Theoretical and Applied Genetics
https://doi.org/10.1007/s00122-020-03613-0

ORIGINAL ARTICLE

The value of early-stage phenotyping for wheat breeding in the age of genomic selection

Daniel Borrenpohl[1] · Mao Huang[1] · Eric Olson[2] · Clay Sneller[1]

Received: 8 October 2019 / Accepted: 15 May 2020

Abstract

Key message **Genomic selection using data from an on-going breeding program can improve gain from selection, relative to phenotypic selection, by significantly increasing the number of lines that can be evaluated.**

Abstract The early stages of phenotyping involve few observations and can be quite inaccurate. Genomic selection (GS) could improve selection accuracy and alter resource allocation. Our objectives were (1) to compare the prediction accuracy of GS and phenotyping in stage-1 and stage-2 field evaluations and (2) to assess the value of stage-1 phenotyping for advancing lines to stage-2 testing. We built training populations from 1769 wheat breeding lines that were genotyped and phenotyped for yield, test weight, Fusarium head blight resistance, heading date, and height. The lines were in cohorts, and analyses were done by cohort. Phenotypes or GS estimated breeding values were used to determine the trait value of stage-1 lines, and these values were correlated with their phenotypes from stage-2 trials. This was repeated for stage-2 to stage-3 trials. The prediction accuracy of GS and phenotypes was similar to each other regardless of the amount (0, 50, 100%) of stage-1 data incorporated in the GS model. Ranking of stage-1 lines by GS predictions that used no stage-1 phenotypic data had marginally lower correspondence to stage-2 phenotypic rankings than rankings of stage-1 lines based on phenotypes. Stage-1 lines ranked high by GS had slightly inferior phenotypes in stage-2 trials than lines ranked high by phenotypes. Cost analysis indicated that replacing stage-1 phenotyping with GS would allow nearly three times more stage-1 candidates to be assessed and provide 0.84–2.23 times greater gain from selection. We conclude that GS can complement or replace phenotyping in early stages of phenotyping.

Abbreviations

AST1	Predictions and selections based on GEBVs using all stage-1 phenotypic data
FHB	Fusarium head blight
FST1	Predictions and selections based on GEBVs using ½ stage-1 phenotypic data-based selecting lines based on family relations
GEBV	Genomic estimated breeding values
GS	Genomic selection
NST1	Predictions and selections based on GEBVs using no stage-1 phenotypic data
NST1-1K	Same as NST1 except predictions made with a just 10% of the markers
PHEN	Predictions and selections based on phenotypes
PS	Phenotypic selection
RST1	Predictions and selections based on GEBVs using ½ stage-1 phenotypic data based on random selection of lines
TP	Training population

Communicated by Jose Crossa.

✉ Clay Sneller
sneller.5@osu.edu

[1] Department of Horticulture and Crop Science, Ohio Agriculture Research and Development Center, The Ohio State University, 1680 Madison Av, Wooster, OH 44691, USA

[2] Department of Plant, Soil, and Microbial Science, Michigan State University, 1066 Bogue St, East Lansing, MI 48824, USA

Introduction

Genomic selection (GS) was first proposed by Meuwissen et al. (2001). In GS, a training population (TP) is formed if individuals who are phenotyped and genotyped with molecular markers. Those data are co-analyzed to build a prediction model. The model can then be used to predict the trait value of unphenotyped individuals that are related

Published online: 01 June 2020

to the TP and have been genotyped with the same markers. The GS model calculates genomic estimated breeding values (GEBVs) for the unphenotyped individuals that can be used for selection in the same manner as phenotypic data.

Much early research on GS focused on requirements to optimize TPs and GS models. Many questions remain, however, regarding GS implementation in plant breeding programs. In general, two applications of GS have been proposed. The first application is for the population improvement phase of a program where recombination creates new genetic variation and testable progeny are generated. GS in this phase can increase the rate of genetic gain by reducing the duration of a recurrent selection breeding cycle (Bassi et al. 2015; Gaynor et al. 2017; Heffner et al. 2010; Jannink et al. 2010). GS can also be applied to the product development phase of breeding where progeny are evaluated for performance in a stage-gate process. GS could enhance field trial selection by integrating genotypic and phenotypic data as opposed to advancing lines using phenotypic data alone (Bernal-Vasquez et al. 2017; Longin et al. 2015; Gaynor et al. 2017; He et al. 2016; Marulanda et al. 2016; Tolhurst et al. 2019). Field-testing lines is the most expensive phase of a breeding program: GS could make it considerably more efficient in terms of money spent and resource allocation. A plant breeding program can potentially use GS in place of phenotypic selection (PS) to advance lines, or it can integrate GS and PS predictions to make more informed selections as compared to PS alone.

GS can be used to enhance the product development phase (e.g., field testing) of breeding by integrating GS into the selection of lines to advance to the next stages of testing (Longin et al. 2015; Marulanda et al. 2016; Gaynor et al. 2017). This application of GS would have the greatest value when applied to early stages of field evaluations (stage-1, stage-2), where phenotypic selection among a large set of lines is based on data from few replications and locations, often resulting in lower accuracy than later stages of testing. Thus, much of the genetic variation generated by a program is evaluated in trials that may produce low entry-mean heritability due to use of few replications and test sites. A program could use GEBVs and phenotypes to make selections, or even replace the earliest stage of PS with GS. In this application, the TP could be breeding lines that have been phenotyped in past trials. The TP is used to obtain the GEBVs of new lines that are either in stage-1 or stage-2 trials or are candidates to enter stage-1 trials. In the first instance, phenotypic and GEBVs could both be used to advance lines from one stage of testing to the next stage, while in the later situation the stage-1 trial would not even be conducted: the candidate lines would be advanced to stage-2 testing based solely on their GEBVs. Using simulations, Marulanda et al. (2016) showed that selecting lines based on GEBVs prior to field testing and reducing the years of field testing prior to selecting lines to be used as parents improved annual genetic gain compared to a traditional phenotyping scheme. Gaynor et al. (2017) also used simulations to evaluate the merits of using GS in the product development and population improvement (rapid cycling) phases of a breeding program. Their results showed that using GS to select superior lines either prior to stage-1 testing or in conjunction with stage-1 testing was superior to PS alone: using GS in both phases of a breeding program provided the greatest gain.

The conclusions of these simulation results have been supported by empirical results in wheat. The primary obstacle to implementing this strategy is that GS, like PS, needs to predict line performance in an unobserved future season. This need is not addressed in many experiments that estimate GS accuracy using cross-validation. In cross-validation, data from the same set of environments are used to build the GS model and in the validation, hence leading to upward bias of GS model prediction accuracy in actual breeding situations (Michel et al. 2016). Michel et al. (2017) assessed the predictive ability of GS and PS by correlating GEBVs and phenotypes from stage-1 trials to phenotypes obtained from multi-environment trials conducted in other seasons. They reported that the accuracy of GS + PS was double that of PS alone for wheat grain yield and protein content. They also reported that GEBVs estimated using data from past years had a higher correlation with future performance than did phenotypes. Similar results have been noted in wheat from breadmaking quality traits (Michel et al. 2016). Belamkar et al. (2018) conducted a similar investigation using wheat grain yield. They found that both GS and PS prediction accuracy between stages of testing was variable from season to season, with GS and PS each having seasons where one had double the prediction accuracy of the other. They recommended that selection in stage-1 trials be done using both PS and GS. Others have also noted that prediction accuracies varied across sets of lines and trials (Sallam and Smith, 2016).

The use of GS could be beneficial in the early stages of field testing even if GS and PS had similar accuracy if genotyping costs less than phenotyping (Rajsic et al. 2016). If GS costs less than PS, then more stage-1 lines could be evaluated using GS than by PS which would result in increased selection intensity. Eliminating a stage of phenotyping could also reduce the duration of a breeding cycle. Our objectives were (1) to compare the prediction accuracy of GS and PS in stage-1 and stage-2 field evaluations and (2) to assess the use of GS in stage-1 testing.

Materials and methods

Phenotypic data

Each season the Ohio State University (OSU) winter wheat breeding program evaluates lines in stage-1 through stage-4 trials, with stage-1 trials being the least advanced trial and

stage-4 trials being the most advanced trial (Fig. 1, Table 1). In a typical cycle, we make 150 crosses among 25 elite parents chosen based on their phenotypes in past trials. The parents are all chosen for their superior trait values, especially yield. There are some common parents among cohorts, and all parents are samples from the adapted gene pool for the upper Midwest. The 150 crosses are advanced to the F4 generation where 100 spikes are randomly selected from each family, and seed of each spike composited to form an F4:5 line. The 15,000 F4:5 lines are grown in a single 1M rows, and 1000 are harvested as F4:6 bulks that are entered into the stage-1 trials. This phase of the breeding takes five seasons and four years. In a typical year, the program phenotypes about 1000 new lines in stage-1, 250 lines in stage-2, 40 lines in stage-3, and 20 lines in stage-4 trials. Stage-1 trials are conducted at one Ohio location (Wooster), stage-2, and stage-3 trials are conducted at three Ohio locations (Wooster, Northwest Agricultural Research Station, North Central Agricultural Research Station), and stage-4 trials are conducted at six Ohio locations (all stage-3 locations plus locations in Crawford, Darke, and Pickaway counties).

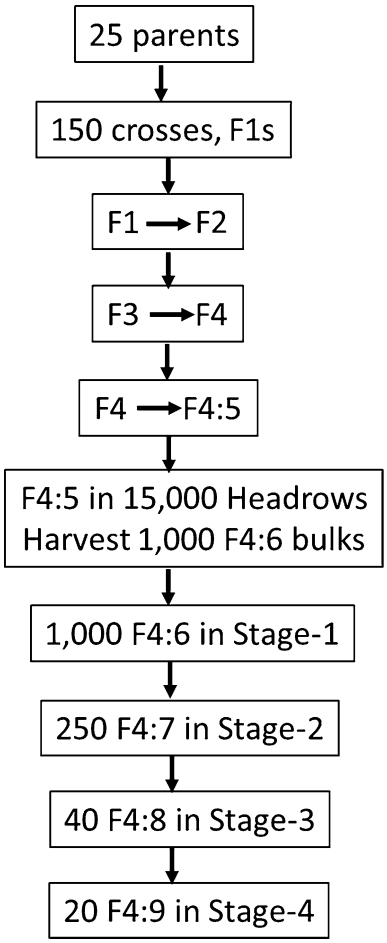

Fig. 1 Schematic of a typical breeding cycle of the Ohio State University winter wheat breeding program. The numbers of crosses and lines varies somewhat from cycle to cycle

Phenotypic data from stage-1 through stage-4 trials from seasons 2013–2018 were used for this analysis (Table 1). Five wheat traits were analyzed: grain yield (tonnes/ha), test weight (kg/hL), height (cm), heading date (Julian days to 50% of the plants attaining Feekes stage 10.1) and resistance to Fusarium head blight (FHB, caused by *Fusarium graminearum*). Grain yield and test weight data were collected in all testing locations, while height, heading date and FHB were only collected in Wooster, OH. Grain yield and test weight were collected by the plot combine and adjusted to 13% moisture. Harvested area of stage-1 yield plots was 2.32 m^2 from 2013 to 2017 and 3.25 m^2 in 2018. Harvested area of stage-2, stage-3, and stage-4 yield plots was 4.64 m^2. Stage-1 and stage-2 trials consist of one replication per environment, stage-3 trials have two replications per environment, and stage-4 trials have three to four replications per environment.

Resistance to FHB was assessed in an inoculated and misted FHB nursery as described by Sneller et al. (2010). Each replication in the FHB nursery consists of a single one-meter row. Stage-1 FHB trials had two replications, while the stage-2, stage-3, and stage-4 FHB trials had three replications. FHB index data were collected approximately 24 days after inoculation and flowering (Feekes stage 10.5)

Table 1 Summary of the number lines in a cohort, the number of lines genotyped, and seasons (years) of testing whose phenotypic data were used in the analysis

Cohort	# Lines in cohort	# Lines genotyped	Stage 1	Stage 2	Stage 3	Stage 4	Other
OH12	807	73	2013 (73)	2014 (73)	2015 (73)	2016 (8)	2017 Stage-4(9) 2018 Stage-4(1)
OH13	889	38	2014 (38)	2015 (38)	2016 (38)	2017 (11)	2018 Stage-4(3)
OH14	715	249	2015 (249)	2016 (249)	2017 (38)	2018 (11)	
OH15	603	251	2016 (251)	2017 (251)	2018 (72)		
OH16	478	473	2017 (473)	2018 (252)			
OH17	775	656	2018 (656)				

The number in parentheses is the number of genotyped lines that were phenotyped in that year and trial

by assessing the percentage of spikelets showing FHB symptoms in three 0.33 m areas per replication.

Grain yield and test weight data were spatially analyzed for within trial variation using P-splines in R package SpATS (v.1.0-9) using the "SpATS" function (Rodríguez-Álvarez et al. 2018). The model used to analyze each trial separately was:

$$\mathbf{y} = \mathbf{X}_g\mathbf{g} + \mathbf{X}_s\mathbf{B}_s + \mathbf{Z}_s\mathbf{s} + \mathbf{Z}_u\mathbf{u} + \mathbf{e}$$

where $\boldsymbol{y}$ is the vector of phenotypic observations, $\boldsymbol{g}$ is the vector of fixed genotypic effects, $\mathbf{X}_g$ is the design matrix of fixed genotypic effects, $\mathbf{X}_s\mathbf{B}_s$ and $\mathbf{Z}_s\mathbf{s}$ form the fixed and random component, respectively, of the mixed model expression of the smooth spatial surface, $\boldsymbol{u}$ is the vector of random row and column effects accounting for discontinuous field variation, $\mathbf{Z}_u$ is the design matrix of random row and column effects, and $\boldsymbol{\varepsilon}$ is the vector of residuals (Velazco et al. 2017). Stage-1 and stage-2 trials were spatially adjusted using 400 knots, and stage-3 and stage-4 trials were spatially adjusted using 100 knots. Best linear unbiased estimates (BLUEs) were estimated for each line within a trial and used in downstream analysis.

Each trait was analyzed with a random effects model using R package lme4 (v. 1.1-21) using the "lmer" function (Bates et al. 2015) to estimate best linear unbiased predictors (BLUPs) for each line. Line BLUPs were estimated for each trait with the following model:

$$y_{ijk} = \mu + g_i + e_j + ge_{ij} + t_k(e_j) + \varepsilon_{ijk}$$

where y_{ijk} is the ijkth phenotypic observation, μ is the overall mean, g_i is the effect of the ith genotype, e_j is the effect of the jth environment, ge_{ij} is the interaction of the ith genotype with the jth environment, $t_k(e_j)$ is the effect of the kth trial nested within the jth environment, and ε_{ijk} is the error of the ijkth observation. All effects were considered random. The same model was also used considering genotypes effects to be fixed to generate best linear unbiased estimators (BLUEs). We obtained the correlation of the BLUEs and BLUPs using the CORR procedure of SAS (SAS, 2017)

Genotypic data

DNA was extracted from 1769 lines OSU wheat lines using the Qiagen DNeasy 96 Plant Kit (Qiaqen Inc., Valencia, CA, USA). Genomic libraries were prepared according to Poland et al. (2012). Genotyping-by-sequencing of 100 base pair single end reads was done at Michigan State University using the Illumina HiSeq 4000 platform. A set of 400 randomly selected lines was first used for single nucleotide polymorphism (SNP) discovery by comparing their sequences to the wheat reference genome (Appels et al. 2018) in the TASSEL-GBS pipeline (Glaubitz et al. 2014). In total, ~ 190,000 SNPs were detected. These same SNPs were called for all the other lines used in analysis. Only SNPs with less than 20% missing data and greater than 0.05 minor allele frequency were retained. Missing marker scores were imputed using an expectation–maximization algorithm in R package rrBLUP (v. 4.6) using the "A.mat" function (Endelman 2011). To ensure even marker genome coverage and reduce marker redundancy, a SNP tagging procedure using a threshold of $r = 0.8$ was conducted to identify the most informative markers (Rinaldo et al. 2005; Huang et al. 2016). After this tagging procedure, a total of 12,037 high-quality and evenly distributed SNPs were used for analysis.

Training populations for genomic selection

In the OSU winter wheat breeding program, $F_{4:5}$ families are planted in a single one-meter row called a headrow. Lines are named by the season when they are selected from a headrow nursery. For example, lines that were harvested from the 2012 headrow nursery are identified as OH12 lines and collectively referred to as the OH12 cohort. The evaluation of a new cohort of lines is initiated every year. Each cohort is then subjected to selection in stage-1 through stage-4 trials (Table 1). This study used data from the OH12 to OH17 cohorts. Not all lines from a cohort were genotyped due to lack of remnant seed when this study was initiated. In total, data from 1740 breeding lines from six cohorts and 29 checks were available for potential use in training populations. Population structure among all 1769 lines was assessed using principal component analysis of the marker data in the R stats package (v. 3.5.3) using the "prcomp" function (R Core Development Team 2019). Euclidean distance between cohorts was calculated using the first three principal components. Euclidean distance was first calculated between all lines individually and then averaged to calculate mean Euclidean distance between cohorts. The OH16 and OH17 cohorts appeared to be split into two groups (Fig. 2): we grouped OH16 and OH17 lines with a PC1 value of < -10 into left groups (OH16L, OH17L) and those with PC1 values > -10 into right groups (OH16R, OH17R). Genetic differentiation between the cohorts was estimated using pairwise F_{ST} according to Weir and Cockerham (1984) using the R package hierfstat (v. 0.04-22) using the "pairwise.WCfst" function (Goudet and Jombart 2015).

We assessed the ability of using phenotypic BLUPs or using GEBVs to predict performance of a cohort in the next stage of testing. Specifically, we estimated line BLUPs and GEBVs for each cohort in stage-1 and stage-2 trials and correlated those values with their phenotypes in stage-2 and stage-3 trials, respectively. This correlation is our definition of a prediction accuracy. Prediction accuracies were estimated for each trait and for each cohort separately, then

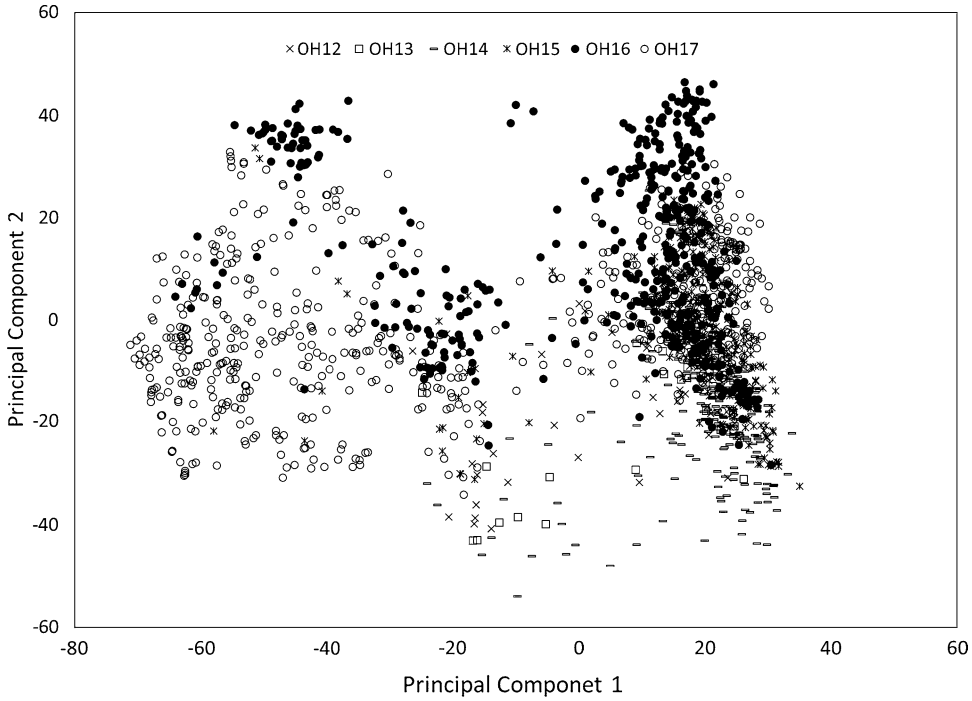

Fig. 2 Plot of the first two principal components of the marker data of 1740 wheat lines from the OH12 to OH17 cohorts

averaged over cohorts to estimate the mean prediction accuracy for each trait and prediction method.

To prevent upward bias of prediction accuracy and simulate a breeding program, we adopted the approach shown in Fig. 3 to form training populations for each cohort that reflect the data that would be available to a breeder. To make selections, a breeder has data from the current selection season and other seasons, but not from the target season which lies in the future. For example, assume we are estimating the stage-1 to stage-2 (stage-1 > 2) prediction accuracy of the OH15 cohort (Fig. 3). The OH15 cohort was in the 2016 stage-1 trial, and selected lines were advanced to the 2017 stage-2 trial: for the OH15 cohort, 2016 is the selection season and 2017 is the target season (Table 1, Fig. 3). First, all phenotypic data from the 2017 target season are deleted from any possible TP for the OH15 cohort. Next, phenotypic data from the 2016 selection season and other seasons (2013, 2014, 2015, and 2018) were used as a TP to build a GS model to predict the performance of OH15 lines in 2017. Any data from OH15 lines from the other seasons (2018 in this example) were removed from the TP. The remaining phenotypic data were used to estimate phenotypic BLUPs for all lines. Then, only phenotypic BLUPs for lines that were genotyped were used as the TP to build a GS model and estimate GEBVs for the OH15 cohort. The phenotypic BLUPs (referred to as PHENs) and GEBVs for OH15 lines were then correlated with their stage-2 phenotypes to calculate the prediction accuracy of PHEN and GEBVs from stage-1 to stage-2 (termed stage-1 > 2). This methodology was employed for all cohorts to estimate prediction accuracies from stage-1 > 2, and from stage-2 to stage-3 (stage-2 > 3). Stage-2 and stage-3 target season phenotypic data were averaged across environments, and the mean was used in the correlation analysis. Any analysis with less than 30 lines was not conducted.

Genomic selection

GS models were built using R package rrBLUP (v. 4.6) to estimate marker effects for each trait using ridge regression BLUP (RR-BLUP) using the "mixed.solve" function (Endelman, 2011). RR-BLUP was chosen primarily for computational efficiency because Huang et al. (2016) found that RR-BLUP had comparable accuracy to other GS models for all wheat traits in the Ohio State University breeding program. The model used to estimate marker effects was:

$$\mathbf{y} = \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of phenotypic BLUPs, $\mathbf{Z}$ is the design matrix of marker values (−1,0,1), $\mathbf{u}$ is the vector of marker effects, and $\boldsymbol{\varepsilon}$ is the vector of residuals.

We evaluated the value of stage-1 phenotypic data by calculating GEBVs with varying amounts of stage-1 data (Table 2). The GEBVs were estimated using: (1) a TP containing all stage-1 phenotypic data from the selection season and other seasons (AST1); (2) a TP using no stage-1 phenotypic data (NST1); or (3) TPs using stage-1 phenotypic data from only ½ of lines in the cohort being

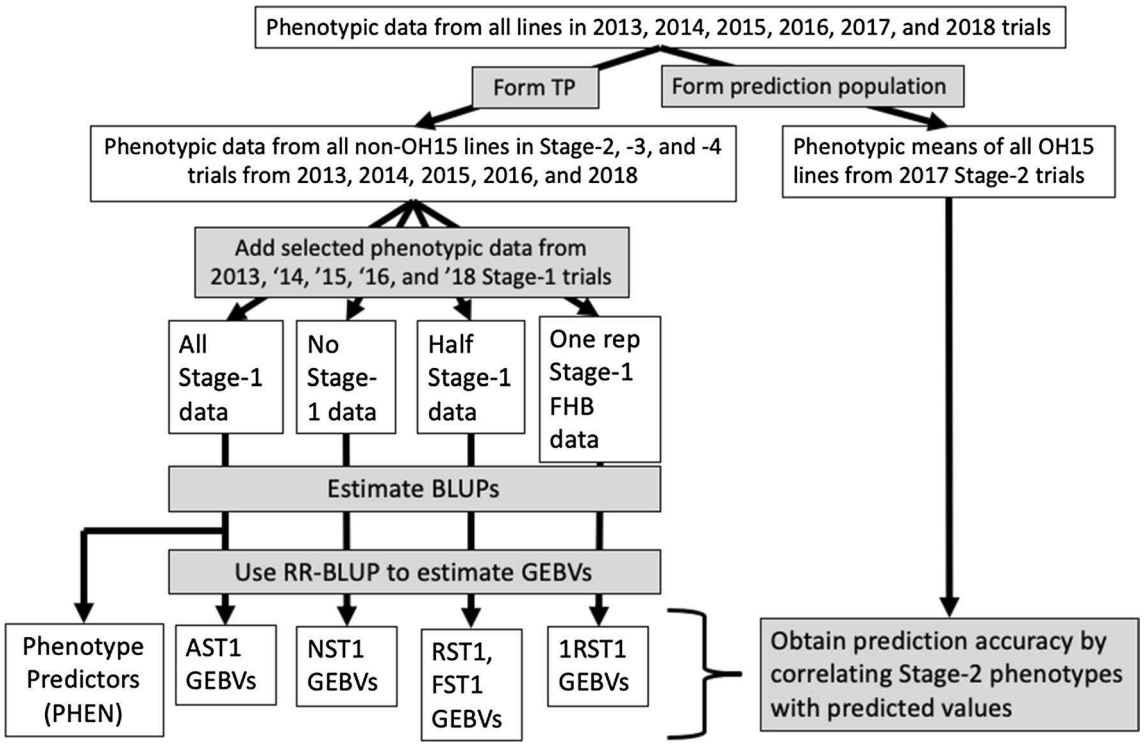

Fig. 3 Diagram of forming prediction populations and the training populations (TP) used to obtain predicted values (PHEN, AST1, NST1, RST1, FST1, 1RST1) of lines in the OH15 cohort for estimating OH15 stage-1 > 2 prediction accuracy. The stage-1 trials of OH15 lines were conducted in 2016, while the stage-2 trials of selected OH15 lines were conducted in 2017

predicted (Table 2). We used two strategies to select ½ the stage-1 lines for inclusion in the TPs: 1) randomly select ½ the stage-1 lines (RST1); and 2) randomly select ½ the stage-1 lines from within each family in the cohort (FST1). We created ten random subsets of lines for the RST1 and FST1 analyses, tested each separately, and compared the average accuracy of the 10 analyses against the accuracy of the other TPs. In addition, two more TPs were used only for FHB index: (1) BLUPs using only stage-1 phenotypic data from only one replication (1RPHEN), and (2), a TP that uses only one stage-1 replication from the selection season (1RST1). Two analyses were done using data from each replication separately. The results from these

Table 2 Summary of methods used to predict the value of lines

Prediction method code	Type of prediction	Phenotypic and marker data used
PHEN	Phenotype BLUP[b]	All stage-1 data, no markers
1RPHEN[c]	Phenotype BLUP	One replication of stage-1 FHB data, no markers
AST1	GS GEBV[a]	All stage-1 data, all markers
NST1	GS GEBV	No stage-1 data, all markers
NST1-1 K	GS GEBV	No stage-1 data, only 1212 markers
RST1	GS GEBV	Data from random selection of ½ of stage-1 lines in cohort, all marker data
FST1	GS GEBV	Data from random selection of ½ of stage-1 lines from each family in cohort, all marker data
1RST1[c]	GS GEBV	One replication of stage-1 FHB data, all marker data

Methods vary by the amount of stage-1 phenotypic data and marker data used in the prediction

[a]*GEBV* genomic estimated breeding value

[b]*BLUP* best linear unbiased predictor

[c]Only used to analyze FHB index

two replications were averaged and compared. The data used for each phenotypic and GS model are summarized in Table 2.

We assessed the prediction accuracy of the NST1 method when using a subset of markers. The NST1-1 K method is a NST1 prediction based on a set of 1123 SNPs as compared to 12,037 SNPs used in NST1. The 1123 markers were selected by repeating the SNP tagging procedure detailed by Rinaldo et al. (2005) using a threshold of $r = 0.0002$: this is a very low threshold but was required to obtain a set of 1123 markers.

We compared the accuracy of GS when using BLUEs or BLUPs. This was done using tenfold cross-validation in a training population of 886 lines that had been tested in stage-2 or higher trials. Data from stage-1 trials were not used in this analysis. The analysis was conducted using rrBLUP.

Value of stage-1 selections in stage-2 trials

We determined what percentage of stage-1 lines ranked as the best 5, 10, 15, and 20% based on predicted values (PHEN, AST1, NST1) were also in the best 5,10, 15, and 20% of lines based on stage-2 phenotypes. Also, the average stage-2 trait value of the best 5, 10, 15, and 20% of the stage-1 lines, based on predicted values, was determined and compared to one another using a *t* test. This was done for grain yield and FHB for the OH14, OH15, and OH16 cohorts.

Cost analysis and gain from indirect selection

Genotyping and phenotyping costs for yield and FHB index were estimated for stage-1 trials. Stage-1 operating costs of testing 1000 lines were estimated for various prediction schemes. We assumed a fixed stage-1 budget that was equal to that of phenotyping 1000 lines with a trial at one location with one replication for grain yield and two replications for FHB. We assumed that 250 stage-1 lines would be advanced to stage-2 testing regardless of selection scheme and selection intensity (k) was calculated for each selection scheme.

We assessed gain from indirect selection where the selection trait (trait X) is the predicted values of stage-1 lines and the target trait (trait Y) is performance in stage-2 stage such that:

$$G_{y.x} = r_g k_x h_x \sigma_{ay}$$

where r_g is the genetic correlation between traits X and Y, k_x is the selection intensity for trait X, h_x is the square root of the heritability of trait X, and σ_{ay} is the additive variance of trait Y. The genetic correlations (r_g) between the predicted values (PHEN or NST1) and stage-2 phenotypes were estimated using R package sommer (v. 3.9.3) using the "cov2cor" function (Covarrubias-Pazaran 2016). The heritability of PHEN from stage-1 trials was calculated as

$$\frac{\sigma_g^2}{\sigma_{g+}^2 \sigma_{\text{error}}^2}$$

We estimated the heritability of the stage-1 GEBVs with R package ASReml-R (Butler et al. 2009) using the genomic BLUP model, which was shown to be equivalent to rrBLUP model (Habier et al. 2007). In this model, an inverse additive relationship matrix was first formatted according to Nazarian and Gezan (2016) and one-time prediction was conducted. Heritability was calculated with the formula:

$$\frac{\sigma_A^2}{\sigma_{A+}^2 \sigma_{\text{error}}^2}$$

where σ_A^2 was the additive genetic variance from genomic BLUP model and σ_{error}^2 is the error variance.

Results

We want to start with some clarification of terms. A "stage-1 line" is any line that is being evaluated for quantitative traits, and in particular for yield, for the first time. This initial evaluation could be in a stage-1 phenotypic trial or the evaluation could be based on only predicted values from GS. We refer to "stage-1 trial" as a field evaluation. A "stage-2 trial" is a trial of selected stage-1 lines and conducted in multiple environments. If the stage-1 lines are selected based solely on GEBVs, then they could be advanced to the stage-2 trial without actually being in a stage-1 phenotyping trial. Thus, a stage-2 trail could consist of lines that have not been phenotyped in a stage-1 trial: we chose to retain the stage-2 identifier for consistency.

We used phenotype BLUPs in the GS analysis instead of BLUES though this raises the question of double shrinkage and some have suggested that using BLUEs is more appropriate sand provides higher GS accuracy that BLUPs (Garrick et al. 2009; Ostersen et al. 2011; Piepho et al. 2008). Huang et al. (2016) found that GS accuracy was identical for all wheat traits in our program using BLUEs or BLUPs. This is expected when BLUEs and BLUPs are correlated (Piepho et al. 2008). The correlation of BLUEs and BLUPs for our traits ranged from 0.87 (height) to 0.98 (FHB). We also estimated the accuracy of GS for each trait in a training population comprised of 886 of the lines that were tested in stage-2 or higher trials using cross-validation. The difference

of GS accuracy ranged from 0.01 (FHB) to 0.04 (height). Given that BLUEs and BLUPs produced nearly identical results we chose to use BLUPs as we considered our stage-1 lines to result from a random selection of lines from a larger population of similar germplasm.

Trait variance components

Variance components were obtained for each trait for all factors. Variation in grain yield, heading date, and test weight was largely attributed to environment (Table 3) as compared to height and FHB index where variation was largely attributed to genetics. The proportion of genetic variance to collective G×E and error variance was low for test weight (0.24) and grain yield (0.53) as compared to FHB index (0.93), height (1.13), and heading date (1.35).

Population structure of OSU breeding lines

Population structure of the 1769 genotyped lines was assessed using principal component analysis (PCA) which suggested two clusters, termed left and right clusters, relative to a PC1 score of -10 (Fig. 2). The left cluster was primarily composed of lines from the OH16 and OH17 cohorts with minimal representation from earlier cohorts (Fig. 2). The right cluster shown in Fig. 2 contained 69.9% of lines, and all OSU cohorts were represented in that cluster. The OH16 and OH17 cohorts were then each divided into two subcohorts (left (L) and right (R)). This formed the OH16L, OH16R, OH17L, and OH17R subcohorts. The mean Euclidean distance between the OH16L and OH17L subcohorts and cohorts in the right cluster was approximately double the mean distance between cohorts in the right cluster (Table 4). The F_{ST} values between cohorts suggest little differentiation among the cohorts in the right cluster or among cohorts in the left cluster (F_{ST} values between cohorts within a cluster were all < 0.06) and moderate differentiation between the right and left side groups (F_{ST} value of 0.092 to 0.125) (Table 4) (Hartl and Clark 1997).

Analysis of prediction accuracy

We assessed several GS approaches using training populations that pertain to applied breeding situations. A breeder must decide whether to phenotype all new stage-1 candidate lines in a stage-1 field trial, a portion of those lines, or to not conduct stage-1 phenotyping at all. In each scenario, GS could be used to either supplement or replace PS. When no stage-1 phenotyping is conducted, then GS predicted values substitute for stage-1 phenotypes for advancing lines to stage-2 trials. Stage-2 phenotyping will be used regardless of the scheme used to select among stage-1 lines.

We assembled TPs for each cohort that varied by the degree that they used stage-1 phenotypic data. These TPs were then used to build GS models used to estimate the GEBVs for lines within a cohort. These GEBVs were correlated to the phenotypes of the same line in the subsequent testing stage. In order to simulate PS, we correlated the phenotypes (PHEN) of the lines from one stage of testing with

Table 3 Summary of variance components for each trait from the random effects model analysis of all lines from all environments and trials

	Grain yield	Test weight	Height	Heading date	FHB index[a]
Genotype	0.233	1.63	31.0	1.98	119.74
Environment	0.697	2.40	17.1	15.60	70.70
G×E	0.145	11.96	6.7	0.55	55.86
Trial	0.146	19.14	7.3	0.70	15.33
Error	0.288	8.27	20.7	0.74	72.18

[a]FHB = index of resistance to Fusarium head blight

Table 4 Mean Euclidean distance within (on diagonal) and between cohorts (below diagonal) calculated from the first two principal component scores and the genetic differentiation between cohorts (above diagonal) using pairwise F_{ST}

	OH12	OH13	OH14	OH15	OH16L	OH16R	OH17L	OH17R
OH12	25.66	0.024	0.026	0.046	0.105	0.044	0.113	0.036
OH13	27.08	25.90	0.026	0.058	0.117	0.056	0.120	0.048
OH14	27.17	27.45	26.12	0.038	0.118	0.045	0.125	0.042
OH15	36.83	35.89	35.71	31.57	0.113	0.035	0.119	0.045
OH16L	61.39	64.42	67.43	69.34	33.36	0.092	0.058	0.111
OH16R	39.42	40.31	39.02	35.20	63.17	29.91	0.112	0.027
OH17L	68.22	68.67	74.50	72.02	48.54	73.02	32.82	0.113
OH17R	32.11	32.77	31.70	30.39	63.71	29.05	71.75	24.23

their phenotypes from the subsequent stage of testing. The correlations of predicted values with the phenotypes in the subsequent stage of testing are defined as prediction accuracies. Across all traits, the correlation of predicted value of lines in stage-2 and their phenotypes in stage-3 (stage-2 > 3) was greater than the correlation between stage-1 predicted values and their stage-2 phenotypes (stage-1 > 2) (Table 5).

For grain yield, prediction accuracy using GS (e.g., AST1, NST1, etc., methods) was slightly greater than for phenotypes (PHEN) (Table 5). The inclusion of stage-1 phenotypic data in the training population did not improve stage-1 > 2 or stage-2 > 3 GS prediction accuracy as accuracy with NST1, RST1, FST1, and AST1 methods was nearly equal to each other and to the accuracy of PHEN (Table 5). Prediction accuracy varied by cohort ("Appendix"). All prediction methods had an accuracy of ~ 0.00 for stage-1 > 2 for OH12 and OH14 cohorts except NST1, which had an accuracy of 0.17 and 0.12, respectively. Stage-1 > 2 prediction accuracy ranged from ~ 0.00 to 0.44 across prediction methods and cohorts, while stage-2 > 3 prediction accuracy ranged from ~ 0.00 to 0.53 across prediction methods and cohorts.

Test weight data were only available to estimate stage-2 > 3 prediction accuracies (Table 5). The OH15 stage-2 > 3 prediction accuracy was low, ranging from − 0.17 to − 0.14 across prediction methods ("Appendix"). Excluding the OH15 stage-2 > 3 prediction accuracy, accuracy ranged from 0.32-0.55 for PHEN and 0.10–0.58 across GS prediction methods.

For height, PHEN had greater accuracy than all GS prediction methods for stage-1 > 2 and stage-2 > 3 (Table 5). Using all (AST1) or half (RST1, FST1), the stage-1 phenotypic data versus no stage-1 data (NST1) significantly increased GS prediction accuracy from stage-1 > 2, but not from stage-2 > 3 (Table 5). The prediction accuracy, PHEN and GS prediction methods that used stage-1 phenotypic data

Table 5 Summary of stage-1 to stage-2 (stage-1 > 2) and stage-2 to stage-3 (stage-2 > 3) mean prediction accuracy for each prediction method and trait

Trait	Prediction method[b]	Stage-1 > 2 mean prediction accuracy ± s.e. (r)[a]	Stage-2 > 3 mean prediction accuracy ± s.e. (r)[a]
Grain yield	PHEN	0.14 ± 0.08	0.30 ± 0.11
	AST1	0.17 ± 0.08	0.32 ± 0.11
	NST1	0.17 ± 0.02	0.35 ± 0.13
	NST1-1 k	0.14 ± 0.04	0.31 ± 0.10
	RST1	0.16 ± 0.07	
	FST1	0.16 ± 0.07	
Test weight	PHEN		0.30 ± 0.16
	AST1		0.22 ± 0.16
	NST1		0.20 ± 0.13
Height	PHEN	0.44 ± 0.07	0.50 ± 0.05
	AST1	0.34 ± 0.05	0.42 ± 0.04
	NST1	0.15 ± 0.06	0.39 ± 0.05
	RST1	0.34 ± 0.04	
	FST1	0.34 ± 0.04	
Heading date	PHEN	0.59 ± 0.04	0.72 ± 0.02
	AST1	0.59 ± 0.03	0.69 ± 0.03
	NST1	0.37 ± 0.07	0.70 ± 0.03
	RST1	0.54 ± 0.05	
	FST1	0.54 ± 0.04	
FHB index	PHEN	0.31 ± 0.13	0.49 ± 0.04
	AST1	0.37 ± 0.12	0.45 ± 0.06
	NST1	0.30 ± 0.02	0.42 ± 0.08
	NST1-1K	0.25 ± 0.01	0.41 ± 0.08
	RST1	0.32 ± 0.11	
	FST1	0.32 ± 0.11	
	1RPHEN	0.27 ± 0.11	
	1RST1	0.33 ± 0.11	

[a]Means are averaged over the individual analyses for each of the six cohorts for Stage-1 > 2 and five cohorts for Stage-2 > 3

[b]Summary of prediction methods is provided in Table 2

ranged from 0.28 to 0.57 for stage-1 > 2, while NST1 predictions ranged from 0.04 to 0.28 ("Appendix"). Stage-2 > 3 prediction accuracy ranged from 0.28 to 0.64 across prediction methods.

Heading date had the highest prediction accuracy of any trait. NST1 had lower stage-1 > 2 accuracy than other prediction methods but was equivalent to other prediction methods from stage-2 > 3 (Table 5). Similar to height, using all or half the phenotypic data from the stage-1 trials increased GS stage-1 > 2 prediction accuracy compared to NST1 (Table 5). For stage-1 > 2 prediction accuracy, PHEN and GS prediction methods using stage-1 phenotypic data ranged from 0.43 to 0.68, while NST1 ranged from 0.19 to 0.57. Stage-2 > 3 prediction accuracy ranged from 0.56 to 0.77 across all prediction methods ("Appendix").

All prediction methods provided nearly equal accuracy for FHB index (Table 5). For the OH14 cohort, stage-1 > 2 prediction accuracy for PHEN and AST1 was ~ 0.00, while NST1 accuracy was 0.28. Excluding the results of the OH14 cohort, stage-1 > 2 accuracy ranged from 0.15 to 0.58 across all prediction methods. Stage-2 > 3 accuracy ranged from 0.34 to 0.60 across all prediction methods.

Across all traits and cohorts, the accuracy of NST1 appeared superior to PHEN when the accuracy of PHEN was low (Fig. 4). Prediction accuracy was estimated for grain yield and FHB index for 17 stage-1 > 2 and stage-2 > 3 comparisons. For six of these, the prediction accuracy of PHEN was < 0.2 with an average prediction accuracy of 0.039, while the average prediction accuracy of NST1 in these six cases was 0.165 (Fig. 4).

A low-density marker set of 1123 markers (versus 12,037 markers) was tested with the NST1 prediction method for grain yield and FHB index. This prediction method was termed NST1-1 K. The NST1-1 K methods resulted in slightly lower stage-1 > 2 and stage-2 > 3 prediction accuracy for grain yield and FHB index compared to NST1 which used all markers SNPs (Table 5).

Selection coincidence analysis

The correlation of stage-1 AST1 and PHEN predictions exceeded 0.81 in all three cohorts, while NST1 predictions were not highly correlated with PHEN in the OH14 ($r = 0.12$), OH15 ($r = 0.25$), or OH16 ($r = 0.29$) cohorts. This suggests that NST1 would select different lines than AST1 or PHEN.

We observed the percentage of stage-1 lines ranked as the best 5, 10, 15, or 20% based on stage-1 predicted values (PHEN, AST1, NST1) that were also ranked in the same percentile based on stage-2 phenotypes. This was done for grain yield and FHB for three cohorts (OH14, OH15, and OH16). None of the stage-1 GS or PS prediction rankings were very good at selecting a high percentage

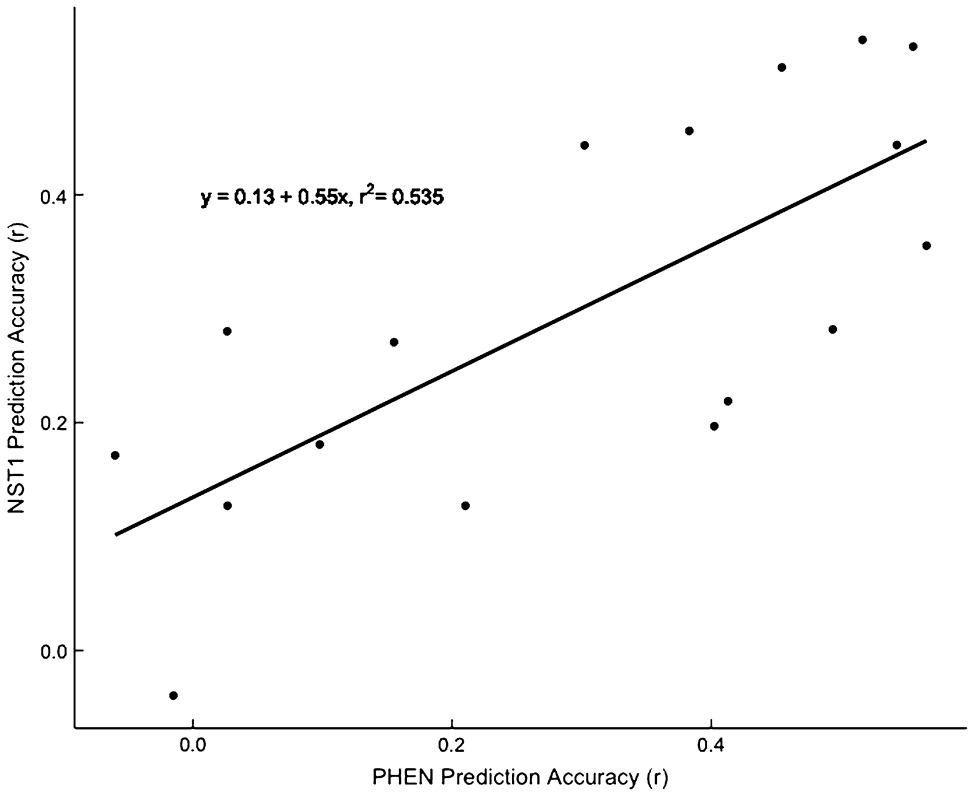

Fig. 4 Regression of genomic selection prediction accuracy using no stage-1 data (NST1) versus accuracy using only phenotypic (PHEN) for grain yield and Fusarium head blight (FHB) index for all stage-1 > 2 and stage-2 > 3 analyses

of lines that were in the same percentile in the stage-2 trials (Table 6). For example, on average, only 29.5% of the lines ranked in the best 20% for yield based on stage-1 phenotypes (PHEN) were also in the top 20% based on stage-2 phenotypes. The percentage of lines ranked in the same percentile increased as selection pressure decreased from 5 to 20%. On average, ranking stage-1 lines based on PHEN and AST1 gave similar results for yield and both were superior to NST1. For FHB, a greater percentage of lines were similarly ranked between stage-1 and stage-2 when using AST1 stage-1 predictions than when using PHEN stage-1 prediction: AST1 and PHEN predictions were both superior to using NST1 predictions.

We compared the mean stage-2 phenotype of lines selected as the best 5, 10, 15, and 20% of stage-1 lines based on predicted values (PHEN, AST1, NST1). The mean of the selected lines was expressed as a percentage of the mean stage-2 phenotype for that cohort (Table 7). For example, on average, lines selected in the best 5% of stage-1 lines based on PHEN yielded 6.6% above the mean in the stage-2 trials. In contrast, on average, lines selected in the best 5% of stage-1 lines based on NST1 yielded 3.0% above the mean in the stage-2 trials. On average, using PHEN or AST1 predictions produced similar stage-2 yields while NST1 was slightly inferior, though just one (OH16 yield) of the phenotypic differences between NST1 and PHEN selected lines were significant at $P < 0.05$ (Table 7). Low values are desired for FHB index. For FHB, lines ranked by stage-1 AST1 predictions had lower FHB index in stage-2 trials than lines ranked by PHEN, and both methods were superior than NST1, though just one of those difference (OH15) was significant.

Cost analysis and gain from indirect selection

The costs to evaluate 1000 stage-1 candidate lines using different selection schemes were compared. The PHEN selection scheme represents the current phenotypic selection scheme used by the OSU winter wheat breeding program and was considered as the base cost: it consists of one plot for grain yield and two plots for FHB index. The cost to phenotype one yield plot and one FHB plot was estimated to be $12.50 and $5.50, respectively. The costs included labor, equipment depreciation, and supplies for seed preparation, seed packaging, arranging packets for planting, planting, plot maintenance, data collection, harvesting, travel, data processing, and data analysis. Genotyping costs were estimated to be $8.00 per line assuming the use of an amplicon-based marker platform (Buckler et al. 2016) that would provide data on ~ 1200 SNPs. The comparison of the NST1 and the NST1-1K prediction accuracy (Table 5) showed that a low-cost assay of ~ 1200 markers was statistically equivalent to the accuracy obtained using 12,000 markers. The genotyping cost includes expenses (labor, supplies, space rental, shipping) to grow plants in a greenhouse, sample tissue, DNA isolation, genotyping, and data analysis. The current phenotyping program of 1000 stage-1 lines would cost $23,500 per season. Obtaining genotype data for the AST1

Table 6 Percentage of stage-1 lines selected in the best 5, 10, 15, and 20% based on one of three prediction methods (PHEN, AST1, NST1) that were in the same percentile when ranked by stage-2 phenotypic data

Cohort	Selected %	Yield			FHB index		
		PHEN[a]	AST1	NST1	PHEN	AST1	NST1
OH14	Top 5%	7.7	7.7	0.0	7.7	15.4	0.0
OH14	Top 10%	19.2	15.4	7.7	3.8	11.5	15.4
OH14	Top 15%	18.4	15.8	13.2	10.5	21.1	21.1
OH14	Top 20%	23.1	13.5	19.2	15.4	17.3	26.9
OH15	Top 5%	23.1	23.1	0.0	0.0	23.1	0.0
OH15	Top 10%	23.1	23.1	15.4	19.2	34.6	11.5
OH15	Top 15%	34.2	34.2	23.7	28.9	39.5	26.3
OH15	Top 20%	30.8	36.5	28.8	34.6	48.1	26.9
OH16	Top 5%	7.7	15.4	7.7	7.7	15.4	0.0
OH16	Top 10%	30.8	23.1	23.1	30.8	34.6	11.5
OH16	Top 15%	28.9	26.3	15.8	39.5	44.7	21.1
OH16	Top 20%	34.6	23.1	21.2	46.2	48.1	25.0
AVG	Top 5%	12.8	15.4	2.6	5.1	17.9	0.0
AVG	Top 10%	24.4	20.5	15.4	17.9	26.9	12.8
AVG	Top 15%	27.2	25.4	17.5	26.3	35.1	22.8
AVG	Top 20%	29.5	24.4	23.1	32.1	37.8	26.3

The analysis was done by cohort for grain yield and Fusarium head blight (FHB) index

[a]Summary of prediction methods is provided in Table 2

Table 7 Mean stage-2 phenotype of lines, expressed as a percent of the mean of all stage-2 lines from that cohort, selected in the best 5, 10, 15, or 20% of stage-1 lines as ranked by one of three prediction methods (PHEN, AST1, NST1)

Cohort	Selected %	Yield			FHB		
		PHEN[a]	AST1	NST1	PHEN	AST1	NST1
OH14	5	100.6	101.2	101.7	96.5	85.4	88.1
OH14	10	101.2	100.6	101.1	101.7	87.9	85.1
OH14	15	100.0	100.6	101.8	99.0	87.6	86.1
OH14	20	99.7	100.6	102.0	93.6	89.1	85.9
OH15	5	109.8	107.5	102.2	47.3	35.5	82.8
OH15	10	106.7	104.8	104.5	51.5	45.0	77.5*
OH15	15	104.2	104.6	103.3	65.7	53.8	82.8
OH15	20	103.0	104.5	103.1	62.7	59.8	82.8
OH16	5	109.5	110.1	105.1	57.6	53.4	69.8
OH16	10	108.1	107.4	103.2	62.9	53.7	73.5
OH16	15	107.8	105.4	102.6*	60.6	55.0	76.7
OH16	20	105.8	104.8	101.7	64.9	59.6	83.0
Average	5	106.6	106.3	103.0	67.2	58.1	80.3
Average	10	105.3	104.2	102.9	72.0	62.2	78.7
Average	15	104.0	103.5	102.6	75.1	65.5	81.9
Average	20	102.8	103.3	102.3	73.7	69.5	83.9

The analysis was done by cohort for grain yield and Fusarium head blight (FHB) index. Note that low values are desired for FHB

*Indicates a mean that is significantly ($P<0.05$) different than the value in the PHEN column for that trait and row based on a *t*test

[a]Summary of prediction methods is provided in Table 2

prediction-based phenotyping and genotyping adds $8000 to the PHEN cost for a total cost of $31,500 for AST1. NSTI requires only genotyping and would cost $8000 per season. The accuracy results presented in Table 5 show no advantage of AST1 over PHEN or NST1. No further analysis of AST1 was performed as AST1 costs more than PHEN and offers no accuracy advantage.

We compared the value of NST1 and PHEN in the theoretical framework of indirect selection ($G_{y.x}$) where we use trait X (NST1 or PHEN predictions) to improve trait *Y* (trait value in stage-2):

$$G_{y.x} = r_g k_x h_x \sigma_{ay}$$

We assumed equal funding for both schemes (e.g., $23,500 per season) and that σ_{ay} is a constant between the PHEN and NST1 selection schemes. The NST1 scheme requires no stage-1 phenotyping, so all the $23,500 from the PHEN scheme is used to genotype a total of 2938 lines that are candidates for stage-2 testing. We assumed that 250 lines would be advanced to stage-2 trials using PHEN or NST1. For NST1, advancing 250 of 2938 lines to stage-2 results in a selection intensity of $k_x = 1.827$, while advancing 250 of 1000 lines using PHEN produces a selection intensity of $k_x = 1.271$ (Table 8). For grain yield, the genetic correlation of PHEN predictions with stage-2 phenotypes was lower than the correlation of NST1 predictions with stage-2 phenotype (Table 8). The opposite was observed for FHB. Estimates of heritability of PHEN and NST1 predictors were nearly equal for grain yield and for FHB index. Given these estimates, $G_{x.y}$ was 2.23 times greater for grain yield using NST1 than PHEN (Table 8). The genetic correlations for PHEN and NST1 were not significantly different for grain yield or FHB index. If we assume that r_g is equal (0.21) for PHEN and NST1, then $G_{x.y}$ for grain yield using NST1 is 1.38 times greater than using PHEN. For FHB index, $G_{x.y}$ for NST1 is just 84% of the gain we would predict for PHEN, though it is 1.44 times greater for NST1 than PHEN if we assume equal r_g (0.43) for both methods, as suggested by their standard errors.

Discussion

We found that stage-1 > 2 and stage-2 > 3 prediction accuracy with GS using training populations with no stage-1 phenotypic data (NST1) was equivalent to selection based solely on phenotypes (PHEN), or to GS training populations that use stage-1 phenotypic data (AST1, RST1, FST1) for grain yield and FHB index (Table 5). NST1 predictions produced lower accuracy than other predictions for height and heading date, though this appears to be due to low variance of NST1 GEBVs compared to the variance of PHEN values (data not shown). The correlation of NST1 and PHEN

Table 8 Comparison of gain in stage-2 grain yield or FHB index from indirect selection ($G_{y.x} = r_g k_x h_x \sigma_{ay}$) based on using either genomic selection with no stage-1 phenotypic data (NST1) or using stage-1 phenotypic data only (PHEN)

Trait	Selection method[c]	Trial size[b]	% Selected	h_x	Selection intensity[a] (k_x)	Genetic correlation (r_g)	$G_{x.y}$ relative to PHEN
Grain yield	PHEN	1000	25.0	0.76	1.271	0.13 ± 0.22	1
	NST1	2938	8.5	0.73	1.827	0.21 ± 0.04	2.23
FHB index	PHEN	1000	25.0	0.81	1.271	0.43 ± 0.27	1
	NST1	2938	8.5	0.81	1.827	0.26 ± 0.05	0.87

[a]Selection intensity was calculated assuming 250 lines are advanced from stage-1 to stage-2

[b]Trial size was calculated assuming a budget of $23,500

[c]Summary of prediction methods is provi±ded in Table 2

stage-1 predictions was generally low, and lines selected by PHEN had slightly better performance in stage-2 trials than selections based on NST1 (Tables 6, 7). Based solely on performance of lines in stage-2 trials, we would conclude that AST1 is the superior selection method due to its slight advantages for selecting for FHB resistance (Tables 6, 7). The AST1 methods though are the most expensive due to the cost of both phenotyping and genotyping.

The accuracy of GS and PS varied considerable by cohort though when the accuracy of PS was low, the accuracy of GS was also low (Fig. 4). This suggests that the value of GS relative to PS may not vary greatly over seasons. This differs from the results of Belamkar et al. (2018) who found that one method was considerable superior to the other depending on the season. The estimated accuracies were generally low likely because they are between unique seasons and not from cross-validation. Higher accuracies would obviously be desired, yet breeders sometimes must deal with instances when data from one season poorly predict performance in the future. Of course, this situation is not known when making current selections and is a major issue restricting genetic gain for many quantitative traits. A breeder must use past experience to decide which approach is most likely to be better and use that approach.

One advantage of GS is that it can cost less than phenotyping and the NST1 method was the least expensive selection scheme. The low cost of the NST1 selection scheme translated into the ability to evaluate nearly three times more lines and to exert higher selection pressure as compared to PHEN or AST1 (Table 8). This would result in greater gains from indirect selection among stage-1 lines for NST1 than PHEN for grain yield and for FHB index under certain assumptions (Table 8). Our results are dependent on our estimated cost of phenotype and genotyping. There is a strong trends of increasing phenotyping costs and decreasing genotyping costs which further favors NST1 over PHEN.

We speculated that using phenotypic data from multiple years to predict the value of stage-1 lines, as is done with the GS methods, would be superior to using stage-1 phenotypes collected from one location and one year as is done by our PHEN method. Our data did not support this hypothesis. In contrast to our results, Michel et al. (2017) reported greater prediction accuracy for GS than for PHEN for grain yield in wheat. Our GS prediction accuracy results were similar to those of Belamkar et al. (2018) in that PHEN and GS prediction accuracy varied by cohort (which is confounded with selection season) while having similar mean predictive ability. We noted a positive trend between PHEN and GS prediction accuracy by cohort as reported by Heffner et al. (2011a). In our study, yield phenotypes in the 2016 stage-2 or stage-3 trials were poorly correlated with their respective predicted values regardless of the prediction method. In other pairs of years, the prediction accuracy of stage-1 > 2 and stage-2 > 3 was not associated. The low prediction accuracy of some cohorts is unlikely due to poor relatedness of that cohort with the applied TP as the genetic differentiation among cohorts was quite low (Table 4). Our results suggest that the genotype by environment interactions in our data set (Table 3) limits our ability to predict the value of lines in untested (e.g., future) environments, even when data from multiple years were incorporated in the GS predictions. This was also noted by Pembleton et al. (2018) in perennial ryegrass.

Marker density had minimal impact on NST1 prediction accuracy for grain yield and FHB index (Table 5). Song et al. (2017) reported that GS accuracy in wheat with 5716 or 1473 markers was nearly identical, while Poland et al. (2012) reported that GS accuracy was nearly identical with 34,749 versus 1827 markers. Abed et al. (2018) reported 2000 SNPs producing equivalent accuracy as 35,000 SNPs for barley traits. Lower marker densities can reduce genotyping cost, which in turn can allow more lines to be genotyped for selection. Identifying a highly predictive and stable subset of markers could potentially increase GS prediction accuracy further (Hoffstetter et al. 2016; Huang et al. 2018), as the 1123 SNPs used in this analysis were only selected on the basis of even genome coverage. Selecting markers based on their association with traits could be useful, though such

markers may be trait and environment specific, and thus less predictive than markers selected for genome coverage. It may prove best to use such selected markers to supplement a set of markers designed to cover the entire genome.

Using the theoretical framework of indirect selection, we found that replacing stage-1 phenotyping with GS predictions that do not use stage-1 phenotypic data provided greater genetic gain than selection using stage-1 phenotypic data (PHEN): utilizing data from other seasons in GS was as good as conducting a stage-1 trial as described here (1 season, 1 plot at one location for yield, 2 plots at one location for FHB index). Utilizing training data from other seasons requires that the lines evaluated in those seasons must be related to the current stage-1 candidates, as occurred in our data sets (Table 4, Fig. 2). The consistency of the relationships between the cohorts in this study facilitates this application of GS. Breeders will need to monitor the relationship among cohorts when deciding what lines to use in their training populations.

Our stage-1 trials were not very predictive of stage-2 yield or FHB index regardless of the prediction method. Using more replications and locations in the stage-1 trial could provide better PHEN results as shown by the stage-2 > 3 prediction accuracies. But employing more plots per line at stage-1 testing also increases the expense per line and will restrict the number of lines assessed when budgets are fixed. Investing more money in moderately predictive phenotyping of stage-1 lines does not seem wise and would further limit the number of lines that can be tested. Low-cost genotyping is needed to realize the advantages of NST1 scheme (Rajsic et al. 2016). With low-cost genotyping, the NST1 prediction scheme provides similar selection efficacy as stage-1 phenotyping along with the opportunity to evaluate many more lines and increase genetic gain because of greater selection intensity.

There are other considerations for replacing stage-1 phenotyping with NST1 selection. In our program, the seed of candidates for stage-1 testing is derived from a single one-meter row that provides a limited amount of seed. In our program, the stage-1 trial also serves as a seed increase, so a stage-2 trial can be planted from the selected stage-1 lines. We need to determine a way to increase the amount of seed harvested from the one-meter rows, so the selected lines could go directly to a stage-2 trial or modify the amount of seed needed for a stage-2 trial. The NST1 was not very predictive of heading date and height, so we will need to obtain accurate phenotypic estimates of these highly heritable traits from the one-meter row nursery. Other traits that we did not assess are also important. Cross-validation results suggest that GS accuracy for soft wheat quality traits is high (Heffner et al. 2011b; Hoffstetter et al. 2016; Huang et al. 2016, 2018), so we hope that the NST1 scheme can work for these traits. In addition, we would need to create the new cohort, isolate DNA, genotype all members, predict their value, and prepare the selected lines for planting in the stage-2 trial in about 60 days. Rapid genotyping and data processing will be essential to attain this turn-around time.

Some advanced GS models could prove beneficial to the NST1 predictions. GS models that estimate marker by environment interactions were not used in this analysis (Crossa et al. 2006; Lopez-Cruz et al. 2015) and may be of limited value as we must predict performance in future environments that have not been sampled. Burgueño et al. (2012) reported that incorporating marker by environment interactions into GS models did not improve the ability to predict the value of untested lines: only the main effect of untested lines could be predicted as was done in this study. Lopez-Cruz et al. (2015) reported similar results. Oakley et al. (2015) reported that a one-step prediction process using data from individual reps, provided greater GS accuracy that using means over replications. Using high-throughput phenotyping platforms to measure secondary traits for use in GS models is another promising approach to improve GS prediction accuracy (Rutkoski et al. 2016; Philomi et al. 2019; Sun et al. 2019). Also, schemes that use both phenotypic and genotypic data could be further investigated to enhance selection in stage-2 trials where phenotypic data are always collected. If NST1 predictions were used to make selections of stage-1 candidate lines in place of stage-1 phenotyping, genotypic data would be available to potentially aid in subsequent stage-2 selections.

In conclusion, GS is proving to be a valuable tool for improving the efficiency of the population improvement and product development phases of breeding field trials (Gaynor et al. 2016). Our results, as well as those of others, show that introducing GS into a program can have a dramatic benefit to a breeding program. But applying GS requires a thorough assessment of the impact of GS on genetic gain, cost per unit of gain, allocation of field-testing resources, and operational considerations in all phases of breeding phases.

Acknowledgements We would like to thank Cassi Sewell, Duc Hua, Brian Sugerman, and all the employees of the OSU winter wheat breeding program that assisted in collecting the data used in this analysis. Without them this work would not have been possible, and for this, we are truly grateful. We acknowledge funding from Ohio Agricultural Research and Development Center, National Institute for Food and Agriculture, and the US Wheat and Barley Scab Initiative.

Author contributions statement DB executed the analyses and wrote the manuscript, MH assisted in the data analyses and editing the manuscript, EO executed the genotyping-by-sequencing, CS initiated the study, directed the project, performed some data analyses, and edited the manuscript. All authors read and approved the final manuscript.

Funding The funding was provided by Ohio Small Grains Marketing Program, Agricultural Research Service (Grant No. 1234567).

Compliance with ethical standards

Conflict of interest The authors declare no conflicts of interest.

Appendix: Prediction accuracy by trait, stage, prediction method, and cohort

Trait	Cohort	Stages	Model	Prediction accuracy
Yield	OH12	2013 Stage-1 to 2014 Stage-2	PHEN	−0.060
Yield	OH13	2014 Stage-1 to 2015 Stage-2	PHEN	0.098
Yield	OH12	2014 Stage-2 to 2015 Stage-3	PHEN	0.383
Yield	OH14	2015 Stage-1 to 2016 Stage-2	PHEN	0.027
Yield	OH13	2015 Stage-2 to 2016 Stage-3	PHEN	−0.015
Yield	OH15	2016 Stage-1 to 2017 Stage-2	PHEN	0.413
Yield	OH14	2016 Stage-2 to 2017 Stage-3	PHEN	0.302
Yield	OH16	2017 Stage-1 to 2018 Stage-2	PHEN	0.210
Yield	OH15	2017 Stage-2 to 2018 Stage-3	PHEN	0.517
Yield	OH12	2013 Stage-1 to 2014 Stage-2	AST1	0.022
Yield	OH13	2014 Stage-1 to 2015 Stage-2	AST1	0.130
Yield	OH12	2014 Stage-2 to 2015 Stage-3	AST1	0.309
Yield	OH14	2015 Stage-1 to 2016 Stage-2	AST1	0.008
Yield	OH13	2015 Stage-2 to 2016 Stage-3	AST1	0.009
Yield	OH15	2016 Stage-1 to 2017 Stage-2	AST1	0.450
Yield	OH14	2016 Stage-2 to 2017 Stage-3	AST1	0.416
Yield	OH16	2017 Stage-1 to 2018 Stage-2	AST1	0.246
Yield	OH15	2017 Stage-2 to 2018 Stage-3	AST1	0.528
Yield	OH12	2013 Stage-1 to 2014 Stage-2	NST1	0.171
Yield	OH13	2014 Stage-1 to 2015 Stage-2	NST1	0.181
Yield	OH12	2014 Stage-2 to 2015 Stage-3	NST1	0.456
Yield	OH14	2015 Stage-1 to 2016 Stage-2	NST1	0.127
Yield	OH13	2015 Stage-2 to 2016 Stage-3	NST1	−0.039
Yield	OH15	2016 Stage-1 to 2017 Stage-2	NST1	0.219
Yield	OH14	2016 Stage-2 to 2017 Stage-3	NST1	0.444
Yield	OH16	2017 Stage-1 to 2018 Stage-2	NST1	0.127
Yield	OH15	2017 Stage-2 to 2018 Stage-3	NST1	0.536
Yield	OH12	2013 Stage-1 to 2014 Stage-2	RST1	0.053
Yield	OH13	2014 Stage-1 to 2015 Stage-2	RST1	0.108
Yield	OH14	2015 Stage-1 to 2016 Stage-2	RST1	0.010
Yield	OH15	2016 Stage-1 to 2017 Stage-2	RST1	0.360
Yield	OH16	2017 Stage-1 to 2018 Stage-2	RST1	0.272
Yield	OH12	2013 Stage-1 to 2014 Stage-2	FST1	0.053
Yield	OH13	2014 Stage-1 to 2015 Stage-2	FST1	0.108
Yield	OH14	2015 Stage-1 to 2016 Stage-2	FST1	0.015

Trait	Cohort	Stages	Model	Prediction accuracy
Yield	OH15	2016 Stage-1 to 2017 Stage-2	FST1	0.356
Yield	OH16	2017 Stage-1 to 2018 Stage-2	FST1	0.266
Yield	OH12	2013 Stage-1 to 2014 Stage-2	NST1-1K	0.267
Yield	OH13	2014 Stage-1 to 2015 Stage-2	NST1-1K	0.163
Yield	OH12	2014 Stage-2 to 2015 Stage-3	NST1-1K	0.429
Yield	OH14	2015 Stage-1 to 2016 Stage-2	NST1-1K	0.141
Yield	OH13	2015 Stage-2 to 2016 Stage-3	NST1-1K	0.026
Yield	OH15	2016 Stage-1 to 2017 Stage-2	NST1-1K	0.071
Yield	OH14	2016 Stage-2 to 2017 Stage-3	NST1-1K	0.296
Yield	OH16	2017 Stage-1 to 2018 Stage-2	NST1-1K	0.074
Yield	OH15	2017 Stage-2 to 2018 Stage-3	NST1-1K	0.490
Yield	OH12	2013 Stage-1 to 2014 Stage-2	NST1-0.5	0.147
Yield	OH13	2014 Stage-1 to 2015 Stage-2	NST1-0.5	0.039
Yield	OH12	2014 Stage-2 to 2015 Stage-3	NST1-0.5	0.418
Yield	OH14	2015 Stage-1 to 2016 Stage-2	NST1-0.5	0.104
Yield	OH13	2015 Stage-2 to 2016 Stage-3	NST1-0.5	−0.077
Yield	OH15	2016 Stage-1 to 2017 Stage-2	NST1-0.5	0.190
Yield	OH14	2016 Stage-2 to 2017 Stage-3	NST1-0.5	0.339
Yield	OH16	2017 Stage-1 to 2018 Stage-2	NST1-0.5	0.088
Yield	OH15	2017 Stage-2 to 2018 Stage-3	NST1-0.5	0.542
Yield	OH12	2013 Stage-1 to 2014 Stage-2	NST1-1.5	0.126
Yield	OH13	2014 Stage-1 to 2015 Stage-2	NST1-1.5	0.112
Yield	OH12	2014 Stage-2 to 2015 Stage-3	NST1-1.5	0.416
Yield	OH14	2015 Stage-1 to 2016 Stage-2	NST1-1.5	0.126
Yield	OH13	2015 Stage-2 to 2016 Stage-3	NST1-1.5	−0.058
Yield	OH15	2016 Stage-1 to 2017 Stage-2	NST1-1.5	0.225
Yield	OH14	2016 Stage-2 to 2017 Stage-3	NST1-1.5	0.419
Yield	OH16	2017 Stage-1 to 2018 Stage-2	NST1-1.5	0.088
Yield	OH15	2017 Stage-2 to 2018 Stage-3	NST1-1.5	0.529
Test Weight	OH12	2014 Stage-2 to 2015 Stage-3	PHEN	0.352
Test Weight	OH13	2015 Stage-2 to 2016 Stage-3	PHEN	0.477
Test Weight	OH14	2016 Stage-2 to 2017 Stage-3	PHEN	0.558
Test Weight	OH15	2017 Stage-2 to 2018 Stage-3	PHEN	−0.170
Test Weight	OH12	2014 Stage-2 to 2015 Stage-3	AST1	0.141
Test Weight	OH13	2015 Stage-2 to 2016 Stage-3	AST1	0.316
Test Weight	OH14	2016 Stage-2 to 2017 Stage-3	AST1	0.583
Test Weight	OH15	2017 Stage-2 to 2018 Stage-3	AST1	−0.178
Test Weight	OH12	2014 Stage-2 to 2015 Stage-3	NST1	0.157

Trait	Cohort	Stages	Model	Prediction accuracy
Test Weight	OH13	2015 Stage-2 to 2016 Stage-3	NST1	0.338
Test Weight	OH14	2016 Stage-2 to 2017 Stage-3	NST1	0.469
Test Weight	OH15	2017 Stage-2 to 2018 Stage-3	NST1	−0.145
Height	OH12	2013 Stage-1 to 2014 Stage-2	PHEN	0.572
Height	OH13	2014 Stage-1 to 2015 Stage-2	PHEN	0.285
Height	OH12	2014 Stage-2 to 2015 Stage-3	PHEN	0.395
Height	OH14	2015 Stage-1 to 2016 Stage-2	PHEN	
Height	OH13	2015 Stage-2 to 2016 Stage-3	PHEN	0.472
Height	OH15	2016 Stage-1 to 2017 Stage-2	PHEN	0.520
Height	OH14	2016 Stage-2 to 2017 Stage-3	PHEN	0.488
Height	OH16	2017 Stage-1 to 2018 Stage-2	PHEN	0.376
Height	OH15	2017 Stage-2 to 2018 Stage-3	PHEN	0.646
Height	OH12	2013 Stage-1 to 2014 Stage-2	AST1	0.281
Height	OH13	2014 Stage-1 to 2015 Stage-2	AST1	0.312
Height	OH12	2014 Stage-2 to 2015 Stage-3	AST1	0.321
Height	OH14	2015 Stage-1 to 2016 Stage-2	AST1	
Height	OH13	2015 Stage-2 to 2016 Stage-3	AST1	0.436
Height	OH15	2016 Stage-1 to 2017 Stage-2	AST1	0.494
Height	OH14	2016 Stage-2 to 2017 Stage-3	AST1	0.409
Height	OH16	2017 Stage-1 to 2018 Stage-2	AST1	0.283
Height	OH15	2017 Stage-2 to 2018 Stage-3	AST1	0.499
Height	OH12	2013 Stage-1 to 2014 Stage-2	NST1	0.057
Height	OH13	2014 Stage-1 to 2015 Stage-2	NST1	0.281
Height	OH12	2014 Stage-2 to 2015 Stage-3	NST1	0.353
Height	OH14	2015 Stage-1 to 2016 Stage-2	NST1	
Height	OH13	2015 Stage-2 to 2016 Stage-3	NST1	0.288
Height	OH15	2016 Stage-1 to 2017 Stage-2	NST1	0.042
Height	OH14	2016 Stage-2 to 2017 Stage-3	NST1	0.492
Height	OH16	2017 Stage-1 to 2018 Stage-2	NST1	0.209
Height	OH15	2017 Stage-2 to 2018 Stage-3	NST1	0.439
Height	OH12	2013 Stage-1 to 2014 Stage-2	RST1	0.433
Height	OH13	2014 Stage-1 to 2015 Stage-2	RST1	0.350
Height	OH14	2015 Stage-1 to 2016 Stage-2	RST1	−0.015
Height	OH15	2016 Stage-1 to 2017 Stage-2	RST1	0.353
Height	OH16	2017 Stage-1 to 2018 Stage-2	RST1	0.228
Height	OH12	2013 Stage-1 to 2014 Stage-2	FST1	0.433
Height	OH13	2014 Stage-1 to 2015 Stage-2	FST1	0.350
Height	OH14	2015 Stage-1 to 2016 Stage-2	FST1	−0.015

Trait	Cohort	Stages	Model	Prediction accuracy
Height	OH15	2016 Stage-1 to 2017 Stage-2	FST1	0.356
Height	OH16	2017 Stage-1 to 2018 Stage-2	FST1	0.225
Heading Date	OH12	2013 Stage-1 to 2014 Stage-2	PHEN	0.434
Heading Date	OH13	2014 Stage-1 to 2015 Stage-2	PHEN	0.686
Heading Date	OH12	2014 Stage-2 to 2015 Stage-3	PHEN	0.664
Heading Date	OH14	2015 Stage-1 to 2016 Stage-2	PHEN	0.588
Heading Date	OH13	2015 Stage-2 to 2016 Stage-3	PHEN	0.726
Heading Date	OH15	2016 Stage-1 to 2017 Stage-2	PHEN	0.607
Heading Date	OH14	2016 Stage-2 to 2017 Stage-3	PHEN	0.752
Heading Date	OH16	2017 Stage-1 to 2018 Stage-2	PHEN	0.619
Heading Date	OH15	2017 Stage-2 to 2018 Stage-3	PHEN	0.744
Heading Date	OH12	2013 Stage-1 to 2014 Stage-2	AST1	0.471
Heading Date	OH13	2014 Stage-1 to 2015 Stage-2	AST1	0.660
Heading Date	OH12	2014 Stage-2 to 2015 Stage-3	AST1	0.614
Heading Date	OH14	2015 Stage-1 to 2016 Stage-2	AST1	0.564
Heading Date	OH13	2015 Stage-2 to 2016 Stage-3	AST1	0.681
Heading Date	OH15	2016 Stage-1 to 2017 Stage-2	AST1	0.621
Heading Date	OH14	2016 Stage-2 to 2017 Stage-3	AST1	0.720
Heading Date	OH16	2017 Stage-1 to 2018 Stage-2	AST1	0.645
Heading Date	OH15	2017 Stage-2 to 2018 Stage-3	AST1	0.734
Heading Date	OH12	2013 Stage-1 to 2014 Stage-2	NST1	0.282
Heading Date	OH13	2014 Stage-1 to 2015 Stage-2	NST1	0.576
Heading Date	OH12	2014 Stage-2 to 2015 Stage-3	NST1	0.670
Heading Date	OH14	2015 Stage-1 to 2016 Stage-2	NST1	0.191
Heading Date	OH13	2015 Stage-2 to 2016 Stage-3	NST1	0.643
Heading Date	OH15	2016 Stage-1 to 2017 Stage-2	NST1	0.284
Heading Date	OH14	2016 Stage-2 to 2017 Stage-3	NST1	0.707
Heading Date	OH16	2017 Stage-1 to 2018 Stage-2	NST1	0.502
Heading Date	OH15	2017 Stage-2 to 2018 Stage-3	NST1	0.774
Heading Date	OH12	2013 Stage-1 to 2014 Stage-2	RST1	0.496
Heading Date	OH13	2014 Stage-1 to 2015 Stage-2	RST1	0.698
Heading Date	OH14	2015 Stage-1 to 2016 Stage-2	RST1	0.418
Heading Date	OH15	2016 Stage-1 to 2017 Stage-2	RST1	0.519
Heading Date	OH16	2017 Stage-1 to 2018 Stage-2	RST1	0.559
Heading Date	OH12	2013 Stage-1 to 2014 Stage-2	FST1	0.496
Heading Date	OH13	2014 Stage-1 to 2015 Stage-2	FST1	0.698
Heading Date	OH14	2015 Stage-1 to 2016 Stage-2	FST1	0.435
Heading Date	OH15	2016 Stage-1 to 2017 Stage-2	FST1	0.516

Trait	Cohort	Stages	Model	Prediction accuracy
Heading Date	OH16	2017 Stage-1 to 2018 Stage-2	FST1	0.573
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	PHEN	0.155
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	PHEN	
FHB Index	OH12	2014 Stage-2 to 2015 Stage-3	PHEN	0.402
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	PHEN	0.027
FHB Index	OH13	2015 Stage-2 to 2016 Stage-3	PHEN	0.543
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	PHEN	0.494
FHB Index	OH14	2016 Stage-2 to 2017 Stage-3	PHEN	0.454
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	PHEN	0.566
FHB Index	OH15	2017 Stage-2 to 2018 Stage-3	PHEN	0.556
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	AST1	0.280
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	AST1	
FHB Index	OH12	2014 Stage-2 to 2015 Stage-3	AST1	0.343
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	AST1	0.066
FHB Index	OH13	2015 Stage-2 to 2016 Stage-3	AST1	0.408
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	AST1	0.536
FHB Index	OH14	2016 Stage-2 to 2017 Stage-3	AST1	0.448
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	AST1	0.588
FHB Index	OH15	2017 Stage-2 to 2018 Stage-3	AST1	0.604
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	NST1	0.271
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	NST1	
FHB Index	OH12	2014 Stage-2 to 2015 Stage-3	NST1	0.197
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	NST1	0.280
FHB Index	OH13	2015 Stage-2 to 2016 Stage-3	NST1	0.444
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	NST1	0.282
FHB Index	OH14	2016 Stage-2 to 2017 Stage-3	NST1	0.512
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	NST1	0.356
FHB Index	OH15	2017 Stage-2 to 2018 Stage-3	NST1	0.530
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	RST1	0.230
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	RST1	0.382
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	RST1	0.065
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	RST1	0.463
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	RST1	0.525
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	FST1	0.230
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	FST1	0.382
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	FST1	0.065
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	FST1	0.444
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	FST1	0.534

Trait	Cohort	Stages	Model	Prediction accuracy
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	1RPHEN	0.169
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	1RPHEN	0.473
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	1RPHEN	0.030
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	1RPHEN	0.401
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	1RPHEN	0.497
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	1RST1	0.276
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	1RST1	0.407
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	1RST1	0.052
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	1RST1	0.442
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	1RST1	0.555
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	NST1-1K	0.234
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	NST1-1K	0.237
FHB Index	OH12	2014 Stage-2 to 2015 Stage-3	NST1-1K	0.166
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	NST1-1K	0.232
FHB Index	OH13	2015 Stage-2 to 2016 Stage-3	NST1-1K	0.491
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	NST1-1K	0.285
FHB Index	OH14	2016 Stage-2 to 2017 Stage-3	NST1-1K	0.488
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	NST1-1K	0.262
FHB Index	OH15	2017 Stage-2 to 2018 Stage-3	NST1-1K	0.480
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	NST1-0.5	0.237
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	NST1-0.5	−0.058
FHB Index	OH12	2014 Stage-2 to 2015 Stage-3	NST1-0.5	0.184
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	NST1-0.5	0.317
FHB Index	OH13	2015 Stage-2 to 2016 Stage-3	NST1-0.5	0.299
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	NST1-0.5	0.249
FHB Index	OH14	2016 Stage-2 to 2017 Stage-3	NST1-0.5	0.457
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	NST1-0.5	0.264
FHB Index	OH15	2017 Stage-2 to 2018 Stage-3	NST1-0.5	0.512
FHB Index	OH12	2013 Stage-1 to 2014 Stage-2	NST1-1.5	0.281
FHB Index	OH13	2014 Stage-1 to 2015 Stage-2	NST1-1.5	0.063
FHB Index	OH12	2014 Stage-2 to 2015 Stage-3	NST1-1.5	0.189
FHB Index	OH14	2015 Stage-1 to 2016 Stage-2	NST1-1.5	0.278
FHB Index	OH13	2015 Stage-2 to 2016 Stage-3	NST1-1.5	0.299
FHB Index	OH15	2016 Stage-1 to 2017 Stage-2	NST1-1.5	0.282
FHB Index	OH14	2016 Stage-2 to 2017 Stage-3	NST1-1.5	0.501
FHB Index	OH16	2017 Stage-1 to 2018 Stage-2	NST1-1.5	0.326
FHB Index	OH15	2017 Stage-2 to 2018 Stage-3	NST1-1.5	0.520

References

Abed A, Pérez-Rodríguez P, Crossa J, Belzile F (2018) When less can be better: how can we make genomic selection more cost-effective and accurate in barley? Theor Appl Genet 131(9):1873–1890

Appels R, International Wheat Genome Sequencing Consortium (IWGSC) et al (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:6403

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J (2015) Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). Plant Sci 242:23–36

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

Belamkar V, Guttieri MJ, Hussain W, Jarquín D, El-basyoni I, Poland J, Lorenz A, Baenziger PS (2018) Genomic selection in preliminary yield trials in a winter wheat breeding program. G3: Genes Genomes Genetics 8(8):2735–2747

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

Buckler E, Ilut D, Wang X, Kretzschmar T, Gore M, Mitchell S. (2016) rAmpSeq: Using repetitive sequences for robust genotyping. bioRxiv, 096628. https://doi.org/10.1101/096628

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

Butler D, Cullis B, Gilmour A, Gogel B (2009) ASReml-R reference manual. The State of Queensland, Department of Primary Industries and Fisheries, Brisbane, Australia

Covarrubias-Pazaran G (2016) Genome assisted prediction of quantitative traits using the R package sommer. PLoS ONE 11(6):e0156744

Crossa J, Burgueno J, Cornelius P, McLaren G, Trethowan R, Krishnamachari A (2006) Modeling genotype x environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Sci 46:1722–1733

Endelman JB (2011) Ridge regression and other Kernels for genomic selection with R package rrBLUP. Plant Genome J 4(3):250

Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. https://doi.org/10.1186/1297-9686-41-55

Gaynor RC, Gorjanc G, Bentley A, Ober E, Howell P, Jackson R, MacKay I, Hickey J (2017) A two-part strategy for using genomic selection to develop inbred lines. Crop Sci 57(5):2372–2386

Glaubitz J, Casstevens T, Lu F, Harriman J, Elshire R, Sun Q, Buckler E (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE. https://doi.org/10.1371/journal.pone.0090346

Goudet J, Jombart T (2015) hierfstat: estimation and tests of hierarchical F-statistics. R package version 0.04-22

Habier D, Fernando R, Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397

Hartl DL, Clark GC (1997) Principles of population genetics. Sinauer Associates, Sunderland

He S, Schulthess A, Mirdita V, Zhao Y, Korzun V, Bothe B, Ebermeyer E, Reif J, Jiang Y (2016) Genomic selection in a commercial winter wheat population. Theor Appl Genetics 1293:641–651

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

Heffner EL, Jannink J-L, Sorrells ME (2011a) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4(1):65

Heffner EL, Jannink J-L, Iwata H, Sorrells E, Sorrells ME (2011b) Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci 51:2597–2606. https://doi.org/10.2135/cropsci2011.05.0253

Hoffstetter A, Cabrera A, Huang M, Sneller C (2016) Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. G3: Genes Genomes Genetics 6(9):2919–2928

Huang M, Cabrera A, Hoffstetter A, Griffey C, Van Sanford D, Costa J, McKendry A, Sneller C (2016) Genomic selection for wheat traits and trait stability. Theor Appl Genetics 129(9):1697–1710

Huang M, Ward B, Griffey C, Van Sanford D, McKendry A, Brown-Guedira G, Tyagi P, Sneller C (2018) The accuracy of genomic prediction between environments and populations for soft wheat traits. Crop Sci 58(6):2274–2288

Jannink J-L, Lorenz A, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings Funct Genomics Proteom 9(2):166–177

Longin C, Mi X, Würschum T (2015) Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. Theor Appl Genetics 128(7):1297–1306

Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink J-L, Singh RP, Autrique E, de los Campos G (2015) Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3: Genes Genomes Genetics 5(4):569–582

Marulanda J, Mi X, Melchinger AE, Xu J-L, Würschum T, Longin C (2016) Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. Theor Appl Genetics 129(10):1901–1913

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

Michel S, Ametz C, Gungor H, Epure D, Grausgruber H, Löschenberger F, Buerstmayr H (2016) Genomic selection across multiple breeding cycles in applied bread wheat breeding. Theor Appl Genetics 129(6):1179–1189

Michel S, Ametz C, Gungor H, Akgöl B, Epure D, Grausgruber H, Löschenberger F, Buerstmayr H (2017) Genomic assisted selection for enhancing line breeding: merging genomic and phenotypic selection in winter wheat breeding programs with preliminary yield trials. Theor Appl Genetics 130(2):363–376

Nazarian A, Gezan S (2016) GenoMatrix: a software package for pedigree-based and genomic prediction analyses on complex traits. J Hered 107(4):372–379. https://doi.org/10.1093/jhered/esw020 **Epub 2016 Mar 29**

Ostersen T, Christensen OF, Henryon M, Neilson B, Su G, Madsen P (2011) Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genet Sel Evol. https://doi.org/10.1186/1297-9686-43-38

Pembleton L, Inch C, Baillie R, Drayton M, Thakur P, Ogaji Y, Spangenberg G, Forster JW, Daetwyler HD, Cogan N (2018) Exploitation of data from breeding programs supports rapid implementation of genomic selection for key agronomic traits in perennial ryegrass. Theor Appl Genetics 131(9):1891–1902

Philomi J, Montesinos-Lopez OA, Crossa J, Mondal S, Perez LG, Poland J, Huerta-Espino J, Crespo-Herrera L, Govindan V, Dreisigacker S, Shretha S, Perez-Rodriguez P, Espinosa FP, Singh RP (2019) Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. Theor Appl Genet 132:177–194. https://doi.org/10.1007/s00122-018-3206-3

Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161(1–2):209–228

Poland J, Brown P, Sorrells M, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7(2):32253

R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rajsic P, Weersink A, Navabi A, Peter Pauls K (2016) Economics of genomic selection: the role of prediction accuracy and relative genotyping costs. Euphytica 210(2):259–276

Rinaldo A, Bacanu S-A, Devlin B, Sonpar V, Wasserman L, Roeder K (2005) Characterization of multilocus linkage disequilibrium. Genetic Epidemiol 28(3):193–206

Rodríguez-Álvarez M, Boer M, van Eeuwijk FA, Eilers P (2018) Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spatial Stat 23:52–71

Rutkoski J, Poland J, Mondal S, Autrique E, Pérez L, 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 Genetics 6(9):2799–2808

Sallam AH, Smith KP (2016) Genomic selection performs similarly to phenotypic selection in barley. Crop Sci 56(6):2871–2881

SAS Institute (2017) Base SAS 9.4 procedures guide: Statistical procedures

Sneller C, Paul P, Guttieri MJ (2010) Characterization of resistance to Fusarium head blight in an eastern US soft red winter wheat population. Crop Sci 50(1):123–133

Song J, Carver B, Powers C, Yan L, Klapste L, El-Kassaby Y, Chen C (2017) Practical application of genomic selection in a double-haploid winter wheat breeding program. Mol Breed 37:117. https://doi.org/10.1007/x11032-017-0715-8

Sun J, Poland J, Mondal S, Crossa J, Juliana P, Singh R, Rutkoski J, Jannink J-L, Crespo-Herrera I, Velu G, Huerta-Espino H, Sorrells ME (2019) High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. Theor Appl Genetics 132(6):1705–1720

Tolhurst D, Mathews K, Smith A, Cullis B (2019) Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model. J Anim Breed Genet 136:279–300

Velazco J, Rodríguez-Álvarez M, Boer M, Jordan D, Eilers P, Malosetti M, van Eeuwijk FA (2017) Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. Theor Appl Genetics 130(7):1375–1392

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38(6):1358–1370