Prediction of *Fusarium* Head Blight Resistance QTL Haplotypes Through Molecular Markers, Genotyping-by-Sequencing, and Machine Learning

Zachary J Winn (zwinn@outlook.com)
NCSU: North Carolina State University https://orcid.org/0000-0003-1543-1527
Jeanette Lyerly
North Carolina State University
Gina Brown-Guedira
North Carolina State University
Richard E. Boyles
Clemson University
Mohamed Mergoum
University of Georgia
Jerry Johnson
University of Georgia
Stephen Harrison
Louisiana State University
Ali Babar
University of Florida
Richard E. Mason
Colorado State University
Russell Sutton
Texas A and M University: Texas A&M University
J. Paul Murphy
North Carolina State University

Research Article

Keywords: Plant-Breeding, Genotyping, QTL, Machine-Learning, Resistance, Host-Defense

Posted Date: February 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1359831/v1

Abstract

Breeders screen germplasm with molecular markers to identify and select individuals that have desirable haplotypes. The objective of this research was to investigate if QTL haplotypes can be accurately predicted using SNPs derived by genotyping-by-sequencing (GBS). In the SunGrains program during 2020 (SG20) and 2021 (SG21), 2376 and 3423 lines submitted for GBS were genotyped for the *Fusarium* head blight QTL: *Fhb1, Qfhb.vt-1B*, *Qfhb.nc-1A*, and *Qfhb.nc-4A*. In parallel, data were compiled from the 2011-2019 Southern Uniform Winter Wheat Scab Nursery (SUWWSN), which had been screened for the same QTL, sequenced via GBS, and phenotyped for: severity (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. Observed and predicted QTL haplotypes effects were compared in the SUWWSN. For all models trained using the SG20 and SG21, the observed *Fhb1* haplotype estimated effects for SEV, FDK, DON, plant height, and heading date in the SUWWSN were not significantly different from any of the predicted *Fhb1* call effects. This indicated that machine learning may be utilized in breeding programs to accurately impute QTL haplotypes in earlier generations via a GBS and KASP genotyped training population.

Key Message

Marker assisted and genomic selection are important for cultivar development. We propose a system where a training population genotyped for QTL and genome wide markers may predict QTL haplotypes.

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 United States is *Fusarium graminearum* (Ward et al., 2008)*{Ward, 2008, An adaptive evolutionary shift in Fusarium head blight pathogen populations is driving the rapid spread of more toxigenic Fusarium graminearum in North America}{Ward, 2008, An adaptive evolutionary shift in Fusarium head blight pathogen populations is driving the rapid spread of more toxigenic Fusarium graminearum in North America}*. FHB leads to lower yield and the accumulation of mycotoxins like deoxynivalenol (DON) (McMullen et al., 2012). . Because of the adverse health effects associated with DON consumption, the Food and Drug Administration of the United States 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 is 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; Petersen et al., 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, 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 used to identify the haplotype of a region, thus the identification 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.

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 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 train machine learning models that predicted QTL haplotype calls using GBS data, only. We examined the effect of training size on cross validated prediction accuracies, the

forward validated accuracies in a population with known QTL haplotype calls, and finally we observed the estimated FHB resistance effects of predictive QTL haplotype calls verses known QTL haplotype calls using an historical data.

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 data set based on the Southern Uniform Winter Wheat Scab Nursery (SUWWSN) from the years 2011-2019 (Murphy et al., 2018, 2019; Murphy et al., 2017; Murphy et al., 2015; Murphy et al., 2016; 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 diagnostic for *Fhb1*, *Qfhb.nc-1A*, *Qfhb.nc-4A*, and the *Qfhb.vt-1B* Jamestown haplotype. The 2019-2020 SunGrains panel (SG20) contained 2,376 lines and the 2020-2021 SunGrains panel (SG21) included 3,423lines. 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. The genome wide GBS SNP data and QTL haplotype call data from the SG20 and SG21 panels was used for training of prediction models as well as to identify the effect of training size on prediction accuracy.

The historical data set, based on the SUWWSN was used to compare observed QTL haplotype call effects to predicted QTL haplotype call effects. Models trained on the SG20 and SG21 data sets were used to predict QTL haplotype calls in the SUWWSN. This dataset represented 95 total environments across the southeastern United States over eight years for 418 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 manufacturers 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. Data was filtered by removing any: 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 a 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 post filtering (Browning et al., 2018; Browning & Browning, 2007).

All lines in the SG20 and SG21 panels were genotyped using KASP markers diagnostic for *Fhb1*, *Qfhb.nc-1A*, *Qfhb.nc-4A*, and the *Qfhb.vt-1B* Jamestown haplotype. Marker sequences and 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. Historic data for the four FHB resistance QTL were compiled from the Uniform Winter Wheat Scab Nursery Marker Reports for the SUWWSN from 2011-2020 (Brown-Guedira, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019). For comparability and simplicity of predictions, only resistant (R) or susceptible (S) 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. Linkage disequilibrium was calculated in the regions for all QTL evaluated in the SG20, SG21, and SUWWSN data sets 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 USA Wheat and Barley Scab Initiative repositories. Data collected included adjusted means for heading date, plant height, severity (SEV), percent fusarium damaged kernels (FDK), and concentration of DON content as measured in parts per million. Heading data was recorded in days after January 1st when heading was evident in half the 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 spikelet 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 USA 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 marker effects:

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

Where y is the 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, e is the random environment effect, and ε is the residual error where $\boldsymbol{\varepsilon \sim (0, E\sigma_{\varepsilon}^2)}$.

QTL Prediction - Models

Four separate models were entertained as possible methods of predicting QTL haplotypes: 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 were used to predict the QTL haplotype call in the SunGrains testing and SUWWSN populations. Parameter tuning via K-fold cross validation within the training population was not conducted for NCOR due to the arbitrary nature with which the classifying marker was selected. All models following were implemented using the “caret” package in R statistical software and optimal model parameters were selected by five-fold cross-validation over 1000 iterations (Kuhn, 2008).

KNN functioned by finding K individuals in a training set that were most similar to an unclassified individual in a testing set; the most frequent class among those K neighbors in the training set was then 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 a randomly generated 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 were drawn using N number of random predictor variables to create splits at nodes in each tree. Once the random forest was generated from the training data, classifications were 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 tested 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), was similar to RF in that it draws a random forest comprised of decision trees made from random selected classifying variables; however, unlike RF, GBM took a logistic regression like approach to classification. The GBM algorithm first derived the log of odds from the observed classifications in the training population and calculated a probability via the logistic function. Then, the GBM algorithm calculated pseudo-residuals from the observed class probability of individuals versus the predicted probability derived from the most frequent class. An initial decision tree was then drawn using randomly selected classifier variables to a limited number of leaves. Unlike RF, GBM had multiple observations in a single leaf. Each leaf’s residuals were 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 was 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 tested 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, ect.). The number of leaves per tree were scaled by training population size so that trees contained only 10, 20 or 30 leaves.

QTL Prediction - Procedure

A general visual diagram of the described procedure is provided (Figure 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 tested: 10%, 25%, 50%, 75%, and 90% of the total available data. Data were randomly subset, without replacement, into training-test 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 0-100 for ease of comparison.

Each trained model was used to predict the QTL haplotype calls of the held out test portion of the SunGrains panel and confusion matrices were calculated using the QTL haplotype call predictions and the observed QTL haplotype calls. The same models trained using the SunGrains data 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

effects and means. 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 predicted QTL haplotype call effects and means.

Results

Cross Validated Accuracies in the SG20 and SG21

For *Fhb1*, prediction accuracies for the held out testing portion of the SG20 and SG21 populations, over all training population sizes, years, and iterations, ranged from 0.88-0.96 for NCOR, 0.93-1.0 for KNN, 0.95-1.0 for RF, and 0.95-1.0 for GBM. For *Qfhb.nc-1A*, accuracies across years and training population sizes ranged from 0.50-0.89 for NCOR, 0.85-0.98 for KNN, 0.86-0.98 for RF, and 0.83-0.96 for GBM. For *Qfhb.nc-4A,* accuracies across years, training population sizes, and iterations ranged from 0.86-0.94 for NCOR, 0.87-0.99 for KNN, 0.91-0.99 for RF, and 0.89-0.99 for GBM. For *Qfhb.vt-1B,* accuracies across years, training population sizes, and iterations ranged from 0.91-0.98 for NCOR, 0.96-1.0 for KNN, 0.97-1.0 for RF, and 0.95-1.0 for GBM. 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 Table 1).

Forward Validated Accuracies in the SUWWSN

The SG20, SG21, and SUWWSN populations had similar frequencies for the *Fhb1* resistant haplotype (Table 2). However, for *Qfhb.nc-1A, Qfhb.nc-4A,* and *Qfhb.vt-1B*, 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 are as follows: 0.88 for *Fhb1*, 0.60 for *Qfhb.vt-1B*, 0.55 for *Qfhb.nc-1A*, and *0.67* 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 population sizes, over all iterations, accuracies for *Fhb1* ranged from 0.78-0.92 for NCOR, 0.86-0.99 for KNN, 0.89-0.99 for RF, and 0.93-0.98 for GBM. For *Qfhb.nc-1A*, accuracies ranged from 0.53-0.77 for NCOR, 0.72-0.91 for KNN, 0.75-0.92 for RF, and 0.67-0.88 for GBM. For *Qfhb.nc-4A*, accuracies ranged from 0.87-0.90 for NCOR, 0.77-0.89 for KNN, 0.82-0.90 for RF, and 0.80-0.90 for GBM. For *Qfhb.vt-1B*, accuracies ranged from 0.81-0.92 for NCOR, 0.90-0.98 for KNN, 0.93-0.98 for RF, and 0.92-0.98 GBM. 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 Table 2).

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 underperformed in SG20 in terms of accuracy, and sensitivity in comparison to NCOR for *Qfhb.nc-4A* (Figure 2A, 2B). Machine learning model accuracies were most-often higher than the no-information rate, except for a few cases in the smallest training sizes (10%); however, moderate-to-large training size (50-90%) models were always above the no-information rate. For predictions made by NCOR, there were several instances of accuracies beneath the no-information rate for all QTL. In general, predictive accuracies tended to decrease with smaller training sizes and higher variability in kappa, accuracy, sensitivity, and specificity was observed in smaller training sizes as compared to large training sizes; indicating that the moderately to large sized (50-90%) training populations may produce results with higher specificity, and thus a lower type I error (i.e. lower frequency of false positives).

Machine learning models were generally superior to NCOR in terms of specificity, indicating that NCOR has a high type II error in comparison to machine learning models. Furthermore, NCOR often had a sensitivity comparable to machine learning models, except in smaller training population sizes, which indicated that using the most correlated marker in a training set to predict the QTL haplotype may lead to a high false negative rate and a tendency to classify a line as not containing a resistant haplotype. All machine learning models performed relatively similarly in terms of all confusion matrix coefficients over the 30 iterations, indicating no single superior machine learning model over all QTL tested in terms of forward-validated accuracy.

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. Linkage patterns were highly similar between SG20, SG21 and SUWWSN with regions containing one to three distinct blocks of linkage (Supplementary Figure 1A, 1B). 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 (Figure 3A, 3B).

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 interactions of the entire procedure. However, regardless of this issue, KNN, RF, and GBM all identified (on average) SNPs in or near the boundaries of KASP assays as highly important for the QTL haplotype call prediction in both years. If no markers were available in the region, as is the case for *Fhb1* in the SG21 dataset, markers in close proximity (less than one megabase pair) were indicated as highly important in predicting the QTL haplotype call.

Analysis and Comparison of Imputed Verses Observed QTL Haplotype Call Estimated Effects

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 observed QTL haplotype calls derived from KASP assays 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 non-resistant haplotype.

Across years, predicted QTL haplotype calls estimated group means tended to vary more widely when training population sizes were small (10%) and were more stable and consistent when training sizes were moderate to large (50-90%). Regardless of training population size, estimations of QTL haplotype call group means tended not to vary outside one standard error of the observed QTL haplotype calls estimate group means (Figure 4). This phenomenon remained consistent among QTL, traits, models, training size, 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 consistent between all training sizes; this is most-likely why the estimate remained very consistent between iterations.

Averages of effect sizes and group means for each QTL-by-trait-by-model-by-training-size combinations were calculated (Supplemental Table 3). The average of estimated group means for *Fhb1, Qfhb.nc-1A*, and *Qfhb.vt-1B* resistant and susceptible predicted QTL haplotype calls tended to closely match that of the observed QTL haplotype call group means (Figure 5A, 5B). Stronger separation of group means were observed when making predictions with the SG20 rather than the SG21 regarding FDK and DON for *Qfhb.nc-1A* and FDK for *Qfhb.vt-1B*. Among the QTL tested, *Fhb1* appeared to have the most consistent group mean estimates across years from predicted QTL haplotype calls.

Discussion

Identification of lines which contain resistance QTL for FHB is key to resistant cultivar development. Moreover, identification of lines in earlier generations which contain resistance QTL poses a major advantage. 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, accuracies for predictions in the SUWWSN increased as training population size increased, yet accuracies, sensitivities, and specificities were comparable among the 90 percent and 50 percent of training sizes. This indicated that perhaps a training population of approximately 1,000-1,500 individuals may adequately serve as a training population for this approach. When looking at the variability of estimated group means for predicted QTL haplotype calls, it was observed that large training sizes (90 percent) may be best at mimicking observed QTL haplotype estimated group means. Thus, larger training sizes of 2,000-3,000 individuals may be better suited for producing QTL haplotype call predictions that most closely approximate the “true” group means.

Among the QTL tested, *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-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 (kilobase pairs) in comparison to the other QTL (*Qfhb.nc-1A* = 25 mega base pairs, *Qfhb.vt-1B* = 141 mega base pairs, and *Qfhb.nc-4A* = 32 megabase pairs). From the visualization of LD in the region (Supplemental Figure 1), clearly there is one defined linkage block in the delimited region for *Fhb1*, whereas other QTL tested can have two or more linkage blocks. However, the QTL with the largest delimited region, *Qfhb.vt-1B*, often had comparable accuracies to *Fhb1*, so it appears that linkage nor size of the QTL explain the underperformance of models for *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 tested 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*.

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 KASP and 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 known QTL of interest (Lundberg et al., 2013).

Regardless, we showed that when comparing estimated group means of QTL haplotype calls made through molecular marker assays verses QTL haplotype call predictions, that QTL haplotype predictions remain within one standard error of the group mean prediction, and on average, closely resemble the observed group means. We therefore propose the following schematic wherein predictive QTL haplotype calling is incorporated into a breeding pipeline (Figure 6). Moreover, we hypothesize that this method may be extended to cover not only FHB resistance QTL, but other major effect QTL for which QTL haplotype calling is performed with KASP assays.

Declarations

FUNDING

The research conducted in this study was supported by funds by the North Carolina Small Grains Growers Association.

CONFLICT OF INTEREST

The author claims no conflict of interest.

AVAILABLITY OF DATA

KASP and genotyping-by-sequencing data for SunGrains and Southern Uniform Winter Wheat Nursery lines is unavailable for public use. Phenotypic data for the Southern Uniform Winter Wheat Scab Nursery is 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/Prediction-of-FHB-Resistance-QTL-Haplotypes-Through-Molecular-Markers-GBS-and-ML.git>

AUTHOR CONTRIBUTION STATMENT

Conceptualization, J.L. and Z.J.W.; methodology, J.L. and Z.J.W..; software, Z.J.W.; validation, Z.J.W.; formal analysis, Z.J.W.; investigation, J.L. and Z.J.W.; data curation, 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.; writing–original draft preparation, Z.J.W.; writing–review and editing, 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.; visualization, Z.J.W.; supervision, J.P.M., G.B.G., and J.L.; project administration, J.P.M; funding acquisition, J.P.M.

Tables

Table 1. Name, positions, and sequences of KASP assays used for haplotyping in the present study. Primers are labeled as susceptible (S), resistant (R), or common reverse (CR). Primer position is given in mega base pairs (Mbp).

QTL	Chromosome	KASP Name	Physical Position (Mbp)	Primer	Primer Sequence
Fhb1	3B	TaHRC-Kasp-S	13.64	S	GAAGGTGACCAAGTTCATGCTTTGTCTGTTTCGCTGGGATG
				R	GAAGGTCGGAGTCAACGGATTGCTCACGTCGTGCAAATGGT
				CR	CTTCCAGTTTCTGCTGCCAT
		snp3BS-8	13.96	S	GAAGGTGACCAAGTTCATGCTCACATGCATTTGCAAGGTTGTTATCC
				R	GAAGGTCGGAGTCAACGGATTCACATGCATTTGCAAGGTTGTTATCG
				CR	CAAAGCAGCCTTAGGTCAATAGTTTGAAA
Qfhb.nc-1A	1A	IWA3805	541.90	S	GAAGGTGACCAAGTTCATGCTAACTTTGCTGTCAACTTTGAGGA
				R	GAAGGTCGGAGTCAACGGATTCTAACTTTGCTGTCAACTTTGAGGG
				CR	TTACTGCAACTGATGGGTGCACTTTATAT
		IWA886	569.75	S	GAAGGTGACCAAGTTCATGCTGTAAGCTGCTAGGTCTTGTAGCC
				R	GAAGGTCGGAGTCAACGGATTAAGTAAGCTGCTAGGTCTTGTAGCA
				CR	TACGTGCACGGTCGATCAGTTTCTA
		IWA1587	566.51	S	GAAGGTGACCAAGTTCATGCTCTATCTATATTCTTTGTTCTTCAAGTCCA
				R	GAAGGTCGGAGTCAACGGATTCTATCTATATTCTTTGTTCTTCAAGTCCG
				CR	GATTGTTGCAACTAGCAACAGCTGTTTAT
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
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

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	0.13	0.25	0.29	0.32
SG21	0.14	0.24	0.28	0.31
SUWWSN	0.12	0.40	0.55	0.43

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

QTL	Trait	Effect	SE	Resistant		Susceptible		Wald Statistic	P Value	Significance
				Mean	SE	Mean	SE			
Fhb1	DON	-2.43	0.04	6.85	1.00	9.29	0.82	16.07	0.0001	***
	FDK	-7.68	0.06	21.21	2.35	28.90	1.99	21.07	0.0000	***
	Heading Date	1.00	0.02	123.34	2.19	122.34	2.17	8.95	0.0028	**
	Plant Height	0.13	0.02	34.19	0.68	34.06	0.61	0.05	0.8243	NS
	SEV	-9.44	0.07	24.01	2.15	33.45	1.76	36.06	0.0000	***
Qfhb.vt-1B	DON	-1.79	0.02	7.20	1.01	8.99	0.99	17.35	0.0000	***
	FDK	-1.66	0.05	27.26	2.43	28.92	2.31	2.09	0.1484	NS
	Heading Date	0.10	0.01	120.50	2.97	120.39	2.97	0.54	0.4619	NS
	Plant Height	0.38	0.02	33.61	0.77	33.24	0.75	0.56	0.4542	NS
	SEV	-3.38	0.04	30.27	2.42	33.66	2.34	6.69	0.0097	**
Qfhb.nc-1A	DON	-1.32	0.01	7.80	0.89	9.12	0.89	12.10	0.0005	***
	FDK	-2.49	0.03	28.38	2.31	30.87	2.29	5.93	0.0149	*
	Heading Date	-0.13	0.01	122.85	2.51	122.98	2.51	0.41	0.5229	NS
	Plant Height	0.52	0.01	34.05	0.62	33.53	0.62	3.20	0.0737	NS
	SEV	-2.36	0.03	32.37	2.02	34.73	2.02	6.15	0.0131	*
Qfhb.nc-4A	DON	0.54	0.01	8.87	0.98	8.33	0.96	1.34	0.2463	NS
	FDK	0.51	0.03	29.15	2.38	28.64	2.33	0.07	0.7976	NS
	Heading Date	0.47	0.01	121.82	2.74	121.35	2.74	5.05	0.0246	*
	Plant Height	-0.18	0.01	33.80	0.68	33.99	0.68	0.79	0.3732	NS
	SEV	-0.64	0.03	32.62	2.29	33.26	2.27	1.17	0.2795	NS

References

1. 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.
2. 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.). *Molecular Breeding*, *36*(7), 1-11.
3. Belkasim, S., Shridhar, M., & Ahmadi, M. (1992). Pattern classification using an efficient KNNR. *Pattern Recognition*, *25*(10), 1269-1274.
4. Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
5. 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.
6. 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.

7. 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.
8. 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.
9. Brown-Guedira, G. (2015). 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. Brown-Guedira, G., Griffey, C., Kolb, F., McKendry, A., Murphy, J., & Sanford, D. (2008). Breeding FHB-resistant soft winter wheat: Progress and prospects. *Cereal Research Communications*, *36*(Supplement-6), 31-35.
15. Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, *103*(3), 338-348.
16. Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, *81*(5), 1084-1097.
17. Buerstmayr, M., Steiner, B., & Buerstmayr, H. (2020). Breeding for Fusarium head blight resistance in wheat—Progress and challenges. *Plant Breeding*, *139*(3), 429-454.
18. Butler, D., Cullis, B. R., Gilmour, A., & Gogel, B. (2009). ASReml-R reference manual. *The State of Queensland, Department of Primary Industries and Fisheries, Brisbane*.
19. Carpenter, N. R., 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 Science*, *60*(6), 2919-2930.
20. Cuthbert, P. A., Somers, D. J., 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.). *Theoretical and Applied Genetics*, *112*(8), 1465.
21. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
22. Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PloS one*, *9*(2), e90346.
23. He, C., Holme, J., & Anthony, J. (2014). SNP genotyping: the KASP assay. In *Crop breeding* (pp. 75-86). Springer.
24. Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, *28*(1), 1-26.
25. Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D., & Dangl, J. L. (2013). Practical innovations for high-throughput amplicon sequencing. *Nature methods*, *10*(10), 999-1002.
26. 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 Disease*, *96*(12), 1712-1728.
27. Moose, S. P., & Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant physiology*, *147*(3), 969-977.
28. 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.
29. 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
30. 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
31. 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

32. 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.
33. 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.
34. 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.
35. 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.
36. 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.
37. 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.
38. National Grain and Feed Association. (2011). FDA Mycotoxin Regulatory Guidance. In *A Guide for Gain Elevators, Feed Manufacturers, Grain Processors and Exporters* (pp. 7): National Grain and Feed Association.
39. Perdry, H., & Dandine-Roulland, L. (2018). Gaston—Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models. *R Package*, *83*, 1-29.
40. Petersen, S., Lyerly, J. H., Maloney, P. V., Brown-Guedira, G., Cowger, C., Costa, J. M., . . . Murphy, J. P. (2016). Mapping of Fusarium head blight resistance quantitative trait loci in winter wheat cultivar NC-Neuse. *Crop Science*, *56*(4), 1473-1483.
41. Petersen, S., Lyerly, J. H., McKendry, A. L., Islam, M. S., Brown-Guedira, G., Cowger, C., . . . Murphy, J. P. (2017). Validation of Fusarium head blight resistance QTL in US winter wheat. *Crop Science*, *57*(1), 1-12.
42. Poland, J. A., Brown, P. J., Sorrells, M. E., & 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).
43. R Core Team. (2013). R: A Language and Environment for Statistical Computing. In.
44. Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, *13*(5), 278-289.
45. Sarinelli, J. M., Murphy, J. P., Tyagi, P., Holland, J. B., Johnson, J. W., Mergoum, M., . . . Sutton, R. (2019). Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. *Theoretical and Applied Genetics*, *132*(4), 1247-1261.
46. 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. *Nature genetics*, *51*(7), 1099-1105.
47. Waldron, B., Moreno-Sevilla, B., Anderson, J., Stack, R., & Frohberg, R. (1999). RFLP mapping of QTL for Fusarium head blight resistance in wheat. *Crop Science*, *39*(3), 805-811.
48. Ward, T. J., Clear, R. M., Rooney, A. P., O'Donnell, K., Gaba, D., Patrick, S., . . . Nowicki, T. W. (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 Genetics and Biology*, *45*(4), 473-484.
49. Wright, E. E. (2014). *Identification of Native FHB Resistance QTL in the SRW Wheat Cultivar Jamestown* Virginia Tech].

Figures

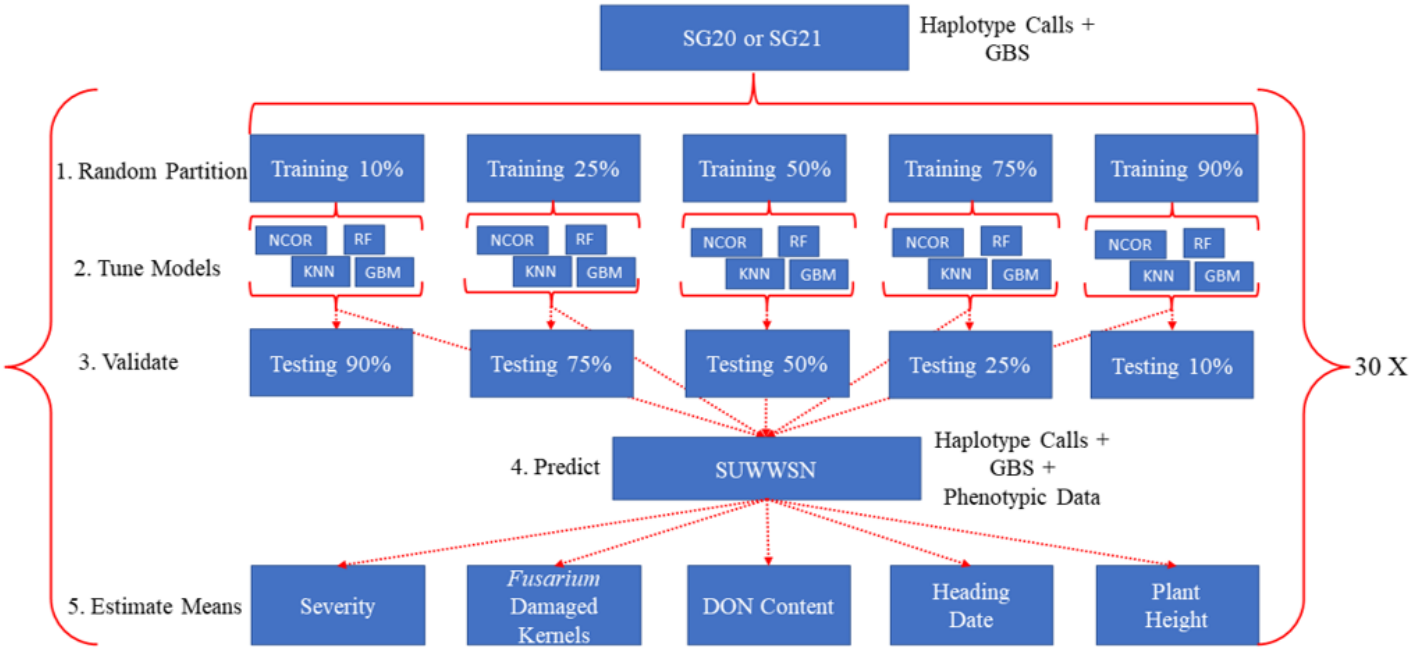

Figure 1

Schematic of total 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 was 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). (5) The QTL haplotype calls predicted in (4) were used to estimate the effect of the predicted QTL haplotype call across the available data for the SUWWSN.

Figure 2

A. 2020 SunGrains trained model forward predictions made on the Uniform Southern Winter Wheat Scab Nursery, specificity, and sensitivity boxplots of the 30 iterations for 10, 50, and 90% training sizes for (I) *Qfhb.nc-1A*, (II) *Qfhb.vt-1B*, (III) *Qfhb.nc-4A*, and (IV) *Fhb1*. Models; 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 sub graph. The y-axis denotes the response.

B. 2021 SunGrains trained model forward predictions made on the Uniform Southern Winter Wheat Scab Nursery, specificity, and sensitivity boxplots of the 30 iterations for 10, 50, and 90% training sizes for (I) *Qfhb.nc-1A*, (II) *Qfhb.vt-1B*, (III) *Qfhb.nc-4A*, and (IV) *Fhb1*. Models; 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 sub graph. The y-axis denotes the response.

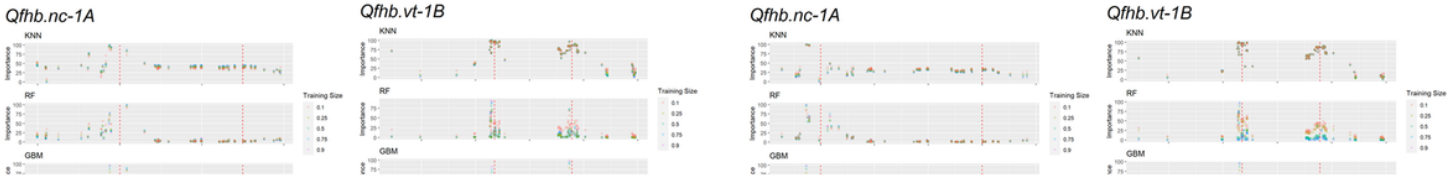

Figure 3

A. 2020 SunGrains average GBS SNP marker importance values. Importance values are scaled between 0-100 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 0-100 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.

Figure 4

Estimation of group (resistant vs susceptible) means using predicted QTL haplotype call vs observed QTL haplotype calls made using KASP assays. Plots show estimations of group means of deoxynivalenol content (DON) for each permutation from 1-30. Each point represents the mean of lines predicted to contain (R) and not contain (S) the resistant haplotype for *Fhb1*. Dark green and dark red bands surrounding point estimates of means show possible values within one standard error. Brightly colored green and red lines denote a one standard error interval for the observed QTL haplotype call means. Predictions of QTL haplotype calls were made by random forest (RF), k-nearest neighbor (KNN), gradient boosting machine (GBM), and using the most correlated marker in the training population (NCOR). (A) depicts the estimated group means derived from models trained with the SunGrains 2019-2020 data (SG20). (B) depicts the estimated group means derived from models trained with the SunGrains 2020-2021 data (SG21).

Figure 5

A. Estimated means of predicted QTL haplotype calls using the SG20 population verses 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 verses 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

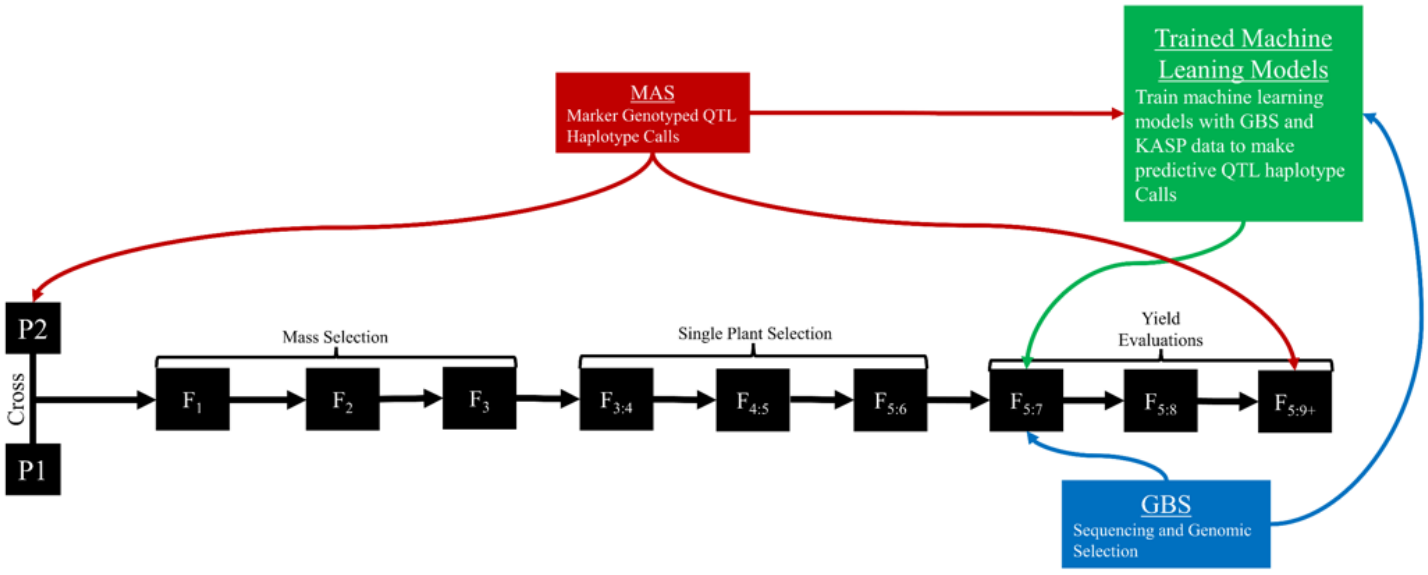

Figure 6

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 makers 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.

Supplementary Files

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

- SUPPLEMENTARYMaterial.docx