016). For instance, the delayed foliar senescence, a trait known as “stay-green” linked to enhanced biomass yield (Thomas and Ougham 2014), can be more precisely scored at advanced phenological stages, when the loss of plant photosynthetic pigments begins (Babar et al. 2006). Therefore, gathering reflectance data throughout the crop life cycle may allow a better characterization of the tested genotypes improving the stability and accuracy of predictions (Aguate et al. 2017). However, combining multiple time-points data yielded only slight improvements in the predictive power of some of the evaluated equations in the present study (Table 1) as well as in reflectance-based models in maize (Aguate et al. 2017) and wheat (Prasad et al. 2007b, Montesinos-López et al. 2017b). These results suggest that, a single evaluation at late grain filling carries almost the same information as multiple flights during this growth stage (Publication I, Publication II, Table 1), potentially reducing the capital, time, and computational resources demanded for phenotyping large breeding populations in a non-destructive manner.

Table 1 Prediction ability for DMY assessed by hyperspectral information collected on two flights conducted after flowering by an UAV as well as by plant height and genomic data for 274 rye hybrids across 4 locations and 2 years.

Parameter	Flight 1	Flight 2	Both flights
---------- Incorporating only hyperspectral data ----------			
All vegetation indices (VI_{all}, N=13)	0.35	0.42	0.42
Hyperspectral bands (HBLUP, N=32 selected wavelengths)	0.54	0.52	0.59
----------Incorporating hyperspectral data and plant height ----------			
VI_{all} + Plant height (PH)	0.55	0.62	0.58
HBLUP + PH (Bivariate_H)	0.55	0.50	0.62
---------- Incorporating hyperspectral data, plant height, and genomic data ----------			
HBLUP + PH + genomic data (Bivariate_G_H)	0.72	0.75	0.75

In contrast to multispectral devices, hyperspectral sensors allow a more exhaustive characterization of phenotypes by collecting numerous narrow wavelengths, however, a high degree of redundancy and intercorrelation among them is typically observed, representing a major issue for hyperspectral data analysis (Thorp et al. 2017, Publication II). Thus, multicollinearity may affect prediction equations based on hyperspectral data

(Dunagan et al. 2007). In statistical modeling, especially in regression analysis, multicollinearity (also called collinearity) refers to the linear relation (i.e., non-independence) between two or more predictors, a situation that can potentially lead to wrong parameter estimation with severe consequences in the performance of models, including incompetence to deal with slight changes in data (instability), inadequate assessment of the relative importance of each variable, imprecise significance testing of predictors, incorrect selection of the most relevant variables, and erroneous extrapolation of results beyond the considered data due to different collinearity patterns (Dormann et al. 2013).

Several strategies have been reported throughout the literature to deal with multicollinearity in hyperspectral datasets. The conventional approach is represented by reducing the reflectance data to VIs (Tattaris et al. 2016). VIs are usually based on individual wavelengths within the red and near-infrared spectral regions, for instance, the widely used Normalized Difference Vegetation Index (NDVI; Rouse Jr et al. 1974, Tucker 1979), however, alternative configurations such as the modified Simple Ratio (mSR; Sims and Gamon 2002) of conventional VIs have been proposed for exploring also the red-edge region (Xie et al. 2018). The rationale behind the use of these specific spectral channels is the unique reflectance fingerprint of vegetation, which is determined by external and internal leaf properties and photosynthetic pigments, therefore, reflecting the plant physiological status (Peñuelas and Filella 1998). The reflectance pattern of healthy plants (Pauli et al. 2016) is characterized by strong radiation absorbance in the visible spectrum (400 – 700 nm; VS) and high reflectance in the infrared radiation (700 to ~1000 nm in our study; IR), with the highest slope in reflectance located in the red-edge region (680 - 750 nm; Filella and Peñuelas 1994) as shown in Fig. 3. Reflectance changes are strong indications of physiological stress, such as lower reflectance in the IR is linked to drought (Filella and Peñuelas 1994, Peñuelas and Filella 1998) as observable for the data collected in 2018 for this study (Fig. 3).

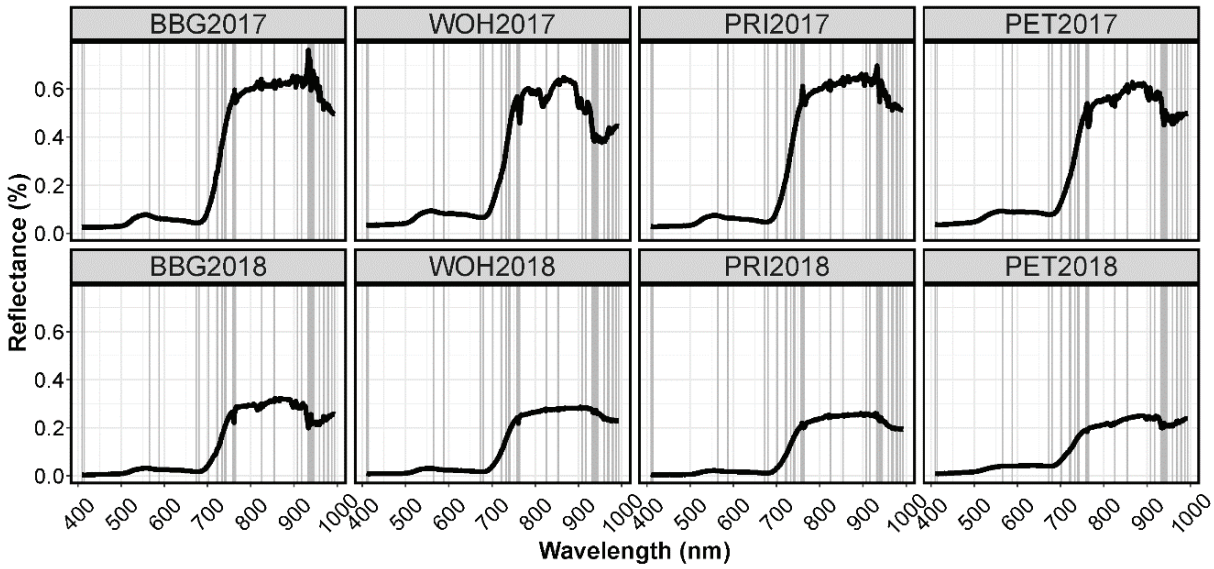


Fig. 3 Canopy spectral reflectance (based on adjusted entry means) collected on two flight dates averaged across 274 winter rye hybrids at each environment. The 32 selected hyperspectral bands after the least absolute shrinkage and selection operator (Lasso) are shown by gray vertical lines. For more information about the adjustment of the hyperspectral data as well as the variable selection procedure applied, please see Publication II.

The hyperspectral data collected for this study consisted of 400 nearly continuous narrow wavelengths ranged from 400 to 993 nm. Of this broad set of spectral data, only a reduced number was used to derive VIs (Publication I), remaining large spectral regions without being exploited. To incorporate full-spectrum data, while effectively reducing data multicollinearity, regularization and variable selection procedures are powerful statistical procedures (Liu and Li 2017). The elastic net (EN; Zou and Hastie 2005), the least absolute shrinkage and selection operator (Lasso; Tibshirani 1996), and ridge regression (RR, Hoerl and Kennard 1970) shrink together the coefficients of correlated variables towards zero, but unlike RR, EN and Lasso set some coefficients to exactly zero, hence, performing also variable selection (Hastie et al. 2009). Therefore, EN and Lasso can assist in handling highly dimensional reflectance datasets while selecting the most relevant spectral features and are preferred over RR and variable projection techniques like principal component analysis and partial least squares (PLS) due to their superior prediction performance and easier model interpretation (Liu and Li 2017).

While in the present study Lasso and EN yielded similar prediction abilities, the first delivered a more parsimonious model and was, therefore, preferred (Publication II). However, this outcome should be interpreted with caution given the advantages of EN over Lasso listed in the literature. The higher number of predictors selected by EN was expected since, unlike Lasso, it performs grouped selection ("oracle" property), i.e., the entire group of highly correlated variables is incorporated into the model if one of them is selected (Zou and Hastie 2005). Moreover, Lasso cannot pick more predictors *p* than the sample size *n* while EN does not show this saturation problem. When selecting among highly correlated variables, Lasso randomly chooses one and ignores the remaining ones whereas EN shows a more stable regularization term (Zou and Hastie 2005). Thus, for other validation scenarios or data rather than considered in this research, the data analysis might profit from the benefits of EN. Likewise, considering the numerous available feature selection procedures with potential applications on spectroscopic data for biomass estimation (Ali et al. 2015), it would be advisable to evaluate the advantages and drawbacks of each of them and choose the one that better fits the data available and the pursued objectives.

Recently, high-dimensional phenomic data has been also used to estimate relationship matrices as routinely done with markers in GS (Rincent et al. 2018, Krause et al. 2019). In both publications, the entire data available was used, remaining the selection of the most relevant wavelengths unexplored, an issue often observed in studies applying traditional methods such as PLS (Thorp et al. 2017). The outcomes of the present study demonstrated that, before kinship matrix estimation, filtering the vast hyperspectral data is worthwhile to effectively address data dimensionality, improve the precision of reflectance-based models, and streamline the analysis of spectral data (Publication II). Also, by hyperspectral data classification, the usefulness of each spectral region for DMY estimation in rye can be assessed. Among the selected wavelengths by Lasso and EN (Publication II), the majority belong to the IR suggesting that this spectral region is the most informative for DMY. Still, some wavelengths from VS were consistently selected, indicating that this region contains some meaningful information.

With the advent of "big data", machine learning models represent, therefore, a sophisticated data analysis tool leading to superior decision making with multiple applications within the agricultural sciences (Liakos et al. 2018). In line with previous

research (Mutanga et al. 2012, Ali et al. 2016, Han et al. 2019), the outcomes of the present study demonstrate that combining machine learning algorithms with remote sensing data of increasing resolution stands for a promising strategy to extract meaningful information to achieve superior biomass prediction abilities. Considering the advent of deep learning in plant science (Singh et al. 2018), future work could explore the advantages of nonlinear models for improving reflectance data analysis.

6.1.2. On the predictability of models based on vegetation indices and full-spectrum data

In this study, the suitability of agronomic traits and VIs as secondary criteria to effectively predict DMY in rye was evaluated. Let r_A be the genotypic correlation between DMY and the secondary trait Y, H_{DMY} the square-root of the heritability of DMY, H_Y the square-root of the heritability of Y, indirect selection for DMY will be preferred over direct selection only if $r_A H_Y > H_{DMY}$ assuming equal selection intensity for both traits (Falconer and Mackay 1996 p.319). This condition was only fulfilled by plant height (PH) and not by other agronomic traits like GY in the present research (Publication I) as well as in a previous study for rye testcrosses (Haffke et al. 2014). However, selecting indirectly for high DMY by focusing only on PH would require complementary breeding efforts for lodging resistance of the selected tall genotypes (Roux et al. 2010, Haffke et al. 2014).

In the present study, 23 previously published VIs (including NDVI and mSR) were evaluated as secondary traits for indirectly estimating DMY of rye hybrids (Publication I). Genetic correlations between VIs and DMY (≤|0.35|) were by far lower than the observed between DMY and other agronomic traits such as GY (0.64) and PH (0.86). Also, highly variable heritability estimates were shown by all VIs (from almost zero to 0.84). Consequently, prerequisites (i.e., $r_A H_Y > H_{DMY}$) for a successful application of indirect selection of DMY based solely on VIs were not met. This study on rye has, therefore, not confirmed previous research on the suitability of VIs for evaluating major plant parameters (e.g., GY, DMY, LAI, nitrogen use efficiency, biotic and abiotic stress condition) of other crop species under field growing conditions (Aparicio et al. 2000, Broge and Mortensen 2002, Hansen and

Schjoerring 2003, Gutiérrez-Rodríguez et al. 2004, Babar et al. 2006, Huang et al. 2007, Prasad et al. 2007a, Prasad et al. 2007b, Tilling et al. 2007, Prasad et al. 2009, Erdle et al. 2011, Monteiro et al. 2012, Zhang et al. 2012, Thorp et al. 2015, Gizaw et al. 2016, Rischbeck et al. 2016, Barmeier and Schmidhalter 2017, Cheng et al. 2017, Frels et al. 2018, Li et al. 2018, Zhang et al. 2019, Zheng et al. 2019, Prey et al. 2020).

A likely explanation for this lack of agreement lies in substantial differences in the experimental data considered, including the characteristics and number of the plant materials, the plot size, the number of environments (equivalent to year–location combinations), and the agronomic management. In this study, 404 elite winter rye hybrids, including a subset of 274 hybrids, were analyzed for GY and DMY, respectively, planted in small-size plots (5-6 m^2) under optimal agronomic conditions (e.g., adequate fertilization, planting density, weed and disease control, as well as supplementary irrigation when necessary). The parents have already been subjected to rigid selection for enhanced agronomic performance, which reduced the diversity of their offspring. Although significant genotypic variation was observed for all agronomic traits and VIs, the variability of this elite breeding material is expected to be substantially lower than the highly diverse panels tested in the abovementioned publications. Most of these studies included genotypes with contrasting phenotypic characteristics, for instance, historical and modern high-yielding cultivars, lines from different breeding programs, different canopy architectures as well as susceptible and resistant varieties against certain abiotic or biotic stresses, guaranteeing, therefore, a superior genetic variability. These studies were also based on by far fewer genotypes (from one up to 75) tested on larger field plots planted in a reduced number of environments (generally from one to four) than in this research thesis. Only Frels et al. (2018) and Gizaw et al. (2016) evaluated a broader panel of genotypes (299 and 402, respectively) in small-size field plots. However, these two publications share with most of those abovementioned not only that distinct subpopulations were included but also that field trials were conducted under contrasting field management (e.g., different irrigation, fertilization treatments, or a combination of both), further increasing the variability observed for the target trait. Hence, to effectively differentiate the performance among tested individuals by using VIs, a wide range of genotypic (Babar et al. 2006) together with environmental (Aparicio et al. 2002) variability should be considered. The negative effect of high uniformity between modern varieties on the prediction ability of VIs-based models has

been observed in this study as well as for GY prediction in barley (Rischbeck et al. 2016). The difficult transfer of results obtained under conditions not fully compatible with practical plant breeding to real phenotyping scenarios, represents, therefore, a major issue slowing down the utilization of remote sensing as a breeding tool (Araus et al. 2018).

Alternatively, several VIs can be pooled with the ultimate aim of reach superior predictive power than the one achievable by using each VI individually as suggested by the findings of this research (Publication I) as well by a previous study in wheat (Montesinos-López et al. 2017b). However, recent publications have shown that VIs cannot capture all the information contained in the extensive data collected by hyperspectral sensors, since the prediction ability of VIs was surpassed by regression and shrinkage methods such as PLS and Bayesian shrinkage using whole-spectrum information for GY in maize (Aguate et al. 2017) and, together with RR, also in wheat (Montesinos-López et al. 2017b, Hernandez et al. 2015). Therefore, statistical procedures mining information across the entire spectrum maximize the benefits of advanced phenotyping in plant breeding (Thorp et al. 2017, Araus et al. 2018). In line with these results, in the present study, HBLUP, a single-kernel model based on a hyperspectral-derived kinship matrix incorporating reflectance data throughout the spectrum, represented a superior alternative to single or combined VIs for DMY prediction (Publication II, Table 1). Moreover, HBLUP surpassed the prediction ability of GBLUP, which is based on a markers-derived kinship matrix, under reduced TRN (Publication II) or for predicting across distinct populations (Publication III). In complete agreement with these outcomes, studies on wheat and poplar have recently observed that the performance of prediction models based on spectroscopic-derived relationship matrix was closer to the conventional marker-based approach (Rincent et al. 2018, Krause et al. 2019).

6.2. Factors influencing the prediction ability of GBLUP and HBLUP

Several factors affect the performance of GS in plant breeding, including the size of the training data, the relatedness between genotypes used for model training and validation, and the trait heritability (Lin et al. 2014). The prospects of HBLUP as a suitable model for dealing with these challenging issues were evaluated and compared to the prediction performance obtained by GBLUP.

6.2.1. The training set size

The phenotypic information in breeding has mutated from being the basis on which the best genotypes are selected (phenotypic selection) to be used to calibrate genomic models aiming the selection of unphenotyped individuals based on their GEBVs (Heffner et al. 2009). This radical paradigm change in plant breeding, however, does not imply that GS can perform well when trained on few genotypes since the negative impact of a reduced TRN on the performance of GS is widely known (VanRaden et al. 2009). Some studies have shown that hyperspectral-based models also yielded higher accuracies when calibrated on larger TRN sizes. For forest biomass estimation, different models based on hyperspectral data profited from larger sample sizes, however, the importance of the TRN size on the prediction accuracies was surpassed by other factors as the choice of the appropriate model and the spectroscopic data type used (Fassnacht et al. 2014). Similarly, a larger TRN was beneficial for regression models based on VS and IR wavelengths used for predicting soil carbon content (Lucà et al. 2017) and near-infrared data for estimating grain nutrients in barley (Wiegmann et al. 2019). The positive correlation between larger TRN and higher predictive power of GBLUP and HBLUP within the same breeding population was also observed in this research (Publication II). Nevertheless, HBLUP was significantly less affected by reduced TRN size, showing, hence, superior performance than GBLUP on small TRN sizes. Thus, HBLUP emerges as a suitable model if larger phenotypic data cannot be afforded.

6.2.2. The genetic and environmental connectivity between training and validation data

In rye breeding, each year numerous genotypes from different genetic backgrounds are tested across several environments, potentially enabling the integration of this data to enlarge TRN (Bernal-Vasquez et al. 2017). The suitability of this strategy, however, depends on the degree of influence of the genotypic and environmental connectivity between TRN and VAL on the performance of genomic and hyperspectral models.

DMY in rye is strongly influenced by genotypic and environmental conditions (Miedaner et al. 2012, Haffke et al. 2014, Publication I). The same holds for spectral fingerprints collected on rye canopies (Publication I, Publication II, Fig.3). Nevertheless, the ten bi-parental families analyzed in the present study could be much more clearly differentiated based on their allelic composition than on their reflectance fingerprints (Publication III). This dissimilar influence of genetic relationships on both data sources resulted in substantial differences in the performance of GBLUP and HBLUP when predicting genotypes poorly correlated to the ones used for model calibration.

The high dependency of genomic predictions on a sufficient genetic correlation between TRN and VAL was observed in this study (Publication III), confirming previous findings in animal (Roos et al. 2009, Pszczola et al. 2012) as well as maize and wheat breeding (Crossa et al. 2014, Heslot et al. 2014). GBLUP performed satisfactorily when used to predict the DMY of progenies closely related to those used for model training (prediction ability = 0.49 - 0.64). Then, a strong decay of about 60% was observed when closely related individuals were excluded from TRN. Similarly, Wang et al. (2014) reported modest prediction ability for GY when GS was applied between two bi-parental rye families, even though they were connected by a common parent. Thus, a sufficiently high degree of genetic relationship is needed for a successful implementation of GS across selection cycles in rye (Auinger et al. 2016, Bernal-Vasquez et al. 2017). Conversely, this prerequisite does not seem to be crucial for hyperspectral-based models to achieve acceptable predictions since HBLUP performed significantly more stable than GBLUP across different degrees of genetic relatedness (Publication III). For instance, the penalization observed for HBLUP when predicting DMY among weakly connected progenies was, on average, only ca. 13% compared to the prediction abilities achieved for highly related genotypes. Thus, when low related populations are to be predicted, spectral data emerges as a competitive information source given the relatively lower influence of genetic backgrounds on the information imprinted on canopy reflectance.

In contrast, predicting DMY in a new environment was highly challenging for hyperspectral- and genomic-based models, mainly due to significant G×E interactions observed for agronomic and reflectance data (Publication I, Publication II, Publication III). The observed low performance of GBLUP in predicting the genotypic performance in an unknown

environment is in line with previous studies showing that the accuracy of genomic models is enhanced by incorporating more environments into the TRN (Utz et al. 2000) as well as by borrowing information from correlated environments (Burgueño et al. 2012, Crossa et al. 2014).

Similarly, VIs and HBLUP performed the highest when data across the series of environments was incorporated into TRN, while the lowest abilities corresponded to TRN without information of the aimed site. Thus, the incorporation of data collected in different environments increases the variation present in TRN, boosting in proportion the prediction ability of hyperspectral-based models (Wiegmann et al. 2019). Nevertheless, this lack of representation of the targeted environment could only be partially compensated by including a higher number of environmental samples into TRN. Moreover, if the environment to be predicted was highly contrasting to the sampled ones, this strategy was indeed counterproductive. These outcomes confirm that enlarging TRN by including highly diverse environments can lead to heterogeneous band-to-trait signals (Montesinos-López et al. 2017a) since reflectance data is strongly influenced by environmental aspects (Gates et al. 1965, Fig. 3). Therefore, the accuracy of predictions can be affected (Rischbeck et al. 2016). Hence, the applicability of hyperspectral models is highly restricted to the environment under which canopies were assessed and predictions beyond these conditions may not be sufficiently reliable (Hernandez et al. 2015, Weber et al. 2012, Rischbeck et al. 2016, Wiegmann et al. 2019). Thus, adequate environmental sampling is crucial for boosting the performance of genomic equations (Utz et al. 2000) and hyperspectral-based models (Lucà et al. 2017, Wiegmann et al. 2019).

6.2.3. Characteristics of the trait under study

While GBLUP performed better than HBLUP for dry matter content (DMC, H^2=0.70 to 0.81), the opposite was observed for mid-heritability traits like DMY (H^2=0.50 to 0.54, Publication III). Thus, the crucial role of the trait heritability in determining the performance of prediction models was confirmed for GBLUP as observed by Jia and Jannink (2012) but not

for HBLUP. Other factors rather than the heritability of the trait under scrutiny might have, therefore, a greater influence on HBLUP.

Across the spectrum, estimates of the heritability of the bands and of the correlation between single bands and different agronomic traits were highly variable (Publication II, Publication III). The exclusion of broad spectral regions due to their relatively lower heritability resulted in significant reductions in the ability of HBLUP to predict DMY (Publication II). Similar observations were made for reflectance-enabled predictions for GY in wheat (Montesinos-López et al. 2017b). Thus, instead of selecting the bands based on their heritability, the spectral regions most relevant to the trait of interest should be considered, which in turn, are determined by their sensitivity to capture changes in the targeted trait based on canopy reflectance. Different biophysical and biochemical plant properties are linked to specific wavelengths across the electromagnetic spectrum (Pauli et al. 2016). While, for instance, the red-edge is the most informative region for assessing biomass-related traits such as chlorophyll concentration and LAI, it lacks sensitivity to assess the water status of non-stressed plants (Filella and Peñuelas 1994). In contrast, the spectral sensitiveness to water status and DMC starts at 950 nm (Jacquemoud et al. 2000). The wavelengths closely linked to DMY were covered by the data of this study (410 - 993 nm), whereas most of the regions sensitive to DMC were not collected, likely explaining the different predictive performance for each trait.

6.3. Conclusions for rye biomass breeding in the "omics" era

In the present breeding scheme in rye, DMY is assessed for the first time in strongly reduced populations in GCA-2 experiments. Given the positive but reduced phenotypic correlation between DMY and GY (r = 0.33 to 0.35; Haffke et al. 2014, Publication I), it is expected, therefore, that there might be individuals with high DMY potential that are discarded before GCA-2 due to their relatively lower GY. Thus, additional breeding efforts, in terms of more effective secondary traits, improved modeling, and optimized breeding pipelines, are imperative for efficient breeding for high-biomass rye hybrids in the context of increasing bioenergy demand in the EU (European Commission 2011a, 2011b). Major advancements in the field of genomics and phenomics open new possibilities to increase selection gain in rye

biomass comparing to classical GY-driven breeding programs. The outcomes of the present study underline that integrating hyperspectral, genomic, and phenotypic data resulted consistently in higher prediction abilities for DMY than when each was individually considered.

Previous studies have shown that PH is a key estimator for DMY prediction in cereal crops (Fernandez et al. 2009, Haffke et al. 2014) and VIs represent a complementary source of information to increase its predictive power to a certain degree (Bendig et al. 2015, Tilly et al. 2015, Yue et al. 2017, Zhang et al. 2017, Li et al. 2018). These findings were corroborated in the study at hand (Publication I) since combined VIs showed lower prediction abilities (0.42) than PH as an individual predictor (0.57) and adding VIs to PH represented only a minor improvement (0.58, Table 1). Interestingly, predictions were improved by a multiple linear regression model integrating PH, GY, and a subset of VIs (0.75). In the context of biomass breeding, the incorporation of the routinely assessed GY into regression equations represents a further opportunity to indirectly assess DMY more effectively as also observed for early prediction of DMY in triticale (Gowda et al. 2011). Recently, instead of incorporating agronomic traits as PH, some authors have incorporated traits derived from remote sensing data (e.g. canopy temperature and VIs) into univariate and multivariate genomic models to achieve superior prediction abilities for GY in wheat breeding (Rutkoski et al. 2016, Sun et al. 2017, Crain et al. 2018, Juliana et al. 2019, Sun et al. 2019). These results are consistent with the objectives of indirect selection since the prediction ability for traits being expensive to assess and showing relatively low heritability can be enhanced by incorporating correlated traits easier accessible and displaying higher heritabilities as demonstrated for genomic models in simulation and empirical studies in plant and animal sciences (Calus and Veerkamp 2011, Jia and Jannink 2012, Guo et al. 2014, Okeke et al. 2017, Schulthess et al. 2018, Velazco et al. 2019). Likewise, Fernandes et al. (2018) observed that the predictive power of single-trait GS for biomass in sorghum was substantially surpassed by a multi-trait GS approach using PH as a secondary trait.

Instead of extending GS models by adding VIs, Krause et al. (2019) combined two relationship matrices, one based on markers and the other on whole-spectrum data (380-850 nm), into multi-kernel GBLUP for leveraging the ability of GS for GY in wheat. In total agreement with this publication, a multi-kernel model (model “G+H”) based also on

reflectance- and marker-derived kinship matrices outperformed the predictive power of the corresponding single-kernel models within the same breeding population in the present study (Publication II). The extension of G+H to a bivariate model ("Bivariate_G+H") by incorporating PH was beneficial in terms of the superior prediction ability of this model (Publication II). The equations here presented are flexible and allow the incorporation of GY and other correlated traits to achieve superior DMY predictions. The reader must be, however, aware that by including several predictors, the complexity of the models and the computational burden substantially increases.

PH, a key biomass component, allows higher selection gains for DMY and represents, therefore, a change in the architecture of crops, which has been strongly oriented to high-yielding dwarf varieties that maximize the harvest index (Fernandez et al. 2009). Nevertheless, the use of PH to predict DMY should be carefully analyzed. Models including PH tended to select taller genotypes, which were not always the highest yielding ones (Publication II). Thus, in the long term, small PH differences of the selected genotypes will accumulate, and additional resources would be needed for breeding against lodging. It is also worth mentioning that the correlations observed in this study for DMY and different agronomic traits (e.g., GY and PH) might vary according to the germplasm under analysis.

In within-cycle predictions, where the genotypes used to calibrate and validate the genomic model are derived in the same selection cycle, higher genotypic connectivity is expected between TRN and VAL than when genotypes correspond to different selection cycles. In the analyzed breeding scheme, this scenario could be, for instance, applied starting at the GCA-1 stage aiming for superior selection decisions for biomass than the traditional GY-based approach. Combining different sources of information resulted here in the best approach for indirectly selecting for superior DMY. However, all analyzed models profited from larger training populations in terms of both higher and more stable predictions. Consequently, while in this scenario, prediction models profit from closer genetic connectivity, TRN size emerges as a crucial factor limiting their performance (Publication II). Thus, to successfully predict unphenotyped individuals, duplicating a substantial proportion of large-sized GCA-1 trials would be needed to calibrate the predictive equations accurately (Publication II). If only a reduced TRN is affordable, HBLUP emerged as a promising approach given its higher prediction power on reduced calibration populations compared to GBLUP (Publication II).

Conversely, large TRNs can be obtained by integrating different selection cycles to predict DMY of genotypes from a subsequent cycle. Here, a reduced genetic similarity between TRN and VAL is expected. The outcomes of this study showed that hyperspectral imaging is an important breeding tool given its relative independence from genetic backgrounds, in opposition to the high impact of this factor on the predictive power of GBLUP (Publication III). In this scenario, the addition of molecular data to HBLUP showed only limited advantages while the superiority of the bivariate model was again confirmed. Thus, the high complementarity observed among molecular, hyperspectral, and phenotypic data allows their integration into multivariate analysis for successfully predict DMY of genotypes not only derived from the same crossings (Publication I, Publication II) but also across distinct populations (Publication III). To successfully predict DMY within and across populations, however, special attention should be given to maintain sufficient environmental connectivity between the data used for model training and validation (Publication I, Publication II, Publication III).

In the breeding scheme at hand, the indirect DMY estimation of the candidates being only tested for GY in GCA-1 trials could be performed by fitting prediction models with data collected on GCA-2 trials from previous selection cycles (Fig. 4). Here, the prediction error can be estimated by forward validation (FV), an alternative procedure to CV for the prediction of breeding values of genotypes tested in later years based on data collected in previous years (Bernal-Vasquez et al. 2017). So, the superfluous duplication of GCA-1 plots is entirely avoided, saving capital and time, and therefore, boosting the benefits of predictive breeding. Moreover, the data collected in the cycles routinely started each year allow the updating of the prediction model, increasing its reliability to estimate the trait of interest (Auinger et al. 2016). Thus, across-cycles prediction based on models combining meaningful hyperspectral, genomic, and agronomic data emerges as a very promising strategy to establish an affordable dual-purpose rye breeding program.

In conclusion, hyperspectral imaging is a valuable breeding tool capable to record phenotypic information that otherwise would be very difficult or expensive to assess by destructive means, as biomass. However, the widespread adoption of this technology in practical plant breeding is still missing. While challenges regarding data acquisition and management as well as the difficulty of extrapolating results obtained in controlled

conditions to real breeding scenarios were confirmed by the outputs of this research, suitable approaches to optimize data collection, reduce data multicollinearity, and improve modeling for incorporating hyperspectral information into predictive equations were presented. The synergetic effect observed among hyperspectral, genomic, and agronomic data can be exploited to achieve better predictions of the target trait. Given the increasing availability of other omics datasets such as transcriptomics, proteomics, and metabolomics future research should focus on assessing the prospects also of these novel data sources as relevant complements to genomics to achieve higher predictive abilities.

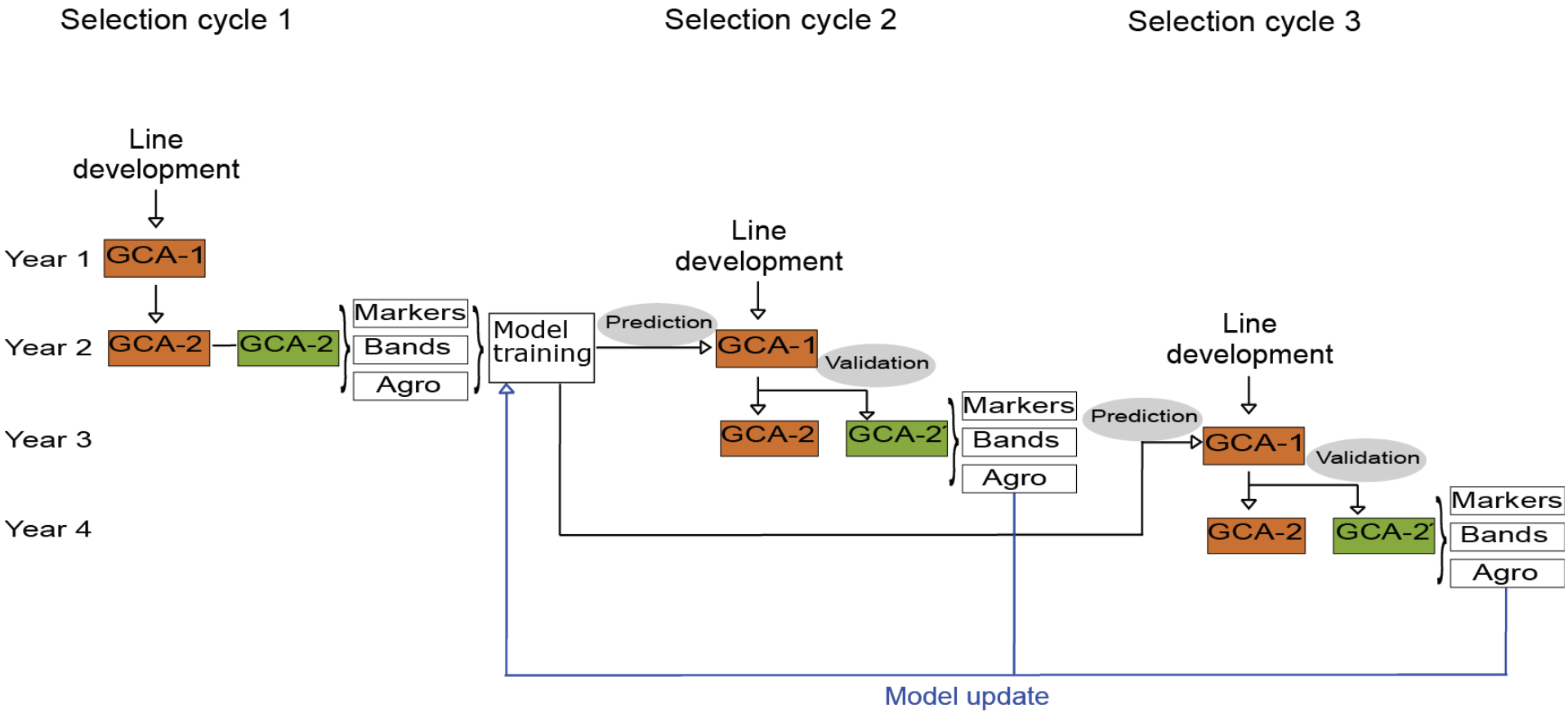


Fig. 4 Selection cycles in a hybrid rye breeding program. Across-cycles prediction could be done by collecting molecular (Markers), hyperspectral (Bands), and agronomic (Agro) data for model training in the second general combining ability trials (GCA-2), whereas prediction and validation are performed among the first GCA trial (GCA-1) and GCA-2 of a subsequent selection cycle, respectively. Brown boxes stand for grain yield trials and green boxes for biomass trials.

7. Summary

Currently, the combination of a growing bioenergy demand and the need to diversify the dominant cultivation of energy maize opens a highly attractive scenario for alternative biomass crops. Rye (*Secale cereale* L.) stands out among other small-grain cereals for its vigorous growth, increased tolerance to abiotic and biotic stressors, as well as high adaptation to acid or sandy soils, being primarily cultivated in Central and Northeastern Europe. In Germany, less than a quarter of the total harvest is used for food production. Consequently, rye arises as a source of renewables with a reduced bioenergy-food tradeoff, emerging biomass as a new breeding objective.

However, rye breeding is mainly driven by grain yield while biomass is destructively evaluated in later selection stages by expensive and time-consuming methods. The overall motivation of this research was to investigate the prospects of combining hyperspectral, genomic, and agronomic data for unlocking the potential of hybrid rye as a dual-purpose crop to meet the increasing demand for renewable sources of energy affordably. A specific aim was to predict the biomass yield as precisely as possible at an early selection stage. For this, a panel of 404 elite rye lines was genotyped and evaluated as testcrosses for grain yield and a subset of 274 genotypes additionally for biomass. Field trials were conducted at four locations in Germany in two years (eight environments). Hyperspectral fingerprints consisted of 400 discrete narrow bands (from 410 to 993 nm) and were collected in two points of time after heading for all hybrids in each site by an uncrewed aerial vehicle.

In a first study, population parameters were estimated for different agronomic traits and a total of 23 vegetation indices. Dry matter yield showed significant genetic variation and was stronger correlated with plant height (r_g=0.86) than with grain yield (r_g=0.64) and individual vegetation indices (r_g=≤|0.35|). A multiple linear regression model based on plant height, grain yield, and a subset of vegetation indices surpassed the prediction ability for dry matter yield of models based only on agronomic traits by about 6 %.

In a second study, instead of vegetation indices, whole-spectrum data was used to indirectly estimate dry matter yield. For this, single-kernel models based on hyperspectral reflectance-derived (HBLUP) and genomic (GBLUP) relationship matrices, a multi-kernel model

combining both matrices, and a bivariate model fitted also with plant height as a secondary trait, were considered. HBLUP yielded superior predictive power than the models based on vegetation indices previously tested. The phenotypic correlations between individual wavelengths and dry matter yield were generally significant ($p < 0.05$) but low ($r_p \leq |0.29|$). Across environments and training set sizes, the bivariate model yielded the highest prediction abilities (0.56 – 0.75). All models profited from larger training populations. However, if larger training sets cannot be afforded, HBLUP emerged as a promising approach given its higher prediction power on reduced calibration populations compared to the well-established GBLUP.

Regarding reflectance data acquisition and management, combining multiple points in time had limited advantages in the predictive power of reflectance-based models. Thus, for practical purposes, hyperspectral imaging can be satisfactorily performed by a single flight after heading. Before its incorporation into prediction models, filtering the hyperspectral data available by the least absolute shrinkage and selection operator (Lasso) was worthwhile to deal with data dimensionally.

In a third study, the effects of trait heritability, as well as genetic and environmental relatedness on the prediction ability of GBLUP and HBLUP for biomass-related traits were compared. While the prediction ability of GBLUP (0.14 - 0.28) was largely affected by genetic relatedness and trait heritability, HBLUP was significantly more accurate (0.41 - 0.61) across weakly connected datasets, particularly for mid-heritable traits as fresh and dry matter yields. In this context, dry matter yield could be better predicted (up to 20 %) by a bivariate model. Nevertheless, due to environmental variances, genomic and reflectance-enabled predictions were strongly dependant on a sufficient environmental relationship between data used for model training and validation.

In summary, to affordably breed rye as a double-purpose crop to meet the increasing bioenergy demands, the early prediction of biomass across selection cycles is crucial. Hyperspectral imaging has proven to be a suitable tool to select high-yielding biomass genotypes across weakly linked populations. Due to the synergetic effect of combining hyperspectral, genomic, and agronomic traits, higher prediction abilities can be obtained by integrating these data sources into bivariate models.

8. Zusammenfassung

Die Kombination eines wachsenden Bioenergiebedarfs und die Notwendigkeit, den vorherrschenden Anbau von Energiemais zu diversifizieren, eröffnen ein äußerst attraktives Szenario für alternative Biomassekulturen. Roggen (*Secale cereale* L.) zeichnet sich, verglichen mit anderen kleinkörnigen Getreidearten, durch ein kräftiges vegetatives Wachstum, eine erhöhte Toleranz gegenüber abiotischen und biotischen Stressfaktoren, sowie eine hohe Anpassung an saure oder sandige Böden aus. Roggen wird hauptsächlich in Mittel- und Nordosteuropa angebaut. In Deutschland wird weniger als ein Viertel der gesamten Ernte für die Lebensmittelproduktion verwendet. Daher gewinnt Roggen durch einen geringeren Zielkonflikt zwischen Bioenergie- und Lebensmittelnutzung an Bedeutung als Quelle für erneuerbare Energien, wobei Biomasse als neues Züchtungsziel auftaucht.

Die Roggenzüchtung konzentriert sich derzeit jedoch hauptsächlich auf den Kornertrag, während die Biomasse in späteren Selektionsstadien durch teure und zeitaufwändige Methoden destruktiv erfasst wird. Die übergeordnete Motivation dieser Arbeit war es, die Aussichten der Kombination von hyperspektralen, genomischen und agronomischen Daten für die Erschließung des Potenzials von Hybridroggen als Zweinutzungspflanze zu untersuchen, um den steigenden Bedarf an erneuerbaren Energiequellen kostengünstig zu decken. Das spezifische Ziel war es, den Biomasseertrag in einem frühen Selektionsstadium so genau wie möglich vorherzusagen. Dazu wurde ein Panel von 404 Elitelinien genotypisiert und als Testkreuzungen für Kornertrag - eine Teilmenge von 274 Genotypen auch für Biomasse-Ertrag – ausgewertet. Feldversuche wurden an vier Standorten in zwei Jahren in Deutschland (entspricht acht Umwelten) durchgeführt. Die hyperspektralen Daten bestanden aus 400 diskreten Banden von 410 bis 993 nm und wurden zu zwei Zeitpunkten nach dem Ährenschieben für alle Testkreuzungen an jedem Ort von einer Drohne gesammelt.

In einer ersten Studie wurden Populationsparameter für verschiedene agronomische Merkmale und insgesamt 23 Vegetationsindizes geschätzt. Der Trockenmasseertrag zeigte eine signifikante genetische Variation und korrelierte stärker mit der Wuchshöhe (r_g=0.86) als mit dem Kornertrag (r_g=0.64) und den einzelnen Vegetationsindizes (r_g=≤|0.35|). Ein multiples lineares Regressionsmodell, welches auf Wuchshöhe, Kornertrag und den besten

Vegetationsindizes basierte, übertraf die Vorhersagefähigkeit für den Trockenmasseertrag von Modellen, die nur auf agronomischen Merkmalen basierten, um etwa 6 %.

In einer zweiten Studie wurde anstelle von Vegetationsindizes das ganze Wellenlängenspektrum verwendet, um den Trockenmasseertrag indirekt abzuschätzen. Hierzu wurden Einzelkernmodelle (*single-kernel models*) basierend auf genomischen (GBLUP) oder hyperspektralen (HBLUP) Beziehungsmatrizen, ein Mehrkernmodell (*multi-kernel model*), das beide Matrizen kombiniert, sowie ein bivariates Modell, welches auch Wuchshöhe als ein sekundäres Merkmal enthielt, analysiert. HBLUP lieferte eine bessere Vorhersagekraft als die Modelle, die auf Vegetationsindizes basierten. Die phänotypische Korrelationen zwischen einzelnen Wellenlängen und dem Trockenmasseertrag waren im Allgemeinen signifikant ($p < 0,05$), jedoch geringfügig ($r_p \leq |0.29|$). Über alle Umwelten und Trainingssatzgrößen hinweg ergab das bivariate Modell die höchsten Vorhersagefähigkeiten (0,56 – 0,75). Alle Modelle profitierten von größeren Trainingspopulationen. Wenn jedoch keine größeren Trainingssätze bereitgestellt werden können, zeigte HBLUP eine höhere Vorhersagefähigkeit als das etablierte GBLUP.

In Bezug auf die hyperspektrale Datenerfassung hatte die Kombination mehrerer Zeitpunkte nur begrenzte Vorteile. Aus praktischen Gründen kann sie daher durch einen einzelnen Flug nach dem Ährenschieben ausreichend gut erfasst werden. Vor der Einbeziehung in Vorhersagemodelle hat sich das Filtern der verfügbaren Hyperspektraldaten durch den *least absolute shrinkage and selection operator* (Lasso) als notwendig erwiesen, um die Dimensionalität der Daten zu verringern.

In einer dritten Studie wurden die Auswirkungen der Heritabilität sowie der Ähnlichkeit innerhalb von Genotypen und Umwelten auf die Vorhersagefähigkeit von GBLUP und HBLUP für biomassebezogene Merkmale verglichen. Während die Vorhersagefähigkeit von GBLUP (0,14 - 0,28) weitgehend durch genetische Verwandtschaft und die Merkmalsheritabilitäten beeinflusst wurde, war HBLUP in wenig verwandten Datensätzen signifikant genauer (0,41 - 0,61), insbesondere für Merkmale mit mittlerer Heritabilität wie Frisch- und Trockenmasseertrag. In diesem Zusammenhang konnte der Trockenmasseertrag durch ein bivariates Modell bis zu 20 % besser vorhergesagt werden. Aufgrund hoher Genotyp-Umwelt-Interaktionen waren genomische und reflexionsbasierte Vorhersagen nur schlecht geeignet, um die Leistung fehlender Umwelten vorherzusagen.

Zusammenfassend ist es für eine kostengünstige Züchtung von Roggen als Zweinutzungspflanze zur Deckung des steigenden Bioenergiebedarfs entscheidend, die Biomasse über Selektionszyklen hinweg frühzeitig vorherzusagen. Die hyperspektrale Bildgebung hat sich als geeignetes Instrument zur Auswahl ertragreicher Biomasse-Genotypen auch in wenig verwandten Populationen erwiesen. Dank des synergetischen Effekts der Kombination von hyperspektralen, genomischen und agronomischen Merkmalen können durch die Integration dieser Datenquellen in bivariaten Modelle höhere Vorhersagefähigkeiten erzielt werden.

9. References

Adão T, Hruška J, Pádua L, Bessa J, Peres E, Morais R, Sousa J (2017) Hyperspectral imaging: a review on UAV-based sensors, data processing and applications for agriculture and forestry. Remote Sens 9(11):1110. doi: 10.3390/rs9111110

Aguate FM, Trachsel S, Pérez LG, Burgueño J, Crossa J, Balzarini M, Gouache D, Bogard M, Campos Gdl (2017) Use of hyperspectral image data outperforms vegetation indices in prediction of maize yield. Crop Sci 57(5):2517–2524. doi: 10.2135/cropsci2017.01.0007

Ali I, Cawkwell F, Dwyer E, Green S (2016) Modeling managed grassland biomass estimation by using multitemporal remote sensing data—a machine learning approach. IEEE J Sel Top Appl Earth Obs Remote Sens 10(7):3254–3264. doi: 10.1109/JSTARS.2016.2561618

Ali I, Greifeneder F, Stamenkovic J, Neumann M, Notarnicola C (2015) Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens 7(12):16398–16421. doi: 10.3390/rs71215841

Aparicio N, Villegas D, Araus JL, Casadesus J, Royo C (2002) Relationship between growth traits and spectral vegetation indices in durum wheat. Crop Sci 42(5):1547–1555. doi: 10.2135/cropsci2002.1547

Aparicio N, Villegas D, Casadesus J, Araus JL, Royo C (2000) Spectral vegetation indices as nondestructive tools for determining durum wheat yield. Agron. J. 92(1):83–91. doi: 10.2134/agronj2000.92183x

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19(1):52–61. doi: 10.1016/j.tplants.2013.09.008

Araus JL, Kefauver SC, Zaman-Allah M, Olsen MS, Cairns JE (2018) Translating high-throughput phenotyping into genetic gain. Trends Plant Sci 23(5):451–466. doi: 10.1016/j.tplants.2018.02.001

Atzberger C (2013) Advances in remote sensing of agriculture: context description, existing operational monitoring systems and major information needs. Remote Sens 5(2):949–981. doi: 10.3390/rs5020949

Auinger H-J, Schönleben M, Lehermeier C, Schmidt M, Korzun V, Geiger HH, Piepho H-P, Gordillo A, Wilde P, Bauer E (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). Theor Appl Genet 129(11):2043–2053. doi: 10.1007/s00122-016-2756-5

Babar MA, Reynolds MP, van Ginkel M, Klatt AR, Raun WR, Stone ML (2006) Spectral reflectance to estimate genetic variation for in-season biomass, leaf chlorophyll, and canopy temperature in wheat. Crop Sci 46(3):1046–1057. doi: 10.2135/cropsci2005.0211

Barmeier G, Schmidhalter U (2017) High-throughput field phenotyping of leaves, leaf sheaths, culms and ears of spring barley cultivars at anthesis and dough ripeness. Front in Plant Sci 8:1920. doi: 10.3389/fpls.2017.01920

Beavis WD (1998) QTL analyses: power, precision, and accuracy. In: Paterson AH (ed) Molecular Dissection of Complex Traits. CRC Press LLC, Boca Raton, FL, pp 145–162

Beckmann JS, Soller M (1986) Restriction fragment length polymorphisms and genetic improvement of agricultural species. Euphytica 35(1):111–124

Ben-Ari G, Lavi U (2013) Marker-assisted selection in plant breeding. In: Altman A, Hasegawa PM (eds) Plant Biotechnology and Agriculture: Prospects for the 21st Century, 1. ed. Academic Press, Amsterdam, Boston, Massachusetts, pp 163–184

Bendig J, Bolten A, Bennertz S, Broscheit J, Eichfuss S, Bareth G (2014) Estimating biomass of barley using crop surface models (CSMs) derived from UAV-based RGB imaging. Remote Sens 6(11):10395–10412. doi: 10.3390/rs61110395

Bendig J, Yu K, Aasen H, Bolten A, Bennertz S, Broscheit J, Gnyp ML, Bareth G (2015) Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. Int J Appl Earth Obs 39:79–87. doi: 10.1016/j.jag.2015.02.012

Bernal-Vasquez A-M, Gordillo A, Schmidt M, Piepho H-P (2017) Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC Genet. 18(1):51. doi: 10.1186/s12863-017-0512-8

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48(5):1649–1664. doi: 10.2135/cropsci2008.03.0131

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47(3):1082–1090. doi: 10.2135/cropsci2006.11.0690

Broge NH, Mortensen JV (2002) Deriving green crop area index and canopy chlorophyll density of winter wheat from spectral reflectance data. Remote Sens Environ 81(1):45–57. doi: 10.1016/S0034-4257(01)00332-7

Bundesanstalt für Landwirtschaft und Ernährung (2019) Bericht zur Markt- und Versorgungslage: Getreide 2019. https://www.ble.de/SharedDocs/Downloads/DE/BZL/Daten-

Berichte/Getreide_Getreideerzeugnisse/2019BerichtGetreide.pdf?__blob=publicationFile&v=3

Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz (2019) Besondere Ernte- und Qualitätsermittlung (BEE). Reihe: Daten-Analysen. https://www.bmel-statistik.de/fileadmin/daten/EQB-1002000-2019.pdf

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52(2):707–719. doi: 10.2135/cropsci2011.06.0299

Busemeyer L, Ruckelshausen A, Möller K, Melchinger AE, Alheit KV, Maurer HP, Hahn V, Weissmann EA, Reif JC, Würschum T (2013) Precision phenotyping of biomass accumulation in triticale reveals temporal genetic patterns of regulation. Sci Rep 3:2442. doi: 10.1038/srep02442

Calderón C, Colla M, Jossart JM, Hemeleers N, Cancian G, Aveni N (2019) Bioenergy Europe statistical report 2019. Report biogas. https://bioenergyeurope.org/article/103-statistical-report-2019-biogas.html

Calus MPL, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol 43(1):1–14. doi: 10.1186/1297-9686-43-26

Catchpole WR, Wheeler CJ (1992) Estimating plant biomass: a review of techniques. Australian Journal of Ecology 17(2):121–131

Cheng T, Song R, Li D, Zhou K, Zheng H, Yao X, Tian Y, Cao W, Zhu Y (2017) Spectroscopic estimation of biomass in canopy components of paddy rice using dry matter and chlorophyll indices. Remote Sens 9(4):319. doi: 10.3390/rs9040319

Cobb JN, Juma RU, Biswas PS, Arbelaez JD, Rutkoski J, Atlin G, Hagen T, Quinn M, Ng EH (2019) Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. Theor Appl Genet 132(3):627–645. doi: 10.1007/s00122-019-03317-0

Crain J, Mondal S, Rutkoski J, Singh RP, Poland J (2018) Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. Plant Genome 11(1). doi: 10.3835/plantgenome2017.05.0043

Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun H-J (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2):713–724. doi: 10.1534/genetics.110.118521

Crossa J, Pérez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokosho C (2011) Genomic selection and prediction in plant breeding. J Crop Improv 25(3):239–261. doi: 10.1080/15427528.2011.558767

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112(1):48–60. doi: 10.1038/hdy.2013.16

Daetwyler HD, Villanueva B, Bijma P, Woolliams JA (2007) Inbreeding in genome-wide selection. J Anim Breed Genet 124(6):369–376

de los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res 92(4):295–308. doi: 10.1017/S0016672310000285

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1):375–385. doi: 10.1534/genetics.109.101501

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitao PJ (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36(1):27–46. doi: 10.1111/j.1600-0587.2012.07348.x

Dunagan SC, Gilmore MS, Varekamp JC (2007) Effects of mercury on visible/near-infrared reflectance spectra of mustard spinach plants (*Brassica rapa* P.). Environ Pollut 148(1):301–311. doi: 10.1016/j.envpol.2006.10.023

Eberhart SA (1970) Factors effecting efficiencies of breeding methods. African soils 15(1/3):655–680

EEG (2012) Gesetz für den Ausbau erneuerbarer Energien (Erneuerbare-Energien-Gesetz - EEG). https://www.erneuerbare-energien.de/EE/Redaktion/DE/Gesetze-Verordnungen/eeg_2012_bf.html. Accessed 02 Nov 2019

EEG (2017) Gesetz für den Ausbau erneuerbarer Energien (Erneuerbare-Energien-Gesetz - EEG). http://www.gesetze-im-internet.de/eeg_2014/EEG_2017.pdf. Accessed 02 Nov 2019

Erdle K, Mistele B, Schmidhalter U (2011) Comparison of active and passive spectral sensors in discriminating biomass parameters and nitrogen status in wheat cultivars. Field Crops Res 124(1):74–84. doi: 10.1016/j.fcr.2011.06.007

European Commission (2011a) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the

Committee of the Regions. A Roadmap for moving to a competitive low carbon economy in 2050. COM(2011) 112 final. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0112:FIN:EN:PDF

European Commission (2011b) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Energy Roadmap 2050. COM(2011) 885 final. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0885:FIN:EN:PDF

European Union (2009) Directive 2009/28/EC of the European Parliament and of the Council of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC. Official Journal of the European Union 5:16–62

Fachagentur Nachwachsende Rohstoffe e.V. (2019) Anbaufläche nachwachsender Rohstoffe in Deutschland nach Kulturarten 2016-2018. https://www.fnr.de/fileadmin/news/fnr/2019/PM_2019-09_Anbauzahlen_II.jpg. Accessed 15 Jul 2020

Fahlgren N, Gehan MA, Baxter I (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Curr opin plant biol 24:93–99. doi: 10.1016/j.pbi.2015.02.006

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4. ed. Longman, Essex

FAO (2020) FAOSTAT database. Food and Agriculture Organization of the United Nations. http://www.fao.org/faostat/en/#data/QC. Accessed 17 Jul 2020

Fassnacht FE, Hartig F, Latifi H, Berger C, Hernández J, Corvalán P, Koch B (2014) Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. Remote Sens Environ 154:102–114. doi: 10.1016/j.rse.2014.07.028

Fernandes SB, Dias KOG, Ferreira DF, Brown PJ (2018) Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theor Appl Genet 131(3):747–755. doi: 10.1007/s00122-017-3033-y

Fernandez MGS, Becraft PW, Yin Y, Lübberstedt T (2009) From dwarves to giants? Plant height manipulation for biomass yield. Trends Plant Sci 14(8):454–461. doi: 10.1016/j.tplants.2009.06.005

Filella I, Peñuelas J (1994) The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status. Int J Remote Sens 15(7):1459–1470. doi: 10.1080/01431169408954177

Frels K, Guttieri M, Joyce B, Leavitt B, Baenziger PS (2018) Evaluating canopy spectral reflectance vegetation indices to estimate nitrogen use traits in hard winter wheat. Field Crops Res 217:82–92. doi: 10.1016/j.fcr.2017.12.004

Furbank RT (2009) Plant phenomics: from gene to form and function. Funct Plant Biol 36(10):5–6

Furbank RT, Jimenez-Berni JA, George-Jaeggli B, Potgieter AB, Deery DM (2019) Field crop phenomics: enabling breeding for radiation use efficiency and biomass in cereal crops. New Phytol 223(4):1714–1727. doi: 10.1111/nph.15817

Furbank RT, Tester M (2011) Phenomics–technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16(12):635–644. doi: 10.1016/j.tplants.2011.09.005

Gates DM, Keegan HJ, Schleter JC, Weidner VR (1965) Spectral properties of plants. Appl Opt 4(1):11–20. doi: 10.1364/AO.4.000011

Geiger HH (1972) Wiederherstellung der Pollenfertilität in cytoplasmatisch männlich sterilem Roggen. Theor Appl Genet 42(1):32–33. doi: 10.1007/BF00306075

Geiger HH, Miedaner T (1999) Hybrid rye and heterosis. In: Coors JG, Pandey S (eds) Genetics and Exploitation of Heterosis in Crops. Crop Science Society, America, Madison, Wisconsin, USA, pp 439–450

Geiger HH, Miedaner T (2009) Rye breeding. In: Carena MJ (ed) Handbook of plant breeding: cereals, vol 3. Springer, New York, pp 157–181

Geiger HH, Schnell FW (1970) Cytoplasmic male sterility in Rye (*Secale cereale* L.). Crop Sci 10(5):590–593

Gizaw SA, Garland-Campbell K, Carter AH (2016) Use of spectral reflectance for indirect selection of yield potential and stability in Pacific Northwest winter wheat. Field Crops Res 196:199–206. doi: 10.1016/j.fcr.2016.06.022

Gowda M, Hahn V, Reif JC, Longin CFH, Alheit K, Maurer HP (2011) Potential for simultaneous improvement of grain and biomass yield in Central European winter triticale germplasm. Field Crops Res 121(1):153–157. doi: 10.1016/j.fcr.2010.12.003

Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G (2014) Comparison of single-trait and multiple-trait genomic prediction models. BMC Genetics 15(1):30. doi: 10.1186/1471-2156-15-30

Gutiérrez-Rodríguez M, Reynolds MP, Escalante-Estrada JA, Rodríguez-González MT (2004) Association between canopy reflectance indices and yield and physiological traits in bread wheat under drought and well-irrigated conditions. Aust J Agric Res 55(11):1139–1147. doi: 10.1071/AR04214

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397. doi: 10.1534/genetics.107.081190

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194(3):597–607. doi: 10.1534/genetics.113.152207

Haffke S, Kusterer B, Fromme FJ, Roux S, Hackauf B, Miedaner T (2014) Analysis of covariation of grain yield and dry matter yield for breeding dual use hybrid rye. Bioenerg Res 7(1):424–429. doi: 10.1007/s12155-013-9383-7

Han L, Yang G, Dai H, Xu B, Yang H, Feng H, Li Z, Yang X (2019) Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data. Plant Methods 15:10. doi: 10.1186/s13007-019-0394-z

Hansen PM, Schjoerring JK (2003) Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. Remote Sens Environ 86(4):542–553. doi: 10.1016/S0034-4257(03)00131-7

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci 50(5):1681–1690. doi: 10.2135/cropsci2009.11.0662

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49(1):1–12. doi: 10.2135/cropsci2008.08.0512

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31(2):423–447

Hernandez J, Lobos GA, Matus I, del Pozo A, Silva P, Galleguillos M (2015) Using ridge regression models to estimate grain yield from field spectral data in bread wheat (*Triticum aestivum* L.) grown under three water regimes. Remote Sens 7(2):2109–2126. doi: 10.3390/rs70202109

Heslot N, Akdemir D, Sorrells ME, Jannink J-L (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127(2):463–480. doi: 10.1007/s00122-013-2231-5

Hoerl AE, Kennard RW (1970) Ridge regression. Biased estimation for nonorthogonal problems. Technometrics 12(1):55–67

Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. Nat Rev Genet 11(12):855–866. doi: 10.1038/nrg2897

Huang W, Lamb DW, Niu Z, Zhang Y, Liu L, Wang J (2007) Identification of yellow rust in wheat using in-situ spectral reflectance measurements and airborne hyperspectral imaging. Precis Agric 8(4-5):187–197. doi: 10.1007/s11119-007-9038-9

Hübner M, Oechsner H, Koch S, Seggl A, Hrenn H, Schmiedchen B, Wilde P, Miedaner T (2011) Impact of genotype, harvest time and chemical composition on the methane yield of winter rye for biogas production. Biomass Bioenerg 35(10):4316–4323. doi: 10.1016/j.biombioe.2011.07.021

Igos E, Golkowska K, Koster D, Vervisch B, Benetto E (2016) Using rye as cover crop for bioenergy production: an environmental and economic assessment. Biomass Bioenerg 95:116–123. doi: 10.1016/j.biombioe.2016.09.023

Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128(1):145–158. doi: 10.1007/s00122-014-2418-4

Jacquemoud S, Bacour C, Poilve H, Frangi J-P (2000) Comparison of four radiative transfer models to simulate plant canopies reflectance: direct and inverse mode. Remote Sens Environ 74(3):471–481. doi: 10.1016/S0034-4257(00)00139-5

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics 9(2):166–177. doi: 10.1093/bfgp/elq001

Jia Y, Jannink J-L (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192(4):1513–1522. doi: 10.1534/genetics.112.144246

Jin S, Su Y, Song S, Xu K, Hu T, Yang Q, Wu F, Xu G, Ma Q, Guan H (2020) Non-destructive estimation of field maize biomass using terrestrial lidar: an evaluation from plot level to individual leaf level. Plant Methods 16:1–19. doi: 10.1186/s13007-020-00613-5

Jong SM de, Pebesma EJ, Lacaze B (2010) Above-ground biomass assessment of Mediterranean forests using airborne imaging spectrometry: the DAIS Peyne experiment. Int J Remote Sens 24(7):1505–1520. doi: 10.1080/01431160210145560

Juliana P, Montesinos-López OA, Crossa J, Mondal S, González Pérez L, Poland J, Huerta-Espino J, Crespo-Herrera L, Govindan V, Dreisigacker S, Shrestha S, Pérez-Rodríguez P, Pinto Espinosa F, Singh RP (2019) Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. Theor Appl Genet 132(1):177–194. doi: 10.1007/s00122-018-3206-3

Krause MR, González-Pérez L, Crossa J, Pérez-Rodríguez P, Montesinos-López O, Singh RP, Dreisigacker S, Poland J, Rutkoski J, Sorrells M, Gore MA, Mondal S (2019) Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat. G3: Genes, Genomes, Genet 9(4):1231–1247. doi: 10.1534/g3.118.200856

Langridge P, Fleury D (2011) Making the most of 'omics' for crop breeding. Trends Biotechnol 29(1):33–40. doi: 10.1016/j.tibtech.2010.09.006

Li J, Shi Y, Veeranampalayam-Sivakumar A-N, Schachtman DP (2018) Elucidating sorghum biomass, nitrogen and chlorophyll contents with spectral and morphological traits derived from unmanned aircraft system. Front in Plant Sci 9:1406. doi: 10.3389/fpls.2018.01406

Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D (2018) Machine learning in agriculture: a review. Sensors 18(8):2674. doi: 10.3390/s18082674

Lin Z, Hayes BJ, Daetwyler HD (2014) Genomic selection in crops, trees and forages: a review. Crop Pasture Sci 65(11):1177–1191. doi: 10.1071/CP13363

Liu W, Li Q (2017) An efficient elastic net with regression coefficients method for variable selection of spectrum data. PloS one 12(2):e0171122. doi: 10.1371/journal.pone.0171122

Lorenz AJ (2013) Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. G3: Genes, Genomes, Genetics 3(3):481–491. doi: 10.1534/g3.112.004911

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink J-L (2011) Genomic selection in plant breeding: knowledge and prospects. Adv Agron 110:77–123. doi: 10.1016/B978-0-12-385531-2.00002-5

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120(1):151–161. doi: 10.1007/s00122-009-1166-3

Lucà F, Conforti M, Castrignanò A, Matteucci G, Buttafuoco G (2017) Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. Geoderma 288:175–183. doi: 10.1016/j.geoderma.2016.11.015

Lundqvist A (1956) Self-incompatibility in rye: I. Genetic control in the diploid. Hereditas 42(3-4):293–348

Mahlein A-K, Oerke E-C, Steiner U, Dehne H-W (2012) Recent advances in sensing plant diseases for precision crop protection. Eur J Plant Pathol 133(1):197–209. doi: 10.1007/s10658-011-9878-z

Marulanda JJ, Melchinger AE, Würschum T (2015) Genomic selection in biparental populations: assessment of parameters for optimum estimation set design. Plant breeding 134(6):623–630. doi: 10.1111/pbr.12317

Meier U (1997) Growth stages of mono- and dicotyledonous plants. Blackwell Wissenschafts-Verlag, Berlin

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Miedaner T, Hübner M, Koch S, Seggl A, Wilde P (2010) Biomass yield of self-incompatible germplasm resources and their testcrosses in winter rye. Plant breeding 129(4):369–375. doi: 10.1111/j.1439-0523.2010.01777.x

Miedaner T, Koch S, Seggl A, Schmiedchen B, Wilde P (2012) Quantitative genetic parameters for selection of biomass yield in hybrid rye. Plant breeding 131(1):100–103

Miedaner T, Korzun V, Bauer E (2019) Genomics-based hybrid rye breeding. In: Miedaner T., Korzun V. (eds) Applications of Genetic and Genomic Research in Cereals. Elsevier, Amsterdam, Netherlands, pp 329–348

Miedaner T, Laidig F (2019) Hybrid breeding in rye (*Secale cereale* L.). In: Al-Khayri JM, Jain SM, Johnson DV (eds) Advances in Plant Breeding Strategies: Cereals. Springer International Publishing, Cham, Switzerland, pp 343–372

Monteiro PFC, Angulo Filho R, Xavier AC, Monteiro ROC (2012) Assessing biophysical variable parameters of bean crop with hyperspectral measurements. Sci Agric 69(2):87–94. doi: 10.1590/S0103-90162012000200001

Montes JM, Melchinger AE, Reif JC (2007) Novel throughput phenotyping platforms in plant genetic studies. Trends Plant Sci 12(10):433–436. doi: 10.1016/j.tplants.2007.08.006

Montesinos-López A, Montesinos-López OA, Cuevas J, Mata-López WA, Burgueño J, Mondal S, Huerta J, Singh R, Autrique E, González-Pérez L, Crossa J (2017a) Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper-spectral image data. Plant Methods 13:62. doi: 10.1186/s13007-017-0212-4

Montesinos-López OA, Montesinos-López A, Crossa J, de los Campos G, Alvarado G, Suchismita M, Rutkoski J, González-Pérez L, Burgueño J (2017b) Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. Plant Methods 13(1):4. doi: 10.1186/s13007-016-0154-2

Mulla DJ (2013) Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosyst Eng 114(4):358–371. doi: 10.1016/j.biosystemseng.2012.08.009

Mutanga O, Adam E, Cho MA (2012) High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. Int J Appl Earth Obs 18:399–406. doi: 10.1016/j.jag.2012.03.012

Mutanga O, Skidmore AK (2004) Narrow band vegetation indices overcome the saturation problem in biomass estimation. Int J Remote Sens 25(19):3999–4014. doi: 10.1080/01431160310001654923

Okeke UG, Akdemir D, Rabbi I, Kulakow P, Jannink J-L (2017) Accuracies of univariate and multivariate genomic prediction models in African cassava. Genet Sel Evol 49(1):88. doi: 10.1186/s12711-017-0361-y

Pauli D, Chapman SC, Bart R, Topp CN, Lawrence-Dill CJ, Poland J, Gore MA (2016) The quest for understanding phenotypic variation via integrated approaches in the field environment. Plant Physiol 172(2):622–634. doi: 10.1104/pp.16.00592

Peñuelas J, Filella I (1998) Visible and near-infrared reflectance techniques for diagnosing plant physiological status. Trends Plant Sci 3(4):151–156. doi: 10.1016/S1360-1385(98)01213-8

Pérez P, de los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Genome 3(2):106–116

Piepho H-P (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49(4):1165–1176. doi: 10.2135/cropsci2008.10.0595

Prabhakara K, Hively WD, McCarty GW (2015) Evaluating the relationship between biomass, percent groundcover and remote sensing indices across six winter cover crop fields in Maryland, United States. Int J Appl Earth Obs 39:88–102. doi: 10.1016/j.jag.2015.03.002

Prasad B, Babar MA, Carver BF, Raun WR, Klatt AR (2009) Association of biomass production and canopy spectral reflectance indices in winter wheat. Can J Plant Sci 89(3):485–496. doi: 10.4141/CJPS08137

Prasad B, Carver BF, Stone ML, Babar MA, Raun WR, Klatt AR (2007a) Genetic analysis of indirect selection for winter wheat grain yield using spectral reflectance indices. Crop Sci 47(4):1416–1425. doi: 10.2135/cropsci2006.08.0546

Prasad B, Carver BF, Stone ML, Babar MA, Raun WR, Klatt AR (2007b) Potential use of spectral reflectance indices as a selection tool for grain yield in winter wheat under great plains conditions. Crop Sci 47(4):1426–1440. doi: 10.2135/cropsci2006.07.0492

Prey L, Hu Y, Schmidhalter U (2020) High-throughput field phenotyping traits of grain yield formation and nitrogen use efficiency: optimizing the selection of vegetation indices and growth stages. Front in Plant Sci 10:1672. doi: 10.3389/fpls.2019.01672

Pszczola M, Strabel T, Mulder HA, Calus MPL (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci 95(1):389–400. doi: 10.3168/jds.2011-4338

Rincent R, Charpentier J-P, 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(12):3961–3972. doi: 10.1534/g3.118.200760

Rischbeck P, Elsayed S, Mistele B, Barmeier G, Heil K, Schmidhalter U (2016) Data fusion of spectral, thermal and canopy height parameters for improved yield prediction of drought stressed spring barley. Eur J Agron 78:44–59. doi: 10.1016/j.eja.2016.04.013

Roos APW de, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. Genetics 183(4):1545–1553. doi: 10.1534/genetics.109.104935

Rouse Jr JW, Haas RH, Schell JA, Deering DW (1974) Monitoring vegetation systems in the Great Plains with ERTS. Third ERTS Symposium, NASA SP-351:309–3017

Roux SR, Wortmann H, Schlathölter M (2010) Rye (*Secale cereale* L.) for biogas production-breeding capability. J Kulturpflanz 62(5):173–182

Rutkoski J, Poland J, Mondal S, Autrique E, Pérez LG, Crossa J, Reynolds M, Singh R (2016) Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3: Genes, Genomes, Genet 6(9):2799–2808. doi: 10.1534/g3.116.032888

Schön CC, Utz HF, Groh S, Truberg B, Openshaw S, Melchinger AE (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167(1):485–498. doi: 10.1534/genetics.167.1.485

Schulthess AW, Zhao Y, Longin CFH, Reif JC (2018) Advantages and limitations of multiple-trait genomic prediction for Fusarium head blight severity in hybrid wheat (*Triticum aestivum* L.). Theor Appl Genet 131(3):685–701. doi: 10.1007/s00122-017-3029-7

Sims DA, Gamon JA (2002) Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. Remote Sens Environ 81(2-3):337–354. doi: 10.1016/S0034-4257(02)00010-X

Singh AK, Ganapathysubramanian B, Sarkar S, Singh A (2018) Deep learning for plant stress phenotyping: trends and future perspectives. Trends Plant Sci 23(10):883–898. doi: 10.1016/j.tplants.2018.07.004

Sun J, Poland JA, Mondal S, Crossa J, Juliana P, Singh RP, Rutkoski JE, Jannink J-L, Crespo-Herrera L, Velu G, Huerta-Espino J, Sorrells ME (2019) High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. Theor Appl Genet 132(6):1705–1720. doi: 10.1007/s00122-019-03309-0

Sun J, Rutkoski JE, Poland JA, Crossa J, Jannink J-L, Sorrells ME (2017) Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. Plant Genome 10(2). doi: 10.3835/plantgenome2016.11.0111

Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M (2017) Plant phenomics, from sensors to knowledge. Current Biology 27(15):R770-R783

Tattaris M, Reynolds MP, Chapman SC (2016) A direct comparison of remote sensing approaches for high-throughput phenotyping in plant breeding. Front in Plant Sci 7. doi: 10.3389/fpls.2016.01131

Thomas H, Ougham H (2014) The stay-green trait. J Exp Bot 65(14):3889–3900. doi: 10.1093/jxb/eru037

Thorp KR, Gore MA, Andrade-Sanchez P, Carmo-Silva AE, Welch SM, White JW, French AN (2015) Proximal hyperspectral sensing and data analysis approaches for field-based plant phenomics. Comput Electron Agric 118:225–236. doi: 10.1016/j.compag.2015.09.005

Thorp KR, Wang G, Bronson KF, Badaruddin M, Mon J (2017) Hyperspectral data mining to identify relevant canopy spectral features for estimating durum wheat growth, nitrogen status, and grain yield. Comput Electron Agric 136:1–12. doi: 10.1016/j.compag.2017.02.024

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1):267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tilling AK, O'Leary GJ, Ferwerda JG, Jones SD, Fitzgerald GJ, Rodriguez D, Belford R (2007) Remote sensing of nitrogen and water stress in wheat. Field Crops Res 104(1-3):77–85. doi: 10.1016/j.fcr.2007.03.023

Tilly N, Aasen H, Bareth G (2015) Fusion of plant height and vegetation indices for the estimation of barley biomass. Remote Sens 7(9):11449–11480. doi: 10.3390/rs70911449

Tsai H-Y, Janss LL, Andersen JR, Orabi J, Jensen JD, Jahoor A, Jensen J (2020) Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. Sci Rep 10(1):3347. doi: 10.1038/s41598-020-60203-2

Tucker CJ (1979) Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens Environ 8(2):127–150

Utz HF, Melchinger AE, Schön CC (2000) Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. Genetics 154(4):1839–1849

VanRaden PM, van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92(1):16–24. doi: 10.3168/jds.2008-1514

Velazco JG, Jordan DR, Mace ES, Hunt CH, Malosetti M, van Eeuwijk FA (2019) Genomic prediction of grain yield and drought-adaptation capacity in sorghum is enhanced by multi-trait analysis. Front in Plant Sci 10:997. doi: 10.3389/fpls.2019.00997

Vieira IC, Dos Santos JPR, Pires LPM, Lima BM, Gonçalves FMA, Balestre M (2017) Assessing non-additive effects in GBLUP model. Genet Mol Res 16(2). doi: 10.4238/gmr16029632

Walter JDC, Edwards J, McDonald G, Kuchel H (2019) Estimating biomass and canopy height with LiDAR for field crop breeding. Front in Plant Sci 10:1145. doi: 10.3389/fpls.2019.01145

Wang Y, Mette MF, Miedaner T, Gottwald M, Wilde P, Reif JC, Zhao Y (2014) The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. BMC genomics 15(1):556. doi: 10.1186/1471-2164-15-556

Wang Y, Mette MF, Miedaner T, Wilde P, Reif JC, Zhao Y (2015) First insights into the genotype–phenotype map of phenotypic stability in rye. J Exp Bot 66(11):3275–3284. doi: 10.1093/jxb/erv145

Weber VS, Araus JL, Cairns JE, Sanchez C, Melchinger AE, Orsini E (2012) Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes. Field Crops Res 128:82–90. doi: 10.1016/j.fcr.2011.12.016

White JW, Andrade-Sanchez P, Gore MA, Bronson KF, Coffelt TA, Conley MM, Feldmann KA, French AN, Heun JT, Hunsaker DJ (2012) Field-based phenomics for plant genetics research. Field Crops Res 133:101–112. doi: 10.1016/j.fcr.2012.04.003

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75(2):249–252. doi: 10.1017/S0016672399004462

Wiegmann M, Backhaus A, Seiffert U, Thomas WTB, Flavell AJ, Pillen K, Maurer A (2019) Optimizing the procedure of grain nutrient predictions in barley via hyperspectral imaging. PloS one 14(11):e0224491. doi: 10.1371/journal.pone.0224491

Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193(2):621–631

Wong CK, Bernardo R (2008) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor Appl Genet 116(6):815–824

World Bioenergy Association (2019) Global Bioenergy Statistics 2019. https://worldbioenergy.org/uploads/191129%20WBA%20GBS%202019_LQ.pdf. Accessed 17 Jul 2020

Xie Q, Dash J, Huang W, Peng D, Qin Q, Mortimer H, Casa R, Pignatti S, Laneve G, Pascucci S (2018) Vegetation indices combining the red and red-edge spectral information for leaf area index retrieval. IEEE J Sel Top Appl Earth Obs Remote Sens 11(5):1482–1493. doi: 10.1109/JSTARS.2018.2813281

Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48(2):391–407. doi: 10.2135/cropsci2007.04.0191

Xue J, Su B (2017) Significant remote sensing vegetation indices: a review of developments and applications. J Sens 2017:1–17. doi: 10.1155/2017/1353691

Yang G, Liu J, Zhao C, Li Z, Huang Y, Yu H, Xu B, Yang X, Zhu D, Zhang X (2017) Unmanned aerial vehicle remote sensing for field-based crop phenotyping: current status and perspectives. Front in Plant Sci 8. doi: 10.3389/fpls.2017.01111

Yue J, Yang G, Li C, Li Z, Wang Y, Feng H, Xu B (2017) Estimation of winter wheat above-ground biomass using unmanned aerial vehicle-based snapshot hyperspectral sensor and crop height improved models. Remote Sens 9(7):708. doi: 10.3390/rs9070708

Zhang J, Pu R, Huang W, Yuan L, Luo J, Wang J (2012) Using in-situ hyperspectral data for detecting and discriminating yellow rust disease from nutrient stresses. Field Crops Res 134:165–174. doi: 10.1016/j.fcr.2012.05.011

Zhang X, Xu F, He Y, Li X, Chen D, Wang G, Shi I (2019) Estimation of corn canopy chlorophyll content using derivative spectra in the O2–A absorption band. Front in Plant Sci 10:1047. doi: 10.3389/fpls.2019.01047

Zhang Z, Masjedi A, Zhao J, Crawford MM (2017) Prediction of sorghum biomass based on image based features derived from time series of UAV images. 2017 IEEE International Geoscience and Remote Sensing Symposium. doi: 10.1109/IGARSS.2017.8128413

Zheng Q, Huang W, Cui X, Dong Y, Shi Y, Ma H, Liu L (2019) Identification of wheat yellow rust using optimal three-band spectral indices in different growth stages. Sensors 19(1):35. doi: 10.3390/s19010035

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2):301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Acknowledgments

I would like to thank the following people and organizations for their support, especially in the context of the COVID-19 pandemic, without whose help this thesis would never have been possible:

First, I would like to express my deepest gratitude to my supervisor Prof. Dr. Thomas Miedaner for allowing me to work under his auspices. I am especially thankful for the freedom and support offered me to conduct this research, the possibility to attend international conferences, his endless patience in proofreading the manuscripts and discussing any matter of relevance to me, including the next steps in my professional career. Many thanks also for encouraging my creativity, productivity, and well-being not only when we were physically present at the Institute, but also in the challenging remote work era.

This research was funded by the German Federal Ministry of Food and Agriculture (BMEL) through the German Agency for Renewable Resources (FNR), grant number FKZ 22019716. I would like to thank them for their generous aid.

I am also greatly indebted to Prof. Dr. Hans-Peter Piepho for his always prompt answers to all my questions regarding statistics, valuable suggestions as co-author, and for serving on my graduate committee. Many thanks to Prof. Dr. Klaus Pillen for being part of the examination committee. Thanks are also due to Prof. Dr. H. F. Utz, who gave me much valuable advice in the early stages of this work.

Further, I wish to thank all the co-authors and project partners from KWS LOCHOW GmbH, who worked alongside me during this research. I gratefully acknowledge Dr. Angela-Maria Bernal-Vasquez and Dr. Patrick Thorwarth for the numerous and fruitful discussions, all the good advice on statistical and coding-related issues, as well as contributions to my publications. Sincere thanks to Dr. Andres Gordillo, Christian Jebsen, and Dr. Philipp Steffan for the excellent joint work and commitment to this project. I also thank Dr. Peer Wilde for his helpful contributions to this project. Many thanks to Dr. Viktor Korzun for his assistance regarding molecular data. Likewise, I would like to thankfully recognize the excellent support of the technical staff at each experimental site, especially to Hans-Otto Wegener, Jörn-Claus Gudehus, Karsten Sell for seed production and conducting field trials.

Special thanks to all my colleagues at the Rye Research Group as well as all members of the working groups of Prof. Dr. T. Würschum, Dr. W. Leiser, Dr. H. P. Maurer, Dr. V. Hahn, and PD. Dr. C. F. H. Longin for the excellent working environment at the State Plant Breeding Institute. Thanks also to Ms. Kurka, Ms. Schrader, and Ms. Kösling for helping me to solve administrative issues.

I extend my sincere thanks to KWS Saat SE & Co. KGaA, particularly to Dr. Lisa-Marie Braune for her flexibility and support during the final stages of this thesis.

During my time in Hohenheim, I have met so many great people who have made this experience a wonderful milestone in my life. Many thanks to all my M.Sc. and Ph.D. colleagues and friends, for the excellent moments shared during work and after, the team spirit, the great discussions, constructive feedback, and the words of support, especially in the moments when I needed them most.

Many thanks to M.Sc. Thea Mi Weiss and Dr. Matías Delgadino for proofreading this thesis.

This thesis is in memory of Agric. Eng. (M.Sc.) Eduardo Ruiz Posse who has aroused my interest in remote sensing and together with other Professors at the National University of Córdoba (Argentina) has guided me in the early stages of my career. My sincere gratitude to the outstanding academic staff involved in the Crop Sciences Master's Program (University of Hohenheim), especially to Prof. Dr. A. E. Melchinger for having deepened my interest in plant breeding and having encouraged me to follow my scientific vocations.

I also wish to sincerely thank the support of my family, which I am lucky to have both in Argentina and Germany. To my Argentine family, to my parents, I am immensely grateful for all the sacrifices you made for your children and for always encouraging me to pursue my dreams. Many thanks to my brother and my bigger family for always trusting me. To my German family, thank you for opening the doors of your home and your heart to me.

Last but not least, I would like to express my deepest gratitude to my wife Sabrina for her unconditional and continuous support since the beginning of this academic journey, which took us from Córdoba to Stuttgart. Thank you for giving me the opportunity, freedom, and especially the time to finish this work.

Curriculum Vitae

PERSONAL INFORMATION

Name:	**RODRIGO JOSÉ GALÁN**
Date and place of birth:	June 24, 1988. Córdoba, Argentina.

EDUCATION

10/2017 – 09/2020	**Ph.D. CANDIDATE – PLANT BREEDING** UNIVERSITY OF HOHENHEIM (State Plant Breeding Institute)
10/2015 – 09/2017	**CROP SCIENCES (M.Sc.)** UNIVERSITY OF HOHENHEIM ▪ Plant Breeding and Seed Science Specialization.
02/2006 – 09/2011	**AGRICULTURAL ENGINEERING, 5 yr degree.** NATIONAL UNIVERSITY OF CÓRDOBA (Argentina).
09/2010 – 03/2011	**ACADEMIC EXCHANGE** UNIVERSITY OF GÖTTINGEN

WORK EXPERIENCE

10/2020 – Today	**MAIZE BREEDER** KWS SAAT SE & Co. KGaA
09/2016 – 10/2016	**PLANT BREEDING INTERNSHIP (Full-time)** UNIVERSITY OF HOHENHEIM (Prof. Dr. A. E. Melchinger)
08/2013 – 03/2015	**PRODUCT DEVELOPER - Seeds Traits and Oils (ST&O)** DOW AGROSCIENCES ARGENTINA S.A.
08/2012 – 08/2013	**FIELD PROMOTER** DOW AGROSCIENCES ARGENTINA S.A.

MEMBERSHIPS

- Society for Plant Breeding “Gesellschaft für Pflanzenzüchtung e.V.” (GPZ), Germany.
- 2019-2020: Ph.D. Students' Representative at the “Hohenheim Research Center for Global Food Security and Ecosystems” (GFE).

Declaration

Annex 3

Declaration in lieu of an oath on independent work

according to Sec. 18(3) sentence 5 of the University of Hohenheim's Doctoral Regulations for the Faculties of Agricultural Sciences, Natural Sciences, and Business, Economics and Social Sciences

1. The dissertation submitted on the topic

"Integration of hyperspectral, genomic, and agronomic data for early prediction of biomass yield in hybrid rye (*Secale cereale* L.)"

is work done independently by me.

2. I only used the sources and aids listed and did not make use of any impermissible assistance from third parties. In particular, I marked all content taken word-for-word or paraphrased from other works.

3. I did not use the assistance of a commercial doctoral placement or advising agency.

4. I am aware of the importance of the declaration in lieu of oath and the criminal consequences of false or incomplete declarations in lieu of oath.

I confirm that the declaration above is correct. I declare in lieu of oath that I have declared only the truth to the best of my knowledge and have not omitted anything.

Place, Date

Signature