Theoretical and Applied Genetics (2022) 135:3177–3194
https://doi.org/10.1007/s00122-022-04178-w

ORIGINAL ARTICLE

Profiling of *Fusarium* head blight resistance QTL haplotypes through molecular markers, genotyping-by-sequencing, and machine learning

Zachary J. Winn[1] · Jeanette Lyerly[1] · Brian Ward[2] · Gina Brown-Guedira[1,3] · Richard E. Boyles[4] · Mohamed Mergoum[5] · Jerry Johnson[5] · Stephen Harrison[6] · Ali Babar[7] · Richard E. Mason[8] · Russell Sutton[9] · J. Paul Murphy[1]

Received: 14 February 2022 / Accepted: 6 July 2022 / Published online: 24 July 2022

Abstract

Key message **Marker-assisted selection is important for cultivar development. We propose a system where a training population genotyped for QTL and genome-wide markers may predict QTL haplotypes in early development germplasm.**

Abstract Breeders screen germplasm with molecular markers to identify and select individuals that have desirable haplotypes. The objective of this research was to investigate whether QTL haplotypes can be accurately predicted using SNPs derived by genotyping-by-sequencing (GBS). In the SunGrains program during 2020 (SG20) and 2021 (SG21), 1,536 and 2,352 lines submitted for GBS were genotyped with markers linked to the *Fusarium* head blight QTL: *Qfhb.nc-1A, Qfhb.vt-1B*, *Fhb1*, and *Qfhb.nc-4A*. In parallel, data were compiled from the 2011–2020 Southern Uniform Winter Wheat Scab Nursery (SUWWSN), which had been screened for the same QTL, sequenced via GBS, and phenotyped for: visual *Fusarium* severity rating (SEV), percent *Fusarium* damaged kernels (FDK), deoxynivalenol content (DON), plant height, and heading date. Three machine learning models were evaluated: random forest, k-nearest neighbors, and gradient boosting machine. Data were randomly partitioned into training–testing splits. The QTL haplotype and 100 most correlated GBS SNPs were used for training and tuning of each model. Trained machine learning models were used to predict QTL haplotypes in the testing partition of SG20, SG21, and the total SUWWSN. Mean disease ratings for the observed and predicted QTL haplotypes were compared in the SUWWSN. For all models trained using the SG20 and SG21, the observed *Fhb1* haplotype estimated group means for SEV, FDK, DON, plant height, and heading date in the SUWWSN were not significantly different from any of the predicted *Fhb1* calls. This indicated that machine learning may be utilized in breeding programs to accurately predict QTL haplotypes in earlier generations.

Communicated by Jochen Reif.

✉ Zachary J. Winn
zjwinn@ncsu.edu

[1] Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC 27695, USA

[2] Department of Horticulture and Crop Science, The Ohio State University, Wooster, OH 44691, USA

[3] Plant Science Research, USDA-ARS SEA, Raleigh, NC 27695, USA

[4] Pee Dee Research and Education Center, Clemson University, Florence, SC 29506, USA

[5] Department of Crop and Soil Sciences, University of Georgia, Athens, GA 30602, USA

[6] Department of Agronomy, Louisiana State University, Baton Rouge, LA 70803, USA

[7] Agronomy Department, University of Florida, Gainesville, FL 32611, USA

[8] Soil and Crop Sciences Department, Colorado State University, Fort Collins, CO 80523, USA

[9] AgriLife Research, Texas A&M University, College Station, TX 77843, USA

Introduction

Wheat (*Triticum aestivum* L) is a worldwide diet staple and a key player in global food security. *Fusarium* head blight (FHB) is a fungal disease caused by pathogenic *Fusarium* species; the most frequent pathogenic *Fusarium* species implicated in FHB infection in the USA is *Fusarium graminearum* (Ward et al. 2008). FHB leads to lower yield and the accumulation of mycotoxins such as deoxynivalenol (DON) (McMullen et al. 2012). Due to the adverse health effects associated with DON consumption, the Food and Drug Administration of the USA has limited the total amount of DON present in finished wheat products destined for human consumption to 1 part per million (ppm) (National Grain and Feed Association 2011). Thus, increasing FHB resistance and reducing DON accumulation in wheat are a crucial goal for wheat breeding programs.

Several strategies have been proposed for the development of FHB-resistant wheat cultivars. Phenotypic selection of resistant lines planted over several years in inoculated nurseries remains a mainstay; however, the application of marker-assisted selection (MAS) and genomic selection (GS) has brought forth new techniques for selection of resistant lines (Buerstmayr et al. 2020). In the case of FHB resistance, which is a highly polygenic trait in wheat, GS has been recommended as the method of choice, due to the low prediction accuracies achieved with only MAS (Arruda et al. 2016). However, large effect FHB resistance loci, like *Fhb1*, produce significant resistance responses, and the identification of lines which contain moderate-to-large effect loci can assist in selection (Brown-Guedira et al. 2008).

Fhb1 was one of the first large effect FHB resistance loci identified (Cuthbert et al. 2006; Waldron et al. 1999) and was a major target for introgression into locally adapted lines. Additional FHB resistance QTL have been observed in soft red winter wheat lines adapted to the Southeastern United States. Two separate resistance QTL, *QFHB.vt-1B.1* and *QFHB.vt-1B.2*, were identified and validated in the cultivar Jamestown (Carpenter et al. 2020; Wright, 2014). Moreover, *Qfhb.nc-1A* and *Qfhb.nc-4A* identified in the cultivar NC-Neuse were mapped to chromosomes 1A and 4A, respectively, and validated in separate populations (Petersen et al. 2016, 2017).

Time and cost associated with genotyping are limiting factors in genomics-assisted breeding. While next-generation sequencing platforms, like genotyping-by-sequencing (GBS), have become more affordable over time (Rhoads & Au, 2015), the volume of genotyping required in breeding programs still poses a large financial obligation. Furthermore, the use of single-marker genotyping platforms for MAS, like Kompetetive allele-specific polymerase chain reaction (KASP) markers (He et al. 2014), requires the set up and execution of a single reaction per marker, per line. In MAS, more than one marker is frequently used to identify the haplotype of a QTL region; thus, the classification of a single QTL in several hundred lines could potentially require thousands of individual reactions, which can create a bottleneck in the workflow and thus impede genetic gain.

Many breeding programs utilize GBS data in earlier yield testing generations to assess the genetic potential of later development lines. In the SunGrains small grains cooperative involving seven public universities in the Southern United States, GBS is performed annually on F_7 lines to derive whole-genome single-nucleotide polymorphism (SNP) marker data. Concurrently, KASP marker panels to detect the haplotypes for upwards of 60 QTL are annually executed by the USDA Eastern Regional Small Grains Genotyping Laboratory on selected germplasm in the F_9 generation. This leaves a potential gap of two generations where MAS QTL haplotype call data are unavailable for lines for which whole-genome SNP data are available.

The objective of this research was to evaluate the utility of several methods for producing predictive FHB resistance QTL haplotype calls for lines which have only GBS data, in order to assist in earlier-generation selection. GBS-derived SNP data and QTL haplotype call data for FHB resistance were utilized to impute KASP assay genotypes and train machine learning models to predict QTL haplotype calls using GBS data. We examined the effect of training size on cross-validated prediction accuracies and the forward validated accuracies in a population with known QTL haplotype call frequencies. Finally, we observed the FHB resistance phenotypes of lines based on predicted QTL haplotype calls versus known QTL haplotype calls using an historical dataset.

Methods

Germplasm

Data utilized in this research can be broadly categorized into two distinct sets: (1) a multiyear contemporary cultivar development set of F_7 lines from the SunGrains cooperative program and (2) an historic dataset of elite soft red winter wheat lines based on the Southern Uniform Winter Wheat Scab Nursery (SUWWSN) from the years 2011–2020 (Murphy et al. 2015, 2016, 2017, 2018, 2019, 2020; Murphy & Navarro 2010, 2011, 2012, 2013, 2014).

All F_7 generation SunGrains lines from the 2019–2020 and 2020–2021 seasons were simultaneously genotyped via GBS for genome-wide SNP markers and KASP assays currently used by the USDA Eastern Regional Small Grains

Genotyping Lab to characterize *Fhb1*, *Qfhb.nc-1A*, *Qfhb.nc-4A*, and the Jamestown haplotype associated with *QFHB.vt-1B.1* and *QFHB.vt-1B.2* (*Qfhb.vt-1B*). The 2019–2020 SunGrains panel (SG20) contained 1536 lines and the 2020–2021 SunGrains panel (SG21) included 2352 lines. Each panel was representative of the SunGrains program composition (North Carolina State University, Clemson University, The University of Georgia, Louisiana State University, The University of Arkansas, and Texas A&M University), and each panel was reflective of the total Southeastern United States soft red winter wheat germplasm in the SunGrains programs. The genome-wide GBS SNP data and QTL haplotype call data from the SG20 and SG21 panels were used for training of prediction models as well as to identify the effect of training size on prediction accuracy.

The historical dataset based on the SUWWSN, which also included genome-wide SNP markers identified by GBS and QTL calls made from a diverse panel of marker assays for the previously mentioned QTL, was used in forward validation to compare observed to predicted QTL haplotype calls. Models trained on the SG20 and SG21 datasets were used to predict QTL haplotype calls in the SUWWSN. This dataset also included phenotypes collected in 103 total environments across the Southeastern United States over eight years for 436 distinct lines from 16 variety development programs and is comprised of elite soft red winter wheat lines adapted to the Southeastern United States.

Genotyping

Genotyping methods used in this study were similar to those used in Sarinelli et al (2019). Leaf tissue at the four-leaf stage was sampled for each line in the SG20, SG21, and SUWWSN panels, and DNA was extracted using sbeadex plant maxi kits (LGC Genomics, Middlesex, UK) as directed by the manufacturer's protocol. Genotyping-by-sequencing was performed as described in Poland et al (2012). Libraries were constructed at 96 plex densities and each library was processed on an Illumina HiSeq 2500. SNP discovery using raw data was done via the Tassel-5GBSv2 pipeline version 5.2.35 (Glaubitz et al. 2014).

Reads were aligned to the RefSeq 1.0 wheat genome assembly (Appels et al. 2018) using the Burrows–Wheeler aligner (BWA) version 0.7.12 (Li & Durbin, 2009). Data were filtered by removing: taxa with 85% or more missing data, SNPs at a minor allele frequency of 5% or lower, SNPs that had any heterozygous call frequency of 10% or higher, SNPs with 20% or more missing data, SNPs with an average read depth of less than 1 or more than 100, and SNPs that did not align with the reference sequence. Imputation via Beagle 5.2 was conducted postfiltering (Browning et al. 2018; Browning & Browning 2007).

All lines in the SG20 and SG21 panels were genotyped using KASP markers for: *Qfhb.nc-1A*, *Qfhb.vt-1B, Fhb1,* and *Qfhb.nc-4A*. Marker sequences and estimated genomic locations of the KASP assays used to genotype the SG20 and SG21 population are provided (Table 1). Composite calls of either resistant (R) or susceptible (S) were recorded based on the results of the marker assays for each region. Lines which were found in the allelic state of the resistant parent for the QTL across all assayed markers were given a resistant call and individuals which were either heterozygous or homozygous susceptible for at least one marker were considered susceptible.

Historic data for the four FHB resistance QTL were compiled from the Uniform Winter Wheat Scab Nursery Marker Reports for the SUWWSN from 2011 to 2020 (Brown-Guedira 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019). It is important to note that these QTL calls were not made with the same markers as in the SG20 or SG21 panel and that they were not made with a consistent marker panel due to changes in marker systems over time. For comparability and simplicity of predictions, only resistant (R) or susceptible (S) haplotype calls were recorded for the QTL in the SUWWSN panels. All lines which received a heterozygous haplotype call, an ambiguous haplotype call, or did not receive a haplotype call for a QTL were removed from the dataset prior to use in predictions.

A separate round of imputation was performed using all available SG20 and SG21 KASP data to estimate LD of KASP markers in the region of resistance QTL. All available KASP SNP calls were converted to be compatible with the variant call format (Danecek et al. 2011) and appended to existing variant calling format files containing the unimputed GBS SNP data for each SunGrains panel within its respective year, and the SUWWSN. KASP assay genotypes were then imputed in the SUWWSN using either the SG20 or SG21 panel using Beagle v5.4. This was to compare the LD of the KASP markers in the SUWWSN to the LD in the SunGrains panels within their respective year.

Linkage disequilibrium was calculated in the regions for all QTL evaluated in the SG20, SG21, and SUWWSN datasets using the function "LD()" in the package "gaston" in R (Perdry & Dandine-Roulland 2018). Boundaries of QTL regions were delimited by using the most proximal and distal megabase pair position of markers used to haplotype regions.

Phenotypic data

Phenotypic data for the SUWWSN were compiled from published scab reports sourced from the US Wheat and Barley Scab Initiative repositories. Data collected included adjusted means for heading date, plant height, visual *Fusarium* severity rating (SEV), percent fusarium

Table 1 Name, positions, and sequences of KASP assays used for haplotyping in the present study

QTL	Chromosome	KASP Name	Physical Position (Mbp)	Primer	Primer sequence
Qfhb.nc-1A	1A	IWA3805	541.9	S	GAAGGTGACCAAGTTCATGCTAACTTTGCTGTCAACTTTGAGGA
				R	GAAGGTCGGAGTCAACGGATTCTAACTTTGCTGTCAACTTTGAGGG
				CR	TTACTGCAACTGATGGGTGCACTTTATAT
		IWA886	569.75	S	GAAGGTGACCAAGTTCATGCTGTAAGCTGCTAGGTCTTGTAGCC
				R	GAAGGTCGGAGTCAACGGATTAAGTAAGCTGCTAGGTCTTGTAGCA
				CR	TACGTGCACGGTCGATCAGTTTCTA
Qfhb.vt-1B	1B	IWB43992	336.46	S	GAAGGTGACCAAGTTCATGCTCATTACTGTCGATATGGATCTTGTGC
				R	GAAGGTCGGAGTCAACGGATTACATTACTGTCGATATGGATCTTGTGT
				CR	TGCTGCTTGAAAAGAAATGCAGGATACTT
		IWA6259	348.13	S	GAAGGTGACCAAGTTCATGCTAACAATAACAGCGCACCAGCACT
				R	GAAGGTCGGAGTCAACGGATTACAATAACAGCGCACCAGCACC
				CR	GGTGGCAATAAATCTGTGTCATTCAGTAT
		IWA7594	477.41	S	GAAGGTGACCAAGTTCATGCTACGGTGTTAGATATGTCACATACTCA
				R	GAAGGTCGGAGTCAACGGATTCGGTGTTAGATATGTCACATACTCC
				CR	GGCACTCTTGAAAGGAAGGGTGCA
Fhb1	3B	TaHRC	13.64	S	GAAGGTGACCAAGTTCATGCTTTGTCTGTTTCGCTGGGATG
				R	GAAGGTCGGAGTCAACGGATTGCTCACGTCGTGCAAATGGT
				CR	CTTCCAGTTTCTGCTGCCAT
		snp3BS-8	13.96	S	GAAGGTGACCAAGTTCATGCTCACATGCATTTGCAAGGTTGTTATCC
				R	GAAGGTCGGAGTCAACGGATTCACATGCATTTGCAAGGTTGTTATCG
				CR	CAAAGCAGCCTTAGGTCAATAGTTTGAAA
Qfhb.nc-4A	4A	IWA2900	543.87	S	GAAGGTGACCAAGTTCATGCTAGGAGGCCTGCATGCACGC
				R	GAAGGTCGGAGTCAACGGATTCAGGAGGCCTGCATGCACGT
				CR	CTTGCACAACCACACGCAGAGGAA
		IWA402	566.65	S	GAAGGTGACCAAGTTCATGCTATATCAATTAAATGCTACATCATGAACATAGT
				R	GAAGGTCGGAGTCAACGGATTATCAATTAAATGCTACATCATGAACATAGC
				CR	TTTAGGAATGGAAGGAGTATCATTCACCA
		IWA2793	575.63	S	GAAGGTGACCAAGTTCATGCTCACAATTTCCCGCTCAGCG
				R	GAAGGTCGGAGTCAACGGATTCCTCACAATTTCCCGCTCAGCA
				CR	GATCTCACCGATCACCTCATGAAGAT
		IWA482	580.79	S	GAAGGTGACCAAGTTCATGCTGATCAATTGGTTCCTGTGATATCATTC
				R	GAAGGTCGGAGTCAACGGATTATGATCAATTGGTTCCTGTGATATCATTT
				CR	TGGGACAACACATTCTTGGGCCATT

Primers are labeled as susceptible (S), resistant (R), or common reverse (CR). The estimated primer position is given in megabase pairs (Mbp) where the distance is in reference to the most distal portion of the short arm of the chromosome

damaged kernels (FDK), and concentration of DON content as measured in parts per million. Heading data were recorded in days after January 1 when heading was evident in 50% of plants within a plot. Plant height was measured in centimeters from the base of the plants in the center of a plot to the tip of spikes, excluding awns. Severity was taken as a percent of the number of spikelets symptomatic for FHB over the total number of spikelets in a subsample of spikes within a plot. FDK was measured by comparing seed samples to standards of known scabby seed percentages to assign ratings. Concentrations of DON were recorded via mass spectrometry and gas chromatography. For all data preparation and phenotyping protocols, refer to the US Wheat and Barley Scab Initiative web portal [scabusa.org].

Phenotypic data analysis—software and models

All data analysis was performed in R statistical software version 4.1.1 (R Core Team 2013). Adjusted means from SUWWSN reports were checked for assumptions of normality by visual comparison of distributions. To detect population structure in the SUWWSN, a principal component analysis (PCA) was conducted using the genome-wide GBS-derived SNP data via the "prcomp()" function from the "stats" package. Principal components (PCs) which accounted for three percent or more of the total variation were used as fixed effects in estimation of marker effects. All mixed linear models were run using the "asreml" package version 4.1.0.160 (Butler et al. 2009). Adjusted means for recorded SUWWSN traits were used with observed and predicted QTL haplotype calls in mixed linear models to estimate group means:

$$y_{ijkl} = \mu + M_i + P_j + g_k + e_l + \varepsilon_{ijkl}$$

where y is the phenotypic response, μ is the population mean, M is the fixed marker call effect, P is the fixed PC effect, g is the random genotype effect where g is independent and identically distributed across all levels, e is the random environment effect where e is independent and identically distributed across all levels, and ε is the residual error where ε is independent and identically distributed across all levels.

QTL haplotype prediction as a categorical response—models and confusion matrix coefficients

Four separate methods of predicting QTL haplotypes as categorical responses were entertained: naïve classification via the most correlated marker in the training set (NCOR), K-nearest neighbors (KNN), random forest (RF), and gradient boosting machines (GBM).

For NCOR, the topmost correlated GBS SNP to the QTL haplotype calls in the SunGrains training population was used to predict the QTL haplotype call in the SunGrains testing (cross-validation) and SUWWSN (forward validation) populations. To identify the topmost correlated GBS SNP to the QTL haplotype call, calls were turned into a numeric dummy variable and a Pearson's correlation was calculated between the numeric QTL haplotype call dummy variable and a numeric matrix of SNPs. The highest correlated marker was then used as the prediction for the QTL haplotype call by taking the allelic state associated with resistance based on the correlation coefficient and using it to denote resistance, while heterozygous calls or homozygous calls not associated with resistance were denoted as susceptible. Parameter tuning via K-fold cross-validation within the training population was not conducted for NCOR due to the arbitrary nature and lack of hyperparameters for this method.

All models following were implemented using the "caret" package in R statistical software and optimal model hyperparameters were selected by fivefold cross-validation over 1000 iterations and optimal hyperparameters were selected by highest achieved cross-validation tuning accuracy (Kuhn 2008). KNN functions by finding K individuals in a training set that are most similar to an unclassified individual in a testing set; the most frequent class among those K neighbors in the training set is used as the prediction for the class of the individual in the testing set (Belkasim et al. 1992). Here, we use the top 100 most correlated SNPs to the QTL haplotype calls in the training population as the variables which define neighbors. To avoid ties in decision making, up to 25 possible neighbors were considered for classifying, starting from one individual, and proceeding in odd numbered intervals (e.g., 1, 3, 5 … 25).

RF is a machine learning model that classifies through the random generation of decision trees. In RF models used for classification, random vectors of observations are drawn out of the training population with replacement, and N number of randomly selected classifying variables are used to split at nodes within trees. A multitude of trees are drawn using N number of random predictor variables to create splits at nodes in each tree. Once the random forest is generated from the training data, classifications are made in the testing population by assigning the most frequently predicted category observed among all trees in the forest for an individual (Breiman 2001). We used the top 100 most correlated markers to the QTL haplotyping calls in the training population as possible classifying variables and N number of random predictors used to split nodes within trees were assessed from one to 100 markers in groups of five (e.g., 1, 5, 10 … 100). The number of trees generated in the random forest was optimized by the "caret" package.

GBM, also known as stochastic gradient boosting (Friedman 2001), is similar to RF in that it draws a random forest comprised of decision trees made from random selected classifying variables; however, unlike RF, GBM uses a logistic regression-like approach to classification. The GBM algorithm first derives the log of odds from the observed classifications in the training population and calculates a probability via the logistic function. Then, the GBM algorithm calculates pseudo-residuals from the observed class probability of individuals versus the predicted probability derived from the most frequent class. An initial decision tree is then drawn using randomly selected classifier variables to a limited number of leaves.

Unlike RF, GBM places multiple observations in a single leaf. Each leaf's residuals are totaled, converted to a log of odds, scaled by a learning rate, and added back to the original log of odds calculated from the frequencies of the observed classifications. A probability is then derived from the newly calculated log of odds, and this process was repeated over N number of trees (Friedman 2001). For QTL haplotype classification, we used the top 100 most correlated markers to the QTL haplotype calls in the training population as random classifying variables. We evaluated three learning rates (0.001, 0.01, and 0.1) and only one-way interactions among variables were considered. Generated random forests contained 100–1000 decision trees proceeding in groups of 100 (e.g., 100, 200, etc.). A number of leaves per tree were scaled by training population size so that trees contained only 10, 20 or 30 leaves.

For all methods listed, confusion matrices were calculated, and confusion matrix coefficients were derived. Accuracy for all models was calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. Sensitivity, specificity, and precision were calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Kappa values were reported as unadjusted Cohen's kappa statistic (McHugh 2012). The calculation for kappa is as follows:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where κ is Cohen's kappa, $\Pr(a)$ is the probability of the observed agreement, and $\Pr(e)$ represents the expected rate of chance agreement. Kappa may be loosely interpreted as how much better a method of classification performs in comparison with randomly guessing, where a value of 1 corresponds to perfect categorization and 0 corresponds to a classification method that performs no better than random chance. In the cases of all models assessed, we define the no-information rate as the largest proportion of the observed classes in the dataset and we use this as a baseline for an informative model.

QTL haplotype prediction as a categorical response-procedure

A general visual diagram of the described procedure for non-imputation and machine learning methods is provided (Fig. 1). For each QTL assessed, only GBS-derived SNP markers located on the QTL's chromosome of origin were considered (e.g., only SNPs on chromosome 3B for *Fhb1*). All observed QTL haplotype calls in the contemporary SunGrains and historic SUWWSN data were bound with imputed marker matrices from the QTL's respective chromosome. Within the SunGrains panels, five training sizes were evaluated using 10%, 25%, 50%, 75%, and 90% of the total available data. Data were randomly subset, without replacement, into training–testing splits, and the training data observed QTL haplotype calls were used in a correlational study of all GBS SNP markers on the QTL's chromosome of origin. Only the top 100 most correlated markers were used as predictors to increase computational efficiency. Importance of predictor variables was calculated via the "varImp()" function in "caret" for KNN, RF, and GBM models, and all importance values were scaled between 0 and 100 for ease of comparison.

For cross-validation, each trained model was used to predict the QTL haplotype calls of a random subset of 150 lines drawn without replacement from the held-out test portion of the SunGrains panel in its respective year. Confusion matrices were calculated using the QTL haplotype call predictions and the observed QTL haplotype calls. For forward validation, models trained using the SunGrains data in their respective year were used to predict the SUWWSN QTL haplotype calls. Predicted and observed QTL haplotype calls were used in calculation of confusion matrices. Predicted and observed QTL haplotype calls in the SUWWSN were used in previously mentioned mixed linear models to estimate QTL haplotype call estimated group means for SEV, FDK, DON, plant height, and heading date. Due to the random subsetting without replacement portion at the beginning of this procedure, the experiment was repeated 30 times to obtain distributions and averages of calculated confusion matrix coefficients and estimated group means.

QTL haplotype calling via imputation of KASP assays

As a baseline for comparison to categorization methods, the imputation algorithm Beagle was used to impute all 11 KASP-derived markers. Resistant versus susceptible calls were made by assessing the allelic state of each KASP assay in a QTL region. If a QTL region had all diagnostic KASP markers in the allelic state of the resistant parent, it was given a resistant call; if any of the diagnostic KASP markers

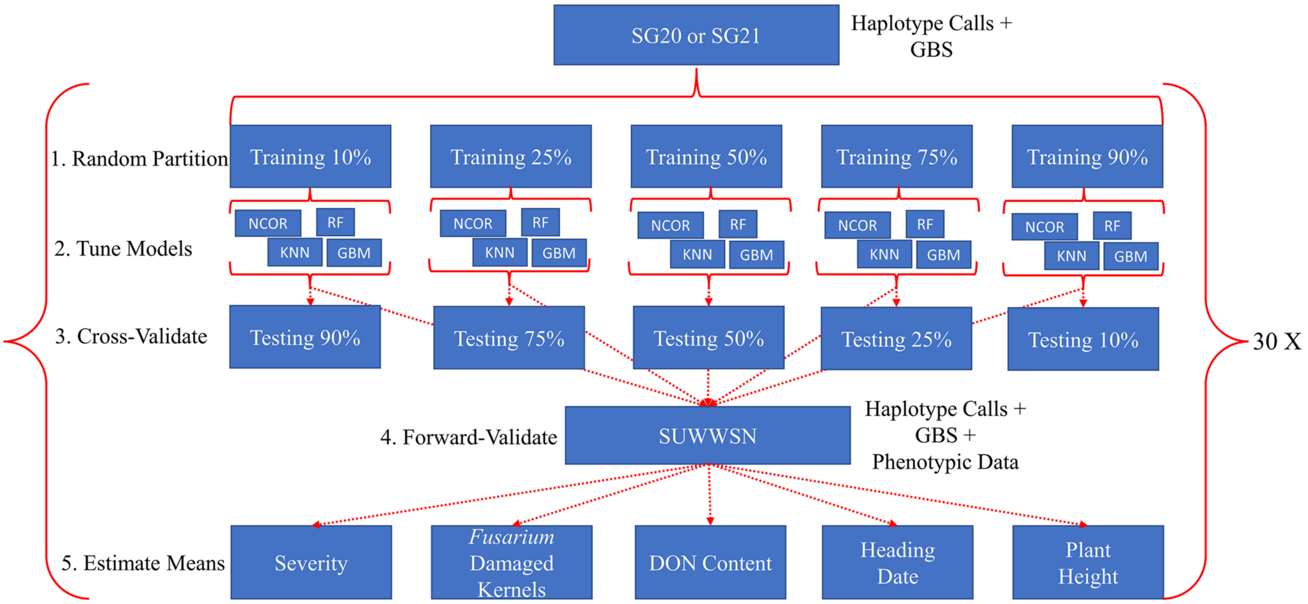


Fig. 1 Schematic of non-imputation analysis. The total analysis was conducted 30 times to obtain averages of estimates and model performance criterion. (1) The SunGrains 2020 (SG20) or SunGrains 2021 (SG21) data were randomly partitioned into different sizes. (2) The training population created in (1) was used to train and tune parameters for gradient boosting machine (GBM), k-nearest neighbor (KNN), naive classification with the most correlated marker (NCOR), and random forest (RF). (3) The trained models from (2) were used to predict the classes of the held-out testing portion of either SG20 or SG21. (4) The trained models from (2) were used to predict the QTL haplotype calls of the Southern Uniform Winter Wheat Scab Nursery (SUWWSN) and calculate confusion matrices and coefficients. (5) The QTL haplotype calls predicted in (4) were used to estimate the group means of the predicted QTL haplotype call across the available data for the SUWWSN

in the region were found not in the allelic state of the resistant parent, they were given a susceptible call. As previously stated, to facilitate imputation with Beagle, KASP SNP calls were converted to be compatible with the variant call format (Danecek et al. 2011). Each KASP SNP was assigned a genomic position approximated through the topmost result from the Basic Local Alignment Search Tool (BLAST) on the chromosome of origin in the Chinese Spring Reference genome v1.0 using the provided sequences of markers (Alaux et al. 2018; Appels et al. 2018).

Each KASP was imputed using all GBS SNPs falling within 100 Mb of its borders. Imputation and phasing were performed using Beagle v5.4 (Browning et al. 2021, 2018) without use of a reference panel, setting window size to encompass the entirety of each QTL and its 100 Mb flanking region. Beagle was supplied with a recombination distance map of GBS markers aligned to the v1.0 Chinese Spring reference genome in a population of 906 recombinant inbred lines (RILs) derived from the cross Synthetic W7984 x Opata (Gutierrez-Gonzalez et al. 2019).

To produce results that were comparable between imputation with Beagle and categorizations made with non-imputation and machine learning methods (NCOR, KNN, RF, and GBM), the following procedure was performed. For each year of the SunGrains material, KASP calls were randomly set to missing data using the same training/testing split proportions and number of replications as described for the non-imputation models. The dataset of lines with artificially missing KASP calls and lines with complete KASP call information was then integrated into the VCF file containing GBS data for that particular year. As all QTLs reside on separate chromosomes and therefore segregate independently, each genotype was uniformly assigned to the training or test set across all QTLs for each replication of cross-validation.

In each year, SUWWSN lines were also added to the VCF file. Therefore, KASP calls were imputed in the SUWWSN material, though these lines were not part of the cross-validation. This was done so that the random subset of lines would produce imputed calls for KASP markers in both the held-out test portion of the SG20 or SG21 data for cross-validation and simultaneously produce imputed KASP calls for the SUWWSN for forward validation and estimation of QTL group means. For cross-validation within the SG20 or SG21, 150 lines in the held-out testing portion were randomly selected without repeat and calls made from imputed KASP markers were used to create confusion matrices and calculate confusion matrix coefficients as previously stated. QTL calls made from imputed KASP were used to make calls in the SUWWSN and were used to calculate forward validated confusion matrix coefficients. QTL calls made from imputed KASP markers in the SUWWSN were used to estimate group means for previously mentioned traits taken in the SUWWSN.

Results

Cross-validated accuracies in the SG20 and SG21

Averaged cross-validated accuracies for *Qfhb.nc-1A* across training sizes and years ranged from 0.91 to 0.97 for Beagle, 0.86 to 0.91 for GBM, 0.87 to 0.94 for KNN, 0.77 to 0.85 for NCOR, and 0.89 to 0.95 for RF. Averaged cross-validated accuracies for *Qfhb.vt-1B* across training sizes and years ranged from 0.98 to 0.99 for Beagle, 0.97 to 0.99 for GBM, 0.96 to 0.98 for KNN, 0.92 to 0.95 for NCOR, and 0.97 to 0.99 for RF. Averaged cross-validated accuracies for *Fhb1* across years and training population sizes ranged from 0.93 to 0.97 for Beagle, 0.95 to 0.98 for GBM, 0.95 to 0.98 for KNN, 0.90 to 0.95 for NCOR, and 0.96 to 0.99 for RF. Average cross-validated accuracies for *Qfhb.nc-4A* across years and training sizes ranged from 0.90 to 0.97 for Beagle, 0.90 to 0.95 for GBM, 0.87 to 0.95 for KNN, 0.87 to 0.88 for NCOR, and 0.93 to 0.97 for RF.

All models observed, except for NCOR, had overlapping distributions for accuracies, indicating no superior method for prediction of QTL haplotype calls. Naive classification with the most correlated marker produced lower prediction accuracies and exceptionally low specificity, indicating a tendency to greatly overclassify lines as possessing the resistant form of a QTL. For all QTL assessed, exceptionally small training sizes (≈ 150 lines) produced machine learning models with lower specificity and higher sensitivity, indicating tendency to overclassify as possessing the resistant form of the QTL. Beagle tended to have higher specificity in small training sizes when compared to NCOR and machine learning models. Average kappa, accuracy, sensitivity, specificity, and precision values over 30 iterations were calculated for each year-by-QTL-by-model-by-training-size combination (Supplemental Information 1). Visualizations of SG20 and SG21 accuracy, sensitivity, and specificity value distributions are provided (Supplemental Information 2, 3).

Forward validated accuracies in the SUWWSN

Table 2 Frequency of the resistant haplotype for each QTL assessed in the present study for the SunGrains 2019–2020 (SG20), SunGrains 2020–2021 (SG21) and Southern Uniform Winter Wheat Scab Nursery (SWWSN) populations

Population	*Fhb1*	*Qfhb.vt-1B*	*Qfhb.nc-1A*	*Qfhb.nc-4A*
SG20	0.17	0.21	0.26	0.32
SG21	0.17	0.23	0.26	0.30
SUWWSN	0.14	0.38	0.72	0.41

The SG20, SG21, and SUWWSN populations had similar frequencies for the *Fhb1*-resistant haplotype (Table 2). However, for *Qfhb.nc-1A, Qfhb.vt-1B*, and *Qfhb.nc-4A*, the resistant haplotype was found at a substantially higher frequency in the SUWWSN than in either the SG20 or SG21 populations. The no-information rate in the SUWWSN for the tested QTL is as follows: 0.86 for *Fhb1*, 0.62 for *Qfhb.vt-1B*, 0.72 for *Qfhb.nc-1A*, and 0.59 for *Qfhb.nc-4A*.

In terms of forward validated prediction kappa, accuracy, sensitivity, and specificity values; results for *Fhb1* were often comparable to that of cross-validated values. For all other QTL, forward validated accuracies were modestly lower than cross-validated accuracies. Across the SG20 and SG21 years and training population sizes, averaged across all iterations, accuracies for *Qfhb.nc-1A* ranged from 0.84 to 0.86 for Beagle, 0.75 to 0.83 for GBM, 0.77 to 0.85 for KNN, 0.72 to 0.76 for NCOR, and 0.78 to 0.86 for RF. Accuracy across years and training sizes averaged over the 30 iterations for *Qfhb.vt-1B* ranged from 0.96 to 0.98 for Beagle, 0.94 to 0.97 for GBM, 0.95 to 0.98 for KNN, 0.89 to 0.90 for NCOR, and 0.95 to 0.97 for RF. Averaged accuracies for *Fhb1* over years and training sizes ranged from 0.94 to 0.97 for Beagle, 0.90 to 0.97 for GBM, 0.93 to 0.96 for KNN, 0.78 to 0.82 for NCOR, and 0.92 to 0.97 for RF. Averaged accuracies for *Qfhb.nc-4A* over years and training sizes ranged from 0.84 to 0.88 for Beagle, 0.86 to 0.87 for GBM, 0.81 to 0.89 for KNN, 0.88 to 0.89 for NCOR, and 0.86 to 0.89 for RF. Averages for kappa, accuracy, specificity, sensitivity, and precision were calculated over 30 iterations for every year-by-QTL-by-model-by-training-size combination (Supplementary Information 4).

For all QTL within year and training size, machine learning models outperformed NCOR consistently with minimal overlap for predictive accuracy and specificity, except for *Qfhb.nc-4A*. Machine learning models and imputation via Beagle underperformed in SG20 in terms of accuracy and sensitivity in comparison with NCOR for *Qfhb.nc-4A* (Fig. 2A, B). However, NCOR underperformed in terms of specificity for *Qfhb.nc-4A* in comparison with Beagle and machine learning models, which indicates that NCOR has a propensity to overclassify lines as not possessing the resistant form of the QTL. Machine learning model accuracies were most often higher than the no-information rate, except for a few cases in the smallest training sizes (≈ 150 lines); however, moderate-to-large training size models (≈ 750 to 1175 lines) were always above the no-information rate and most often produced models with comparable accuracies and specificities to imputation of single markers with Beagle.

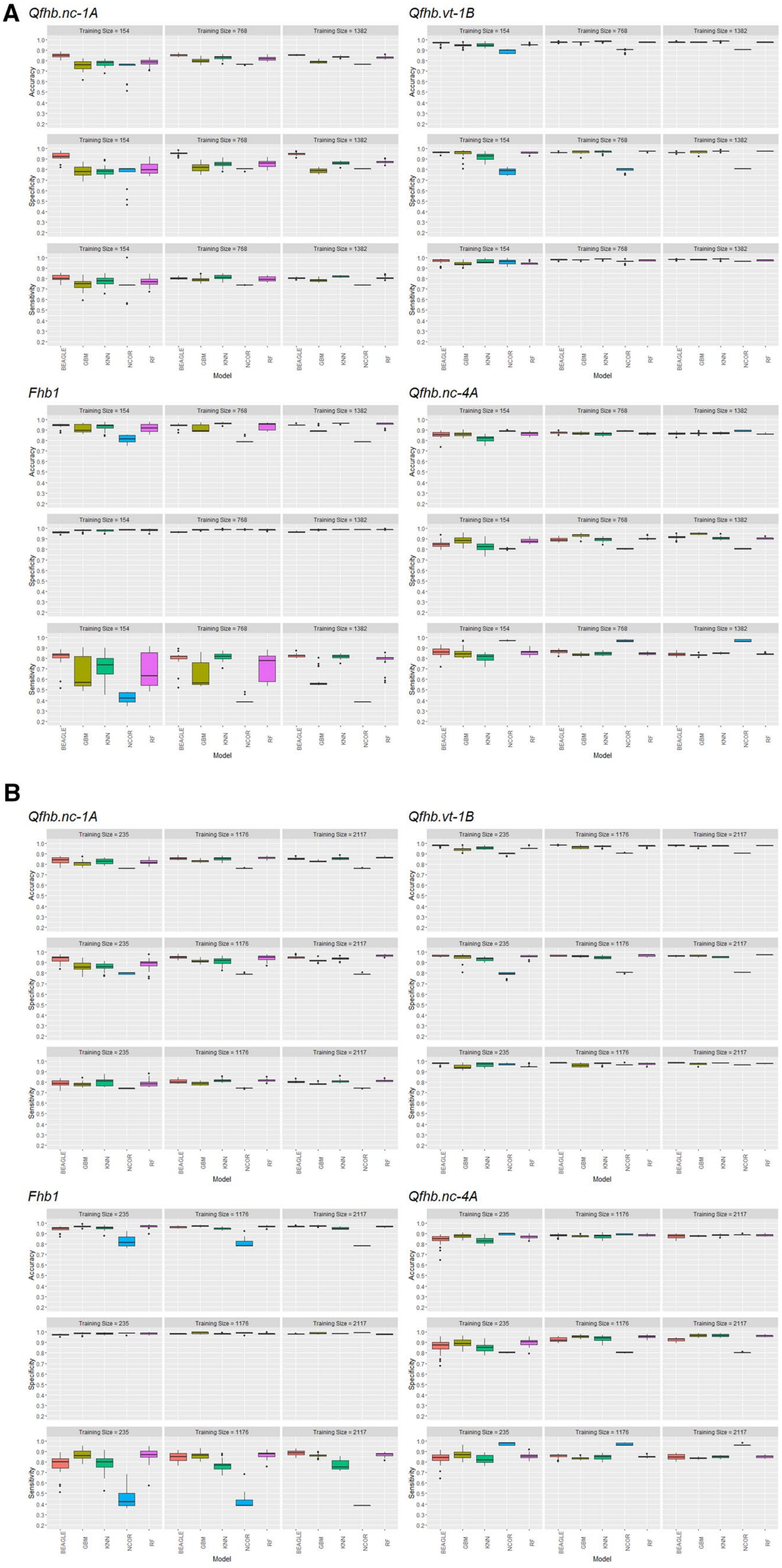
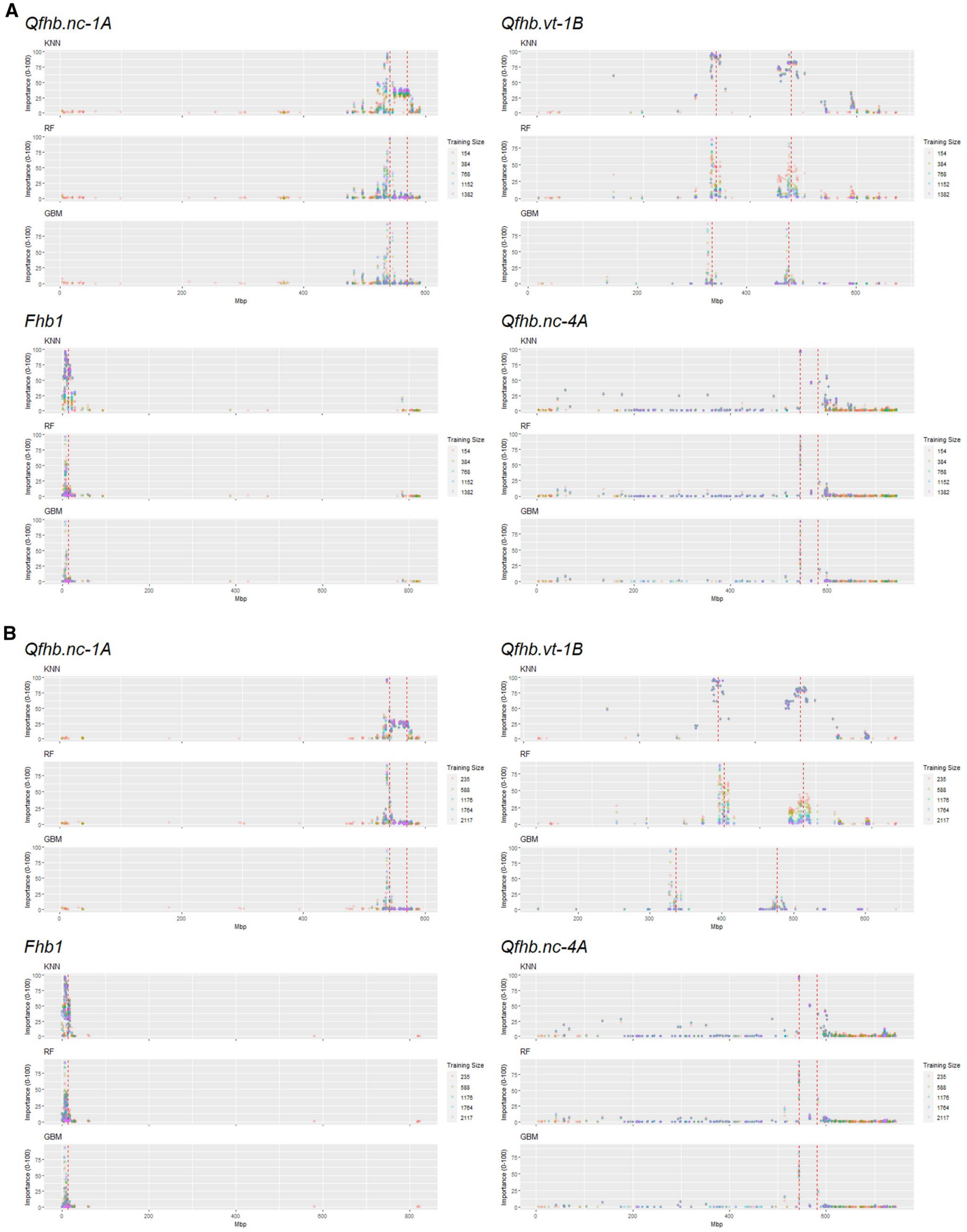


Fig. 2 **A** Accuracy, specificity, and sensitivity boxplots of the 30 iterations for 10, 50, and 90% training sizes (denoted by actual number of individuals in training) of the 2020 SunGrains trained model forward predictions made on the Uniform Southern Winter Wheat Scab Nursery for *Qfhb.nc-1A*, *Qfhb.vt-1B*, *Qfhb.nc-4A*, and *Fhb1*. Models, beagle (BEAGLE), gradient boosting machine (GBM), k-nearest neighbor (KNN), naive classification with the most correlated marker (NCOR), and random forest (RF), are denoted by color and listed on the x-axis. Training sizes are denoted by gray banners in each subgraph. The y-axis denotes the response. **B** Accuracy, specificity, and sensitivity boxplots of the 30 iterations for 10, 50, and 90% training sizes (denoted by actual number of individuals in training) of the 2021 SunGrains trained model forward predictions made on the Uniform Southern Winter Wheat Scab Nursery for *Qfhb.nc-1A*, *Qfhb.vt-1B*, *Qfhb.nc-4A*, and *Fhb1*. Models, beagle (BEAGLE), gradient boosting machine (GBM), k-nearest neighbor (KNN), naive classification with the most correlated marker (NCOR), and random forest (RF), are denoted by color and listed on the x-axis. Training sizes are denoted by gray banners in each subgraph. The y-axis denotes the response


A
Qfhb.nc-1A
Qfhb.vt-1B
Fhb1
Qfhb.nc-4A
KNN
RF
GBM
Importance (0-100)
Mbp
Training Size
154
384
768
1152
1382
B
Qfhb.nc-1A
Qfhb.vt-1B
Fhb1
Qfhb.nc-4A
KNN
RF
GBM
Importance (0-100)
Mbp
Training Size
235
588
1176
1764
2117
0
25
50
75
100
200
300
400
500
600
800

◂**Fig. 3** **A** 2020 SunGrains average GBS SNP marker importance values. Importance values are scaled between 0 and 100 for interpretability. Training size is denoted by the color of the point. Importance value is denoted by the y-axis. The x-axis denotes the position of the marker in mega base pairs (Mbp). The red vertical lines indicates the interval of KASP markers used in haplotyping. **B** 2021 SunGrains average GBS SNP marker importance values. Importance values are scaled between 0 and 100 for interpretability. Training size is denoted by the color of the point. Importance value is denoted by the y-axis. The x-axis denotes the position of the marker in mega base pairs (Mbp). The red vertical lines indicates the interval of KASP markers used in haplotyping

When comparing the use of GBS markers to categorize lines based on their QTL haplotype call versus using Beagle to impute each marker and then make a call by looking at all imputed markers within a region, it was observed that categorization by machine learning models performed similarly to Beagle when trained with moderately sized training populations (≈ 750 to 1175 lines); however, this trend was not apparent for specificity in *Qfhb.nc-1A* and sensitivity in *Fhb1*.

For specificity in *Qfhb.nc-1A,* specificities of machine learning models appeared to become equivalent to Beagle only in models trained off relatively large populations in the SG21 panel (≈ 1175 to 2100 lines). This indicated that for *Qfhb.nc-1A* models trained on smaller populations (< 1000 lines) tended to overclassify lines as resistant. For *Fhb1*, all models, including NCOR, tended to be highly specific when classifying, indicating that they have a tendency toward identifying true positive cases. However, models trained on relatively small training population sizes (< 750 lines) had generally lower sensitivities than Beagle, indicating that machine learning models trained on small populations had a tendency toward a high false negative rate in *Fhb1* (e.g., identifying resistant lines as susceptible).

In general, imputation and prediction distributions tended to overlap in overall accuracy, except for NCOR, which indicated that using the most correlated GBS marker in a training population to predict a QTL haplotype call leads to low accuracies and a tendency to overclassify or underclassify the resistant allele; therefore, this method cannot be recommended as a method to predict a QTL haplotype. However, when comparing machine learning prediction methods versus imputation via Beagle, as long as the population size was moderately large (< 1,000 lines), machine model categorical predictions tended to produce similar accuracies to conventional imputation.

Linkage disequilibrium and model-derived importance values

Boundaries were set for QTL using the most proximal and distal KASP marker positions (e.g., 540–570 megabase pairs for *Qfhb.nc-1A*) and LD was calculated for the SG20, SG21, and the SUWWSN using imputed data. Linkage disequilibrium patterns were highly similar between GBS datasets in different years; QTL regions appeared to contain similar LD patterns with KASP markers (Supplementary Information 5).

For *Qfhb.nc-1A*, marker IWA886 was in high LD ($r^2 \approx 0.90$) with a large linkage block extending from approximately 548 megabase pairs to 570 megabase pairs; the other marker used to call *Qfhb.nc-1A*, IWA3805, appears to be in minor ($r^2 < 0.25$) LD with the same block. For *Qfhb.vt-1B*, IWA7594 appears in a large linkage block from approximately 453 megabase pairs to 477 megabase pairs; IWA6259 and IWB43992 appear to be in a smaller linkage block from 336 megabase pairs to 348 megabase pairs. The two main linkage blocks for *Qfhb.vt-1B* share moderate LD ($r^2 > 0.25$), indicating that both linkage blocks may be inherited together.

For *Fhb1*, recombination appears high in the region, resulting in many markers being in small linkage blocks or not sharing LD with any markers around them. Markers snp3BS-8 and TaHRC appear to be in high LD with each other ($r^2 > 0.75$), yet only share moderate LD ($0.50 > r^2 > 0.25$) with a linkage block from 13.9 to 14.4 megabase pairs. For *Qfhb.nc-4A*, two of the four markers, IWA482 and IWA2793, are in minor LD ($r^2 < 0.25$) across the entire region for *Qfhb.nc-4A*. IWA402 lies in high LD ($r^2 > 0.75$) with a linkage block that spans from approximately 545 to 570 megabase pairs and IWA2900 lies in high LD ($r^2 > 0.75$) in a block spanning from 543 to 544 megabase pairs.

Importance values were averaged over the 30 iterations for KNN, RF, and GBM. In general, all machine learning models tended to identify SNPs within or near delimited boundaries of KASP assays for the QTL region as highly important in determining the haplotype of a line in the training population (Fig. 3A, B). As expected, the 100 most correlated markers used for training models did not remain consistent across years, training sizes, or iterations, resulting in inconsistent GBS SNP sets between each iteration of the entire procedure. Even so, KNN, RF, and GBM all identified (on average) SNPs in or near the boundaries of flanking KASP markers as highly important for the QTL haplotype call prediction in both years. If no markers were available in the region, markers in close proximity (less than one megabase pair) were indicated as highly important in predicting the QTL haplotype call.

Analysis and comparison of estimated phenotypic means for predicted versus observed QTL haplotype groups

Principle component analysis conducted on genome-wide markers indicated that the first three principal components contributed 10.7, 4.7, and 3.5 percent of the total variation, respectively. All principal components after the first three accounted for less than three percent of the total variation; therefore, only the first three principal components were used as fixed covariates in estimating QTL haplotype call significance.

Analysis of the historical observed QTL haplotypes indicated that *Fhb1*, *Qfhb.vt-1B,* and *Qfhb.nc-1A* produced a significant resistance response (Table 3). *Qfhb.nc-4A* did not produce a significant resistance response for SEV, FDK, or DON; due to *Qfhb.nc-4A*'s insignificance in producing a resistance response, it was precluded from further analysis. *Fhb1* produced a significant effect for heading date and estimated effects indicated that lines which possessed the resistant haplotype for *Fhb1* headed approximately one day later than the nonresistant haplotype.

Across years, estimated phenotypic means for predicted haplotypes tended to vary more widely when training population sizes were small (≈ 150–225 lines) and were more stable and consistent when training sizes were moderate to large (≈ 750–2,100 lines). In general, regardless of training population size, estimations of phenotypic means for predicted haplotypes tended not to vary outside one standard error of the observed QTL haplotype calls' estimated phenotypic means. This phenomenon remained consistent among QTL, traits, models, training sizes, and years. Using the most correlated marker in the training set as the predictor in the testing set led to estimations that remained very consistent between training sizes. The most correlated marker remained mostly consistent between all training sizes; this is most likely why the phenotypic means for predicted haplotypes remained relatively stable between iterations.

Averages of phenotypic means for predicted haplotypes for each QTL-by-trait-by-model-by-training-size combination were calculated (Supplemental Information 6). The average of phenotypic means for predicted haplotypes for *Fhb1, Qfhb.nc-1A*, and *Qfhb.vt-1B* tended to remain within one standard error of the observed QTL haplotype phenotypic means (Fig. 4A, B). For *Qfhb.nc-1A*, most models centered around the observed QTL estimated group means for models trained by SG20 and SG21 populations; this statement also extends to *Qfhb.vt-1B*. However, NCOR tended to produce estimations of the resistant haplotype of *Fhb1* that varied more than one standard deviation from the observed resistant group mean, indicating that using the most correlated marker in a training set to define a

Table 3 Estimated effects and significance of observed QTL haplotype calls in the Southern Uniform Winter Wheat Scab Nursery

QTL	Trait	Effect	SE	Resistant		Susceptible		Wald Statistic	*P* Value	Significance
				Mean	SE	Mean	SE			
Qfhb.nc-1A	DON	– 1.23	0.01	7.81	0.89	9.05	0.89	10.66	0.0011	**
	FDK	– 2.56	0.03	28.11	2.30	30.68	2.29	5.79	0.0161	*
	Heading Date	– 0.04	0.01	122.94	2.49	122.98	2.49	0.04	0.8417	NS
	Plant Height	0.55	0.01	34.05	0.62	33.49	0.62	3.69	0.0546	NS
	SEV	– 2.37	0.03	32.15	2.02	34.51	2.02	6.05	0.0139	*
Qfhb.vt-1B	DON	– 1.66	0.02	7.23	1.02	8.89	0.99	14.63	0.0001	***
	FDK	– 1.50	0.05	27.19	2.43	28.69	2.30	1.40	0.2366	NS
	Heading Date	0.13	0.01	120.56	2.95	120.43	2.95	0.22	0.6401	NS
	Plant Height	0.33	0.02	33.54	0.76	33.21	0.74	0.39	0.5329	NS
	SEV	– 3.20	0.05	30.26	2.43	33.47	2.34	5.46	0.0195	*
Fhb1	DON	– 2.19	0.03	7.05	0.99	9.24	0.82	16.27	0.0001	***
	FDK	– 6.72	0.06	21.89	2.34	28.61	1.99	19.26	0.0000	***
	Heading Date	0.93	0.02	123.32	2.18	122.39	2.16	7.77	0.0053	**
	Plant Height	0.07	0.02	34.12	0.67	34.05	0.61	0.12	0.7246	NS
	SEV	– 8.82	0.07	24.37	2.15	33.19	1.76	34.79	0.0000	***
Qfhb.nc-4A	DON	0.65	0.01	8.92	0.98	8.27	0.96	2.10	0.1469	NS
	FDK	0.91	0.03	29.19	2.38	28.28	2.32	0.35	0.5552	NS
	Heading Date	0.35	0.01	121.81	2.73	121.46	2.73	2.91	0.0881	NS
	Plant Height	– 0.28	0.01	33.75	0.68	34.03	0.67	1.81	0.1787	NS
	SEV	– 0.17	0.03	32.63	2.29	32.80	2.27	0.47	0.4929	NS

Traits displayed are severity (SEV), percent *Fusarium* damaged kernels (FDK), deoxynivalenol content in parts per million (DON), heading date, and plant height. Effects are in reference to inheriting the resistant allele of the listed QTL. P Values are derived from the listed Wald statistic and a Chi-square distribution. Significance is denoted as such: $p < 0.001 =$ ***, $p < 0.01 =$ **, $p < 0.05 =$ *, $p > 0.05 =$ NS

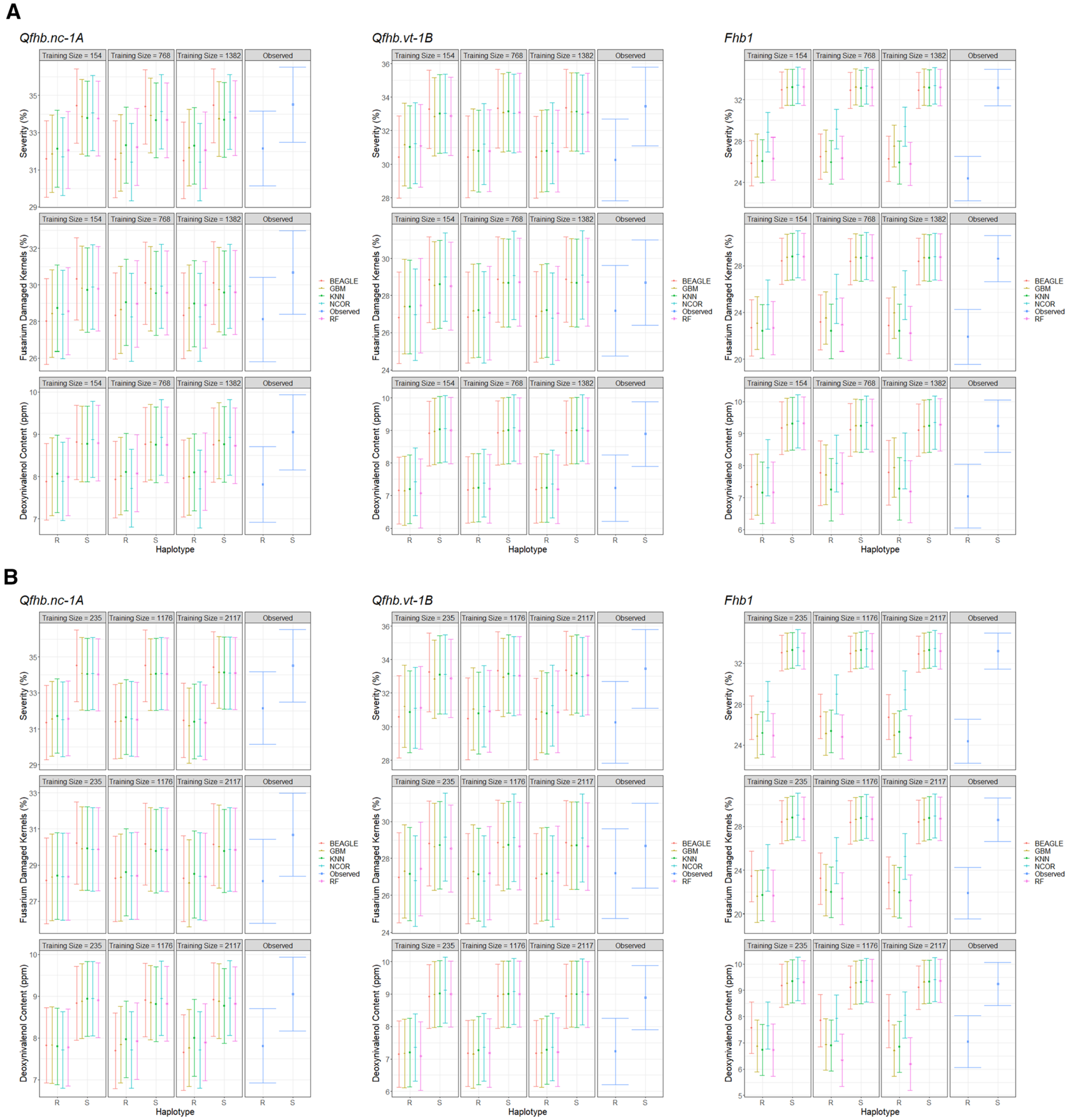


Fig. 4 **A** Estimated means of predicted QTL haplotype calls using the SG20 population versus observed QTL haplotype calls in the SUWWSN averaged over 30 iterations. Each sub-figure is labeled with the QTL to which the results displayed belong. The averaged estimated group means for severity (SEV), percent *Fusarium* damaged kernels (FDK), and deoxynivalenol content (DON) are presented and indicated on the y-axis. The x-axis denotes a haplotype call of resistant (R) or susceptible (S). Line color and point shape denote what model a prediction came from or if the QTL haplotype calls were observed. The training size of the population used to train the models is denoted above in gray banners. Bars surrounding points represent the averaged standard error about the averaged estimated group mean. **B** Estimated means of predicted QTL haplotype calls using the SG21 population versus observed QTL haplotype calls in the SUWWSN averaged over 30 iterations. Each sub-figure is labeled with the QTL to which the results displayed belong. The averaged estimated group means for severity (SEV), percent *Fusarium* damaged kernels (FDK), and deoxynivalenol content (DON) are presented and indicated on the y-axis. The x-axis denotes a haplotype call of resistant (R) or susceptible (S). Line color and point shape denote what model a prediction came from or if the QTL haplotype calls were observed. The training size of the population used to train the models is denoted above in gray banners. Bars surrounding points represent the averaged standard error about the averaged estimated group mean (gm)

QTL haplotype and then estimate the phenotypic means of that QTL tends toward an inaccurate estimate.

Discussion

Identification of lines which contain resistance QTL for FHB is key to resistant cultivar development. Moreover, identification of lines in earlier generations that contain resistance QTL allows for increased efficiency in the selection process. By identifying lines without resistance QTL of interest and removing them from the program, resources can be allocated to those individuals with a more promising QTL profile; this is highly beneficial when apportioning resources during the necessary misted/inoculated FHB screening of advanced lines. Including predictive resistance QTL haplotype calls in earlier-generation selection criterion can therefore potentially increase genetic gain by increasing the selection intensity (Moose & Mumm 2008).

In general, forward validated accuracies for QTL haplotype predictions increased as training population size increased, yet accuracies, sensitivities, and specificities were comparable among models trained by moderately sized populations. This indicated that a training population of approximately 750 to 1175 lines may adequately serve as a training population for categorization via machine learning models. Imputation rather than categorization via machine learning models appears to produce the highest accuracies in predicting QTL haplotypes when using small training population sizes (≈ 150–225 lines). However, there are some key advantages to using the machine learning methods implemented in this study.

In the present study, we approached classifying haplotypes in two different ways: imputation of KASP markers via the Beagle algorithm (which we then used to make a QTL haplotype call) and categorization of a line based off of previously made QTL haplotype calls using both naïve and machine learning classification models. While Beagle produced high specificity and sensitivity in small population sizes, it requires the following data: known individual KASP calls for each line in the training population, known physical position of each marker to a known reference genome, some estimation of recombination frequencies in the regions of the markers themselves, and a dense GBS genotype matrix. Machine learning, while requiring a larger training population, does not require any knowledge of the individual marker calls, their position, or their recombination frequency. All that is required of the machine learning methods implemented in the current study is a population with known QTL calls (a simple categorical response of resistant versus susceptible) and a dense genotypic data matrix that may or may not have known distances and physical positions.

We believe that the minimal information used in the presented machine learning methods provides a distinct benefit to the user. Firstly, not all molecular markers used in QTL haplotyping have a known position in the reference. Given that many of the markers run for different QTL are derived from mapping populations designed prior to the availability of a reference genome and alignment based SNP calling in wheat, physical positions are often not available. Additionally, the most widely used reference genome for wheat is derived from Chinese Spring (Appels et al. 2018). Some of the markers used in genotyping by the Small Grains Eastern Genotyping Lab do not have a known position in the publicly available reference genome due to the absence of the QTL in the Chinese Spring sequence or their presence on translocations derived from related species. Furthermore, markers used to profile QTL may change over time. Changes from marker platforms like simple sequence repeats (SSRs) to KASP could pose a problem when attempting to use an imputation method like Beagle in a historical dataset of lines which have been profiled via multiple different sets of markers. Additionally, the use of a method like Beagle requires the curation of many marker calls over time and across platforms.

Moreover, if marker platforms change over time, this could produce an issue with different missing data thresholds for different markers. Here, we had a complete dataset with no missing genotypic data; however, we did not evaluate what occurs when multiple markers of different missing data thresholds are imputed and used to identify QTL haplotypes. The machine learning methods presented here do not encounter these issues when using a historical dataset, which is more reflective of the datasets available to breeding programs. We therefore recommend that if only a small population of lines genotyped with a consistent marker panel for a QTL are available, the user should use an imputation method like Beagle to derive QTL haplotypes; however, if a user has access to a relatively large historical panel of diverse lines genotyped with different marker sets for a QTL over a long period of time, they can instead choose to treat this imputation issue as a categorization issue and opt to use machine learning instead.

Among the QTL assessed, *Fhb1* had the largest effect for resistance, as well as the highest and most consistent accuracies for prediction. One of the markers used for calling the *Fhb1* haplotype, TaHRC (TaHRC-Kasp-S), is taken directly from Su et al (2019) and occurs at the deletion site that confers the *Fhb1* phenotype. Furthermore, the region of the two markers used to call the *Fhb1* haplotype is noticeably short (≈ 320 kilobase pairs) in comparison with the other QTL (*Qfhb.nc-1A* = 25 megabase pairs, *Qfhb.vt-1B* = 141 megabase pairs, and *Qfhb.nc-4A* = 32 megabase pairs). While the markers that were used to genotype *Fhb1* lie in an area of short and inconsistent LD, the machine learning

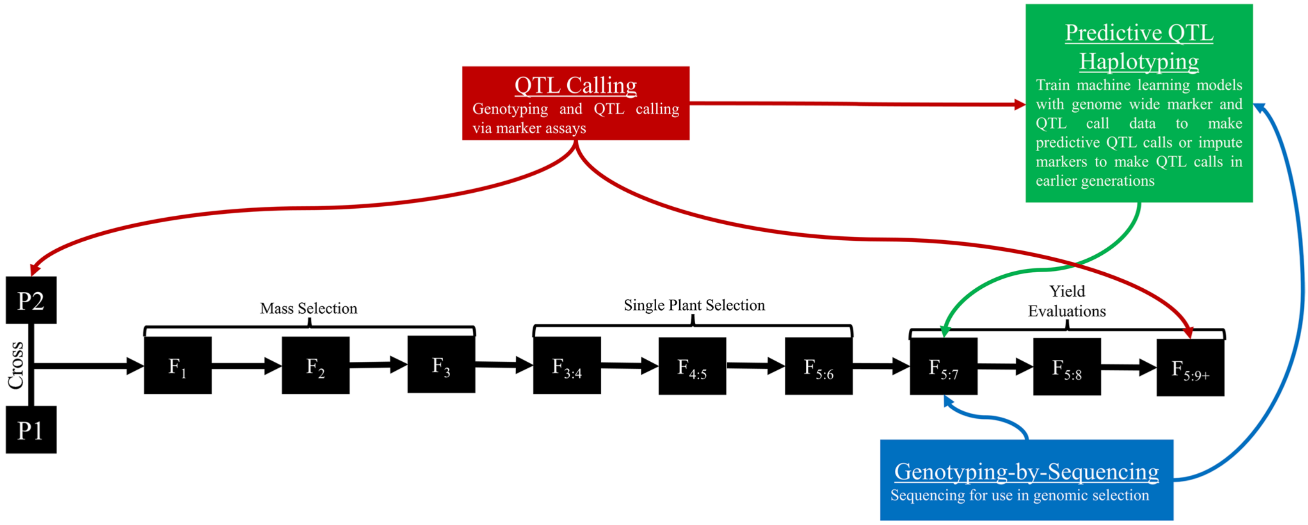

QTL Calling
Genotyping and QTL calling via marker assays
Predictive QTL Haplotyping
Train machine learning models with genome wide marker and QTL call data to make predictive QTL calls or impute markers to make QTL calls in earlier generations
P2
Cross
P1
Mass Selection
Single Plant Selection
Yield Evaluations
F_1
F_2
F_3
$F_{3:4}$
$F_{4:5}$
$F_{5:6}$
$F_{5:7}$
$F_{5:8}$
$F_{5:9+}$
Genotyping-by-Sequencing
Sequencing for use in genomic selection

Fig. 5 A hypothetical general schematic of how predictive QTL haplotyping could be incorporated into a breeding pipeline. All boxes in black and all black text near black boxes relate to the phenotypic breeding program method. Displayed is the mass–selection–pedigree method of a single cross. Red boxes and lines relate to the marker-assisted selection (MAS) pipeline where lines are genotyped using molecular markers to make a QTL haplotype call. Blue boxes and lines relate to the genotyping-by-sequencing (GBS) pipeline. Green boxes and arrows involve data from both the MAS and GBS pipeline to train machine learning models to predict QTL haplotype calls

models produced high accuracies. However, the QTL with the largest delimited region, *Qfhb.vt-1B*, often had comparable accuracies to *Fhb1*, so it appears that neither LD patterns nor physical distance between markers targeting the QTL explain the underperformance of models for *Qfhb.nc-1A* and *Qfhb.nc-4A*.

The markers developed to genotype *Qfhb.nc-1A* and *Qfhb.nc-4A* were designed from SNPs in the regions that were delineated by Peterson et al. (2017). The markers used in the current study to assess *Qfhb.nc-1A* and *Qfhb.nc-4A* were not the original KASP markers designed and evaluated by Peterson et al. (2017). This could explain the results we observed when attempting to assess the effect of the QTL haplotype calls and predicted QTL haplotype calls. Additionally, of the NC-Neuse QTL screened in the current work, it appears that only *Qfhb.nc-1A* produced a significant resistance response in the historic data of the SUWWSN. Perhaps simulation studies of predicting QTL haplotype calls for QTL with differing effect size and LD may aid in understanding the reason for machine learning model underperformance and lack of resistance effect for *Qfhb.nc-4A*. Furthermore, the historical QTL haplotype calls made in the SUWWSN were not made by a consistent marker panel and continue to change as regions are refined and marker platforms are updated. This may be an additional explanation for the lower accuracies observed for *Qfhb.nc-1A* and *Qfhb.nc-4A*.

One caveat of this study is that QTL calls were categorized as resistant or susceptible when, in practice, a heterozygous category may also be present. In cases where multiple KASP assays were considered for a given QTL, we classified individuals containing the resistant allelic state at every KASP in the QTL region as resistant; all other individuals were considered susceptible. While all the material used in this study was either F7 or later in generation, there is still the possibility for heterozygous individuals in the population. Often, we do not know the causal polymorphism in the region of a QTL. If an individual holds the susceptible allelic state for one flanking marker and the resistant state for the second flanking marker, this implies a recombination event. Since we do not have knowledge of the causal polymorphism's location, we cannot guarantee inheritance of the resistance imparted by the region; thus, we took a conservative approach to declaring resistance in these cases.

Potentially heterozygous individuals may have been segregating for the QTL when they were phenotyped in the SUWWSN, which would not be reflective of the homozygous resistant case if the QTL in question was additively inherited. Moreover, if a line were heterozygous at the time of sampling, this would indicate that future generations of the line would segregate at the QTL. For the purposes of this study and the application of these methods heterozygous individuals were classified as susceptible because we consider it a negative selection criterion. Regardless, this means that the categorization methods we have presented were not evaluated for the ability to delineate heterozygous individuals and future studies that apply this method would need to take this margin of error into account.

In the present study, we demonstrated that major FHB resistance QTL, like *Fhb1*, may be accurately and consistently predicted for lines which only have GBS data by using a dataset of just QTL haplotype calls or KASP markers and a GBS genotyped training population. In programs limited to GBS as a sequencing option, this method could be potentially beneficial; however, this method could be circumvented by using an amplicon sequencing technique which could incorporate probes for SNPs associated with known QTL of interest (Lundberg et al. 2013). If target SNPs related to QTL haplotypes were included in a targeted sequencing platform that also includes an adequate number for genome-wide loci, this could remove the KASP genotyping step from the breeding process and ameliorate this potential bottleneck in the breeding pipeline. However, this may be associated with a potential increase in genotyping cost across the entire program and will require frequent additions to the genotyping platform as new targets are identified. The approach we have described may further be used to validate new loci by predicting the presence of alleles known to affect traits for which historical phenotypes are available without additional genotyping.

Regardless, we showed that when comparing estimated group means of QTL haplotype calls made through molecular marker assays versus QTL haplotype call predictions, that QTL haplotype predictions will generally remain within one standard error of the phenotypic means, and on average, resemble the observed group means. We therefore propose the following schematic wherein predictive QTL haplotype calling is incorporated into a breeding pipeline (Fig. 5). Moreover, we hypothesize that this method may be extended to cover not only FHB resistance QTL, but the many major effect QTL for which historical QTL haplotype information and/or phenotypic data are available through the Small Grains Eastern Genotyping Lab genotyping reports.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-022-04178-w.

Author contributions statement J.L. and Z.J.W. were responsible for conceptualization and methodology; Z.J.W. contributed to software, validation, writing and preparing the original draft, and visualization; Z.J.W. and B.W. carried out the formal analysis; J.L., Z.J.W., and B.W. conducted investigation; Z.J.W., J.L., G.B.G., R.E.B., M.M., J.J., S.H., A.B., R.E.M., R.S., and J.P.M. performed data curation; Z.J.W., J.L., B.W., G.B.G., R.E.B., M.M., J.J., S.H., A.B., R.E.M., R.S., and J.P.M. wrote, reviewed, and edited the manuscript; J.P.M., G.B.G., and J.L. supervised the study; J.P.M. was involved in project administration; and J.P.M. and G.B.G. acquired the funding.

Funding The research conducted in this study was directly supported by funds from the North Carolina Small Grains Growers Association and indirectly through the United States Wheat and Barley Scab Initiative (USWBSI) over time in phenotypic evaluation of the SUWWSN and genotyping of lines in the SG20, SG21, and SUWWSN panels.

Data availability KASP and genotyping-by-sequencing data for SunGrains and genotyping-by-sequencing data for the Southern Uniform Winter Wheat Nursery lines are unavailable. Phenotypic data for the Southern Uniform Winter Wheat Scab Nursery are freely available at < scabusa.org > . Haplotype information related to resistance QTL used in the current study for the Southern Uniform Winter Wheat Nursery is freely available at < https://www.ars.usda.gov/southeast-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/ > .

Code availability All code and raw output of code used in the current study may be found at < https://github.com/zjwinn/Profiling-of-FHB-Resistance-QTL-Haplotypes-Through-MM-GBG-and-ML > .

Declarations

Conflict of interest The author claims no conflict of interest.

References

Alaux M, Rogers J, Letellier T, Flores R, Alfama F, Pommier C, International Wheat Genome Sequencing, C (2018) Linking the international wheat genome sequencing consortium bread wheat reference genome sequence to wheat genetic and phenomic data. Genome Biol 19(1): 111. https://doi.org/10.1186/s13059-018-1491-4

Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, Poland J (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361(6403):eaar7191

Arruda M, Lipka A, Brown P, Krill A, Thurber C, Brown-Guedira G, Kolb F (2016) Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (Triticum aestivum L.). Mol Breed 36(7):1–11

Belkasim S, Shridhar M, Ahmadi M (1992) Pattern classification using an efficient KNNR. Pattern Recogn 25(10):1269–1274

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Brown-Guedira G, Griffey C, Kolb F, McKendry A, Murphy J, Sanford D (2008) Breeding FHB-resistant soft winter wheat: progress and prospects. Cereal Res Commun 36(Supplement-6):31–35

Brown-Guedira G (2011) Cooperative uniform winter wheat scab nursery marker report. In: https://www.ars.usda.gov/southeast-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/cooperative-uniform-winter-wheat-scab-nurseries/: USDA-ARS

Brown-Guedira G (2012) Cooperative Uniform winter wheat scab nursery marker report. In: https://www.ars.usda.gov/southeast-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/cooperative-uniform-winter-wheat-scab-nurseries/: USDA-ARS

Brown-Guedira G (2013) Cooperative Uniform winter wheat scab nursery marker report. In: https://www.ars.usda.gov/southeast-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/cooperative-uniform-winter-wheat-scab-nurseries/: USDA-ARS

Brown-Guedira G (2014) Cooperative uniform winter wheat scab nursery marker report. In: https://www.ars.usda.gov/southeast-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/cooperative-uniform-winter-wheat-scab-nurseries/ USDA-ARS

Brown-Guedira G (2015) Cooperative uniform winter wheat scab nursery marker report. In: https://www.ars.usda.gov/south east-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/cooperative-uniform-winter-wheat-scab-nurseries/: USDA-ARS

Brown-Guedira G (2016) Cooperative uniform winter wheat scab nursery marker report. In: https://www.ars.usda.gov/southeast-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/cooperative-uniform-winter-wheat-scab-nurseries/: USDA-ARS

Brown-Guedira G (2017) Cooperative uniform winter wheat scab nursery marker report. In: https://www.ars.usda.gov/southeast-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/cooperative-uniform-winter-wheat-scab-nurseries/: USDA-ARS

Brown-Guedira G (2018) Cooperative uniform winter wheat scab nursery marker report. In: https://www.ars.usda.gov/southeast-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/cooperative-uniform-winter-wheat-scab-nurseries/: USDA-ARS

Brown-Guedira G (2019) Cooperative uniform winter wheat scab nursery marker report. In: https://www.ars.usda.gov/southeast-area/raleigh-nc/plant-science-research/docs/small-grains-genotyping-laboratory/regional-nursery-marker-reports/cooperative-uniform-winter-wheat-scab-nurseries/: USDA-ARS

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81(5):1084–1097

Browning BL, Zhou Y, Browning SR (2018) A one-penny imputed genome from next-generation reference panels. Am J Hum Genet 103(3):338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Browning BL, Tian X, Zhou Y, Browning SR (2021) Fast two-stage phasing of large-scale sequence data. Am J Hum Genet 108(10):1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Buerstmayr M, Steiner B, Buerstmayr H (2020) Breeding for Fusarium head blight resistance in wheat—progress and challenges. Plant Breed 139(3):429–454

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

Carpenter NR, Wright E, Malla S, Singh L, Van Sanford D, Clark A, Chao S (2020) Identification and validation of Fusarium head blight resistance QTL in the US soft red winter wheat cultivar 'Jamestown.' Crop Sci 60(6):2919–2930

Cuthbert PA, Somers DJ, Thomas J, Cloutier S, Brulé-Babel A (2006) Fine mapping Fhb1, a major gene controlling fusarium head blight resistance in bread wheat (Triticum aestivum L). Theor Appl Genet 112(8):1465

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Genomes Project Analysis, G (2011) The variant call format and VCFtools. Bioinformatics 27(15): 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9(2):e90346

National Grain and Feed Association (2011) FDA mycotoxin regulatory guidance. In: A guide for gain elevators, feed manufacturers, grain processors and exporters. National Grain and Feed Association, p 7

Gutierrez-Gonzalez JJ, Mascher M, Poland J, Muehlbauer GJ (2019) Dense genotyping-by-sequencing linkage maps of two synthetic W7984×Opata reference populations provide insights into wheat structural diversity. Sci Rep 9(1):1793. https://doi.org/10.1038/s41598-018-38111-3

He C, Holme J, Anthony J (2014) SNP genotyping: the KASP assay. In: Walker JM (ed) Crop breeding. School of Life Sciences, University of Hertfordshire, Hatfield, pp 75–86

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28(1):1–26

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL (2013) Practical innovations for high-throughput amplicon sequencing. Nat Methods 10(10):999–1002

McHugh ML (2012) Interrater reliability: the kappa statistic. Biochemia Medica 22(3):276–282

McMullen M, Bergstrom G, De Wolf E, Dill-Macky R, Hershman D, Shaner G, Van Sanford D (2012) A unified effort to fight an enemy of wheat and barley: Fusarium head blight. Plant Dis 96(12):1712–1728

Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiol 147(3):969–977

Murphy J, Navarro R (2010) Southern uniform winter wheat scab nursery. In: https://scabusa.org/db/documents.php: U.S. Wheat and Barley Scab Initiative

Murphy J, Navarro R (2011) Southern uniform winter wheat scab nursery. In: https://scabusa.org/pdfs_dbupload/suwwsn11_report.pdf: U.S. Wheat and Barley Scab Initiative

Murphy J, Navarro R (2012) Southern uniform winter wheat scab nursery. In: https://scabusa.org/pdfs_dbupload/suwwsn12_report.pdf: U.S. Wheat and Barley Scab Initiative

Murphy J, Navarro R (2013) Southern uniform winter wheat scab nursery. In: https://scabusa.org/pdfs_dbupload/suwwsn13_report.pdf: U.S. Wheat and Barley Scab Initiative

Murphy J, Navarro R (2014) Southern uniform winter wheat scab nursery. In: https://scabusa.org/pdfs_dbupload/suwwsn14_report.pdf: U.S. Wheat and Barley Scab Initiative

Murphy J, Lyerly J, Petersen S, Poole B (2015) Southern uniform winter wheat scab nursery. In: https://scabusa.org/pdfs_dbupload/suwwsn15_report.pdf: U.S. Wheat and Barley Scab Initiative

Murphy J, Lyerly J, Sarinelli J, Tyagi P, Brown-Guedira G (2016) Southern uniform winter wheat scab nursery. In: https://scabusa.org/pdfs_dbupload/suwwsn16_report.pdf: U.S. Wheat and Barley Scab Initiative

Murphy J, Lyerly J, Acharya R, Sarinelli J, Tyagi P, Page J, Brown-Guedira G (2017) Southern uniform winter wheat scab nursery. In: https://scabusa.org/pdfs_dbupload/suwwsn17_report.pdf: U.S. Wheat and Barley Scab Initiative

Murphy J, Lyerly J, Acharya R, Page J, Ward B, Brown-Guedira G (2018) Southern uniform winter wheat scab nursery. In. https://scabusa.org/pdfs_dbupload/suwwsn18_report.pdf: U.S. Wheat and Barley Scab Initiative

Murphy J, Lyerly J, Acharya R, Page J, Ward B, Brown-Guedira G (2019) Southern uniform winter wheat scab nursery. In: https://scabusa.org/pdfs_dbupload/suwwsn19_report.pdf: U.S. Wheat and Barley Scab Initiative

Murphy J, Lyerly J, Winn Z, Page J, Brown-Guedira G (2020) Southern uniform winter wheat scab nursery. In: https://scabusa.org/pdfs_dbupload/suwwsn20_report.pdf: U.S. Wheat and Barley Scab Initiative

Perdry H, Dandine-Roulland L (2018) Gaston—genetic data handling (QC, GRM, LD, PCA) & linear mixed Models. R Package 83:1–29

Petersen S, Lyerly JH, Maloney PV, Brown-Guedira G, Cowger C, Costa JM, Murphy JP (2016) Mapping of Fusarium head blight resistance quantitative trait loci in winter wheat cultivar NC-Neuse. Crop Sci 56(4):1473–1483

Petersen S, Lyerly JH, McKendry AL, Islam MS, Brown-Guedira G, Cowger C, Murphy JP (2017) Validation of Fusarium head blight resistance QTL in US winter wheat. Crop Sci 57(1):1–12

Poland JA, Brown PJ, Sorrells ME, 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):e32253

R Core Team (2013) R: A language and environment for statistical computing

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinform 13(5):278–289

Sarinelli JM, Murphy JP, Tyagi P, Holland JB, Johnson JW, Mergoum M, Sutton R (2019) Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. Theor Appl Genet 132(4):1247–1261

Su Z, Bernardo A, Tian B, Chen H, Wang S, Ma H, Li T (2019) A deletion mutation in TaHRC confers Fhb1 resistance to Fusarium head blight in wheat. Nat Genet 51(7):1099–1105

Waldron B, Moreno-Sevilla B, Anderson J, Stack R, Frohberg R (1999) RFLP mapping of QTL for Fusarium head blight resistance in wheat. Crop Sci 39(3):805–811

Ward TJ, Clear RM, Rooney AP, O'Donnell K, Gaba D, Patrick S, Nowicki TW (2008) An adaptive evolutionary shift in Fusarium head blight pathogen populations is driving the rapid spread of more toxigenic Fusarium graminearum in North America. Fungal Genet Biol 45(4):473–484. https://doi.org/10.1016/j.fgb.2007.10.003

Wright EE (2014) Identification of Native FHB Resistance QTL in the SRW Wheat Cultivar Jamestown, Virginia Tech

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.