A New Strategy for Using Historical Imbalanced Yield Data to Conduct Genome-Wide Association Studies and Develop Genomic Prediction Models for Wheat Breeding

Chenggen Chu
Texas A&M AgriLife Research Center at Amarillo

Shichen Wang
Genomics and Bioinformatics Service Center

Jackie C. Rudd
Texas A&M AgriLife Research, Amarillo

Amir M.H. Ibrahim
Texas A&M University College Station

Qingwu Xue
Texas A&M AgriLife Research, Amarillo

Ravindra N. Devkota
Texas A&M AgriLife Research Center

Jason A. Baker
Texas A&M AgriLife Research Center

Shannon Baker
Texas A&M AgriLife Research Center

Bryan Simoneaux
Texas A&M University

Geraldine Opena
Texas A&M University

Haixiao Dong
Washington State University

Xiaoxiao Liu
Texas A&M AgriLife Research Center

Kirk E. Jessup
Texas A&M AgriLife Research Center

Ming-Shun Chen
USDA-ARS, Hard Winter Wheat Genetics Research Unit

Kele Hui
Texas A&M AgriLife Research Center

Richard Metz

Genomics and Bioinformatics Service Center, Texas A&M AgriLife Research

Charles D. Johnson

Genomics and Bioinformatics Service Center, Texas A&M AgriLife Research

Zhiwu S. Zhang

Washington State University

Shuyu Liu (SLiu@ag.tamu.edu)

Texas A&M AgriLife Research, Amarillo https://orcid.org/0000-0003-4748-2900

Research Article

Keywords: genetic correlation, wheat breeding, imbalanced data, insect resistance, genomic prediction accuracy

Posted Date: September 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-494076/v2

Version of Record: A version of this preprint was published at Molecular Breeding on March 22nd, 2022. See the published version at https://doi.org/10.1007/s11032-022-01287-8.

Abstract

Using imbalanced historical yield data to predict performance and select new lines is an arduous breeding task. Genome-wide association studies (GWAS) and high throughput genotyping based on sequencing techniques can increase prediction accuracy. An association mapping panel of 227 Texas elite (TXE) wheat breeding lines was used for GWAS and a training population to develop prediction models for grain yield selection. An imbalanced set of yield data collected from 102 environments (year-by-location) over ten years, through testing yield in 40–66 lines each year at 6–14 locations with 38–41 lines repeated in the test in any two consecutive years, was used. Based on correlations among data from different environments within two adjacent years and heritability estimated in each environment, yield data from 87 environments were selected and assigned to two correlation-based groups. The yield best linear unbiased estimation (BLUE) from each group, along with reaction to greenbug and Hessian fly in each line, were used for GWAS to reveal genomic regions associated with yield and insect resistance. A total of 74 genomic regions were associated with grain yield and two of them were commonly detected in both correlation-based groups. Greenbug resistance in TXE lines was mainly controlled by *Gb3* on chromosome 7DL in addition to two novel regions on 3DL and 6DS, and Hessian fly resistance was conferred by the region on 1AS. Genomic prediction models developed in two correlation-based groups were validated using a set of 105 new advanced breeding lines and the model from correlation-based group G2 was more reliable for prediction. This research not only identified genomic regions associated with yield and insect resistance but also established the method of using historical imbalanced breeding data to develop a genomic prediction model for crop improvement.

Introduction

The complex nature of grain yield makes it difficult for precise selection of lines with high yield potential. Accumulation of favorite alleles combinations for yield would lead to lines with improved yields. For identifying favorite alleles, genome-wide association studies (GWAS) (Atwell et al. 2010; Rafalski 2010) have showed advantages over traditional QTL mapping using bi-parental mapping populations. GWAS uses the natural collection of germplasm lines such as landraces, varieties, and breeding lines as mapping panels, and detects historical recombination events and linkage disequilibrium (LD) to identify the non-random association between allele loci and traits (Flint-Garcia et al. 2003). GWAS can identify multiple alleles simultaneously since a wider range of germplasms in a panel would contain more diverse genetic composition (Atwell et al. 2010; Myles et al. 2009; Zhu et al. 2008).

Increasing marker coverage in the genome will enhance the power of GWAS for allele identification. Single nucleotide polymorphisms (SNPs), the variations on a single nucleotide at the specific position, are the most abundant and widely distributed genome markers (Agarwal et al. 2008). With the advances in DNA sequencing technology, the genotype-by-sequencing (GBS) (Elshire 2011) and later double-digested restriction-site associated DNA sequencing (ddRADseq) (Baird et al. 2008; Peterson et al. 2012) are robust approaches of identifying SNPs that are randomly distributed throughout the whole genome and thus are suitable for investigating genome-wide genetic variations. Particularly, by aligning SNP flanking DNA sequences to assembled whole genome sequences of hexaploidy wheat (IWGSC 2014; Zimin et al. 2017), tetraploid wheat (Avni et al. 2017) and *Ae. tauschii* (Jia et al. 2013b), chromosomal location of each SNP can be precisely located to accurately track favorite alleles.

Previously, QTL analysis was widely used to identify genomic regions associated with the traits and then develop molecular markers for marker-assisted selection (MAS). However, most important traits such as grain yield, yield components, end-use quality, etc., are all highly polygenic with each locus contributing only a very small proportion of total phenotypic variance (Collard and Mackill 2008; Jia et al. 2013a; Simmonds et al. 2014; Tyagi et al. 2014), which leads to weak stability and low repeatability in those QTLs and thus limits the application of MAS for accumulating desirable genes in crop improvement. Genomic selection (GS) utilizes a large set of markers covering a whole genome to detect all possible alleles within the LD and their effects on the trait, and to estimate the genomic breeding value of each line to conduct selection in breeding (Bernardo and Yu 2007; Bhat et al. 2016; Daetwyler et al. 2008; Meuwissen et al. 2001; Rutkoski et al. 2016; Sun et al. 2019; Tsai et al. 2020). Therefore, using GWAS to identify favorite alleles and form the training population to develop prediction models for conducting GS will be an efficient way of accumulating desirable alleles for improving yield and other polygenic traits.

To conduct GWAS and GS, the composition of training populations and precision of phenotyping are two additional critical factors affecting prediction accuracy (He et al. 2016; Marulanda et al. 2015; Michel et al. 2017). Using advanced lines from the same breeding program as a training population has showed a positive effect on prediction accuracy (Daetwyler et al. 2008; Endelman et al. 2014). It was demonstrated that a training population including lines from the same family, half sibs and more distant lines could be efficient for a GS scheme (Verges and Van Sanford 2020). Using historical advanced breeding lines developed at different periods together with germplasm lines from the same breeding program may also be a good strategy for association mapping and genomic prediction studies, simply because all those lines could represent all genetic sources in the program with historical recombination maintained to ensure mapping resolution, particularly when the training population is keeping updated as the new germplasm lines were introduced into the program.

Another advantage of using advanced breeding lines in GWAS and genomic prediction is that the lines have been evaluated in many years under different environmental conditions and have phenotypic data already available. This would be especially helpful for traits such as grain yield that requires many resources for phenotyping (Verges and Van Sanford 2020). However, the most significant difficulty in using phenotypic data of the historical breeding lines for GWAS is that the lines are evaluated at different time and that the data were typically imbalanced. Research using imbalanced data in historical breeding lines for GWAS and genomic prediction have been seldom reported. The possibility of using imbalanced data for GS was explored by clustering analysis of data through pre-defined mega-environments based on climatic patterns, farming systems, water regimes and the incidence of biotic and abiotic stress, but this strategy appears ineffective for genomic selection (Dawson et al. 2013). Therefore, it is necessary to identify an appropriate way that can directly utilize the imbalanced historical data for GWAS and genomic prediction.

During the past few decades, the Texas A&M AgriLife Research and Extension Center at Amarillo, TX has released several drought tolerant cultivars such as ‘TAM 105’, ‘TAM 107’, ‘TAM 111’ and ‘TAM 112’ (Lazar et al. 2004; Porter et al. 1980; Porter et al. 1987; Rudd et al. 2014). Among them, TAM 111 and TAM 112 were two broadly planted cultivars in the Great Plains hard red winter wheat regions since 2010 based on planted acreages (NASS, 2011–2013 http://www.nass.usda.gov). A newer release 'TAM 114' has superior bread-making quality and drought tolerance (Rudd et al. 2018). A grain and forage dual purpose awnless wheat 'TAM 204' has higher yield, drought tolerance and a good level of resistance to insects such as greenbug, Hessian fly, and wheat curl mite (Rudd et al. 2019). These widely adapted winter wheat cultivars have been used as germplasm lines in wheat breeding programs in the U.S. and many other countries. To localize their genes conferring the superior traits will greatly improve selection efficiency for improving yield, end-use quality, and tolerance to biotic and abiotic stresses.

In this research, we used a set of 227 elite breeding lines (including the aforementioned released cultivars) developed by Texas A&M AgriLife Research wheat breeding programs in the last ten years as the mapping panel for conducting GWAS to identify favorite alleles and as the training population to build a genomic prediction model for selecting grain yield. By combining correlation analysis with genetic heritability estimation using the imbalanced yield data collected from diverse environments, we successfully developed a data management strategy of using the imbalanced historical yield data for conducting GWAS and building genomic prediction models. In addition, the genomic prediction model was further validated using a set of newly developed advanced breeding lines from Texas wheat breeding programs.

Materials And Methods

Plant materials

The set of 227 Texas elite (TXE, F_9) breeding lines were developed by the two Texas A&M AgriLife Research wheat breeding programs located at Amarillo and College Station, TX during 2009 – 2018, which included 13 released TAM cultivars. Briefly, the lines were originally selected at the F_6 generation according to their performance in the observation yield trials conducted at two locations followed by evaluation preliminary yield trials in six locations and advanced yield trials in ten locations at the F_7 and F_8 generations, respectively. The state-wide yield TXE trials were conducted at 16 locations across Texas with three replicates per location. The superior TXE lines were further evaluated for traits of yield, end-use quality, biotic and abiotic tolerance, and agronomic traits either toward the new cultivar release or as the germplasms to enter the new breeding cycles. Therefore, the set of TXE lines represented the major gene sources in Texas wheat breeding programs and were appropriate materials for building genomic prediction models to improve selection efficiency. In addition, a set of 105 lines entered into the advanced yield trials was evaluated at ten locations with two replications per location, and their yield data were used to validate the genomic prediction model developed from the TXE collection.

Grain yield data analysis

The TXE trials were conducted during 2009 - 2018 were conducted in three replications at 16 locations that represented four typical wheat growing regions (High plains, Rolling plains, Blacklands and South Texas) in Texas (Fig. S1), and grain yield data were collected from 6 to 14 locations each year (Table 1) due to the abandoned harvest in some locations experienced serious damage caused by severe weather. Totally, yield data were from 102 environments defined as year-by-location combinations (Table 1). However, TXE yield data were typically imbalanced with 40 to 66 lines evaluated each year and 38 to 41 lines were also tested in the following year (Table 1). Three cultivars, ‘TAM 112’, ‘TAM 401’ and ‘TAM W-101’ were used as controls across all years.

To manage the imbalanced yield data for using in GWAS and then developing genomic prediction model for grain yield in Texas wheat breeding programs, genetic correlations coefficients among yield data of common TXE lines in different environments were calculated through R package META-R (Alvarado et al. 2020), and the significantly positive correlation indicated that those common TXE lines showed similar trends of reacting to growing condition under those environments. Yield data collected from all environments were then grouped according to their correlations. For example, if data of common lines in datasets A and B were correlated and data of common lines in datasets B and C were correlated, the three datasets A, B and C will be kept in one correlated group though no common lines between A and C for correlation calculation. The best linear unbiased prediction (BLUP) in each correlation group was calculated using the mixed model $y = X\beta + Zu + \varepsilon$ using the R package *lme4* (Bates et al. 2015) with genotype was set as the only random effect, where y represents the vector of observations, β and u mean fixed and random effects, respectively, X and Z are matrices of observations related to fixed and random effects, respectively, and ε is the residual of the model. Since genotype was the only random effect in this model, variation from random effect will be the genetic variance (V_G) and thus can be used to estimate the heritability (H^2) of grain yield in each correlation group using the formula $H^2 = V_G / (V_G + V_e)$, where V_e is the residual variance. If a dataset from one environment was included in heritability estimation and lead to an increase in yield heritability of one correlation-based group, the dataset was kept in that correlation-based group. Such a heritability estimation was conducted for each environment, including the datasets that were non-correlated in correlation analysis in the previous step, to finally determine if the dataset should be kept in the corresponding correlation-based group.

Once correlation-based groups of yield data were finalized through heritability estimation, the R package *lme4* and the mixed model $y = X\beta + Zu + \varepsilon$ were used again to calculate the best linear unbiased estimation (BLUE) of each line in each correlation-based group with genotype was set as the fixed and environment set as the random effects. The yield BLUEs calculated in each group were used for GWAS and developing genomic prediction models.

Evaluation resistance to greenbug and Hessian fly

Growth chamber and greenhouse experiments were conducted to evaluate resistance to greenbug (*Schizaphis graminum* Rondani) and Hessian fly (*Mayetiola destructor* Say) in the TXE lines. Briefly, wheat plants grown in one-gallon pots were maintained in 60 × 60 × 60 cm cages (MegaView Science Co., Ltd., Taichung, Taiwan) equipped with insect proof mesh. Greenbug biotype E and Hessian fly biotype GP colonies were established and maintained on caged wheat plants for approximately six weeks prior to the evaluation experiments and were used as the source for the subsequent assays.

Greenbug infestation was done according to Weng and Lazar (2002). Cultivars ‘TAM 105’ or TAM 111, and ‘TAM 110’ were used as susceptible and resistant controls, respectively. Sixteen lines and two controls were grown in a 30 × 50 cm flat with 20 seeds per line. At the three-leaf stage, about 500 greenbugs were scattered over each test flat and the flats were then kept in a growth chamber at 22 °C with a day length of 8 hours. The plant was scored as either resistant (normal healthy) or susceptible (chlorotic leaf and necrotic stem lesions) 10 -14 days after infestation. Percentage of resistant plants in each line was recorded for GWAS.

Hessian fly infestation was conducted as described in (Chen et al. 2009). Wheat accessions ‘Carol’ (H3), ‘Cardwell’ (H6) and ‘Molly’ (H13) were used as the resistant checks, and ‘Danby’ as the susceptible control. Twenty lines and four checks with 25 seeds per line were planted in one plastic flat (56 × 36 cm) in a greenhouse at 18 ± 3 °C with day length as 14 hours. When the first leaf was fully expanded and the second leaf started emerging, about 200 newly mated female flies were released to each flat covered with a cheesecloth tent (540 × 120 × 40 cm). Resistance rating were conducted three weeks later and the stunted plants having bloated live larvae at stem base were considered as susceptible (S), and the normally healthy plants with small dead larvae or tiny live larvae between leaf sheaths as resistant (R). Percentage of resistant plants per line was calculated for GWAS.

SNP genotyping and marker data management

Whole genomic DNA was extracted from leaf samples using CTAB (cetyl trimethylammonium bromide) method (Stewart and Via 1993) with slight modification (Liu et al. 2013). SNP genotyping was done through ddRADSeq procedure (Peterson et al. 2012). Briefly, genomic DNA was co-digested with two restriction enzymes *PstI* (CTGCAG) and *MspI* (CCGG) and barcoded adapters were then ligated to DNA segments of each individual sample. Adapter oligos were synthesized from Integrated DNA Technologies (IDT), Inc. (Coralville, Iowa), and were mixed in equimolar amounts (30 μM of top and bottom oligos). After denaturing at 95°C for 10 sec, oligos were cooled to 12°C at a rate of 0.1°C per sec. P5-Index adapters were made through annealing the top and bottom oligos (Top oligo (5’ - 3’): AAT GAT ACG GCG ACC ACC GAG ATC TAC ACX XXX XXX XTC TTT CCC T; Bottom oligo (5’ -3’): /5Phos/AXX XXX XXX GTG TAG ATC TCG GTG GTC GCC GTA TCA TT, where XXXXXXXX represents 8-base i5 index sequences). The P5-PstI-Bridge adapters were made by annealing top (Pster_T, 5’ to 3’): /5Phos/ACA CGA CGC TCT TCC GAT CTT GCA and bottom (Pster_B, 5’ to 3’): AGA TCG GAA GAG CGT CGT GTA GGG AAA G oligos. P7-MluCI Adapter was made by annealing top (P7-MluCI_T, 5’ to 3’): AAT TAG ATC GGA AGA GCA CAC GTC TGA ACT CCA GTC AC and bottom (P7-MluCI_B, 5’ to 3’): GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T.

The ddRADSeq libraries were constructed using 96-plex plate with a single random blank well used for quality control, and were then sequenced through an Illumina HiSeq 2000 at the Genomics & Bioinformatics Services of Texas A&M AgriLife Research at college station, TX (Yang et al. 2020), and SNP calls were made using the reference-based Stacks Pipeline (Catchen et al. 2013) using IWGSC v1.0 as the reference genome (IWGSC 2014), which obtained over 247,000 raw SNP data with missing rate below 50%. Considering all TXE lines were at F_9 generation or later that have a very low level of heterozygosity and thus the homozygous SNP readings should be more reliable, which is also approved by comparison of SNP readings of few control lines with 2 - 4 replications included for DNA sequencing and SNP calling (data not shown). Majority of heterozygous SNP readings were more likely due to technique error during sequencing according to their extra high heterozygosity rate. Therefore, heterozygous marker data from TXEs thus were all converted as the missing data, and all SNP data were filtered again using criteria of data-missing rate less than 30% and minor allele frequency (MAF) below 5% through the computer package Tassel v5.0 (http://www.maizegenetics.net/) (Bradbury et al. 2007), which retained over 75,000 SNPs with a higher level of reliability. Genotype imputation with accuracy of 98% was then conducted using computer program Beagle (v5.0) (Browning and Browning 2007) and achieved data-missing rate to less than 10%. Imputed data were filtered again through MAF less than 5% and obtained the final set of 70,525 SNPs were used for GWAS.

In the set of 105 advanced breeding lines from yield trial in 2018, SNP genotyping and marker data management was done using the similar methods as indicated for TXE collections. A total of 384,648 SNPs was called and imputed in the 105 advanced lines with marker data missing rate less than 10%. Among those SNPs, 37,975 were the common set between TXE and advanced breeding lines and were extracted for validating the accuracy of genomic prediction model developed from TXE by comparing the predicted with observed yield in those 105 advanced lines.

Population structure analysis

From the raw 247,000 SNPs in TXE with a data missing rate of less than 50%, a set of 8,401 SNPs with data missing rate less than 18% and heterozygosity less than 5% was considered the most reliable markers for analyzing population structure in TXE lines. The computer program Structure v2.3.4 (https://web.stanford.edu/group/pritchardlab/structure.html) (Falush et al. 2003; Pritchard et al. 2000) was used with the number of presumable sub-populations (*K*) set from three to ten with iteration number equal to ten. For simulation running under each *K*, length of burn-in period was set to 10,000 and number of MCMC replicates was set to 100,000 with the model of admixture and correlated allelic frequency was used. The number of sub-populations was then determined using delta *K* (ΔK) method described in Evanno et al. (2005) through the online tool Structure Harvester (http://taylor0.biology.ucla.edu/structureHarvester/). Meanwhile, phylogenetic tree using 70,525 imputed SNP data through UPGMA (unweighted pair group method with arithmetic mean) hierarchical clustering method was also carried out through Tassel v5.0 (Bradbury et al. 2007) and the clade tree was drawn using the online tool Interactive Tree Of Life (iTOL v5) (Letunic and Bork 2019, https://itol.embl.de/). The phylogenetic tree was used to verify the results obtained from Structure v2.3.4.

Genome-wide association studies

GWAS was carried out using the set of 70,525 imputed SNPs through Tassel v5.0 (Bradbury et al. 2007). Principal component analysis (PCA) was conducted with the number of sub-populations determined by Structure v2.3.4 to generate the Q-matrix that incorporated as the covariate in association analysis, and the fixed and random effect mixed model (MLM) (Liu et al. 2016) was used for association mapping with the K-matrix showing the relationship of all individuals that was used to account for effects due to kinship. For detecting genomic regions associated with grain yield, the yield BLUE of each line calculated using the R package *lme4* in correlation-based groups was used as the trait data and GWAS was separately conducted in each correlation-based group. Bonferroni adjustment using R package simpleM (http://simplem.sourceforge.net/) (Gao et al. 2010) determined the significant threshold of LOD = 4.0 for grain yield and LOD = 6.0 for insect resistance.

SNP allele frequency change during new TXE lines development

According to the time of line development, the 227 TXE lines were divided into three groups using a three-year interval with the first group including 92 lines developed during 2009 – 2011(namely the old group), the second group containing 67 lines from 2012 – 2014, and the third group had 68 lines from 2015 – 2017 (namely the newly developed group). Therefore, comparing allele frequency between the first and the third groups would have a good indication of allele drifting due to breeding selection in Texas wheat breeding programs. Allele frequency change was investigated focusing on the major allele genotype of SNPs in the TXE collection. The frequency of each SNP major allele was respectively calculated in the first and third groups and then to find the difference between the two frequencies.

Genomic prediction model development

For developing genomic prediction models, 7,573 SNP markers from all chromosomes with at least one million bases (Mb) apart were selected for estimating the mean effect of each marker. The R package rrBLUP (ridge regression best linear unbiased prediction, Endelman 2011) was used for developing genomic prediction model. The mixed model $y = \mu + X\beta + \varepsilon$ was used with y as the vector of phenotypic means, μ as the overall mean, X as the marker matrix, β as the vector of marker effects and ε as the vector of residual effects. The genomic estimated breeding values (GEBVs) of each line were calculated by adding the grand mean to the product of genotypic matrix and the vector of mean effect of each marker. The prediction accuracy was measured by the correlation between the predicted and observed yield BLUEs. Genomic prediction models were developed separately in each of the correlation-based groups, and the prediction accuracy was estimated at three times using 60%, 70% and 80% of TXE lines as training sets and 40%, 30% and 20% as testing sets, accordingly. For each training/testing set, prediction accuracy was obtained based on a calculation using 500 repeated runs.

To validate the prediction models developed in each correlation-based group, all TXE lines were used as the training set and 105 advanced breeding lines were used as the testing set. The common SNPs between TXEs and advanced breeding lines with at least one Mb apart on each chromosome were selected. Marker effects were estimated using BLUEs of each correlation-based group through rrBLUP mixed model $y = \mu + X\beta + \varepsilon$. The GEBVs of each advanced breeding line were calculated by adding the grand mean to the product of genotypic matrix and the vector of mean effect of each marker. Prediction accuracy was then measured through the correlation between predicted and observed yield in 105 advanced breeding lines.

Results

Grain yield in two correlation-based groups

Based on correlations among yield of overlapped lines in different environments, yield data from eighteen environments were not correlated with any of the remaining 84 environments (Tables S1 and S2). Data from those 84 environments were divided into two correlation-based groups with groups G1 containing 36 and G2 including 48 environments (Table 1). After heritability estimation, data from five non-correlated environments were added into the group G1 but data from two environments were dropped from the group G2. Therefore, group G1 contained data of 41 environments and group G2 carried data from 46 environments for further analysis, and data from 15 non-correlated environments were abandoned (Table S3). Interestingly, the majority of dataset in correlation-based group G1 included environments from the High Plains and Rolling Plains that normally have low level of rainfall and represented the drought-prone areas of wheat acreage in Texas, whereas the majority of data in group G2 contained environments from Blacklands and South Texas that usually received relatively a higher level of rainfall and represented the wet growing conditions of Texas (Table S3).

Of the two correlation-based groups that respectively included data collected from 41 and 46 environments, the best linear unbiased estimation (BLUE) of each line was calculated through R package *lme4* and the skewed yield distributions were observed in both groups (Fig. 1). In correlation-based group G1, the grain yield of all lines ranged from 2,000 to 4,500 Kg/Ha, but 201 (88.5%) lines have grain yield in the range of 3,250–4,000 Kg/Ha. Whereas in the correlation-based group G2, grain yield of each lines was in the range of 2,250–4,750 Kg/Ha but with more lines distributed in a broader range (Fig. 1). The grain yield in two datasets were not correlated and Bartlett's test also indicated heterogeneous variances in the two datasets. Since the correlation-based groups G1 and G2 were corresponding to dry and wet growing condition in north and south Texas, respectively, the broader trait variation in group G2 corresponding to the less stressful growing conditions in south Texas may more likely explain the genetic variance.

Snp Genotyping And Population Structure In Txe Collection

Of the 70,525 imputed SNPs with data missing rate less than 10%, the B-genome chromosomes carried the most (37,773) SNPs, followed by the chromosomes in the A-genome (23,782) and the D-genome (8,970) (Table S4), indicating the relatively lower level of genetic diversity in the D-genome. Particularly, only 731 and 708 SNPs were identified on chromosomes 4D and 5D, respectively, suggesting that the two chromosomes in TXE lines have the least variations.

Population structure analysis revealed five sub-populations in the TXE collection (Fig. 2) and the population was mainly admixed with all released cultivars (Fig. S2) spread into different sub-populations, which indicated that the TXE collection covered primary gene sources in Texas wheat breeding programs. Phylogenetic tree developed using 70,525 imputed SNPs also suggested the similar population structure (Fig. 3).

Genome wide association studies of grain yield in TXE lines

Association analysis identified 74 genomic regions in two correlation-based groups associated with grain yield with significant level above threshold LOD = 4.0 (Fig. 5, Tables S5 and S6). Of those associations, two regions were commonly detected in both groups and located on chromosome 1D at 229.6 and 345.4 Mb with LOD scores of 4.4 and 4.8 and each contributing around 10% of yield variation. The favorite alleles at two regions had the additive effects of increasing yield by 330.5 and 325.0 Kg/Ha in group G1 and 406.7 and 343.5 Kg/Ha in group G2, respectively. In addition, 17 genomic regions identified only in correlation-based group G1 and 55 regions only in group G2 were associated with grain yield. In Group G1, those genomic regions were located on nine chromosomes, namely 2A, 3A, 3B, 5A, 5D, 6B, 6D, 7B, and 7D. The LOD scores ranged from 4.0 to 5.8 and explained 7–14% of yield variations with favorite alleles having the potential of increasing yield 220.9–816.3 Kg/Ha. There were ten genomic regions with each explaining 10% or more of phenotypic variations located on chromosome 2A at 308.1 Mb, 3A at 12.8 Mb, 3B at 245.3 Mb, 5A at 24.1 Mb, 6B at 32.8 and 682.6 Mb, 6D at 454.1 Mb, 7B at 627.2, 637.8, and 676.8 Mb. In group G2, the significant genomic regions were detected from all chromosomes with the significance varying from LOD = 4.0 to 8.1 and each accounting for 7 to 16% of trait variations with the favorite allele having the additive effects of increasing yield 258.0–516.1 Kg/Ha. The most significant association was located at 14.7 Mb on chromosome 7D and explained 16% of the phenotypic variations with the favorite allele having the potential of increasing yield by 516.1 Kg/Ha (Tables S5 and S6). A total of 16 regions on seven chromosomes each explained over 10% of trait variations in group G2 (Table S5 and Fig. 5).

Genome wide association studies of greenbug and Hessian fly resistance in TXE lines

For reaction to greenbug, 173 and 42 TXE lines were susceptible and resistant, respectively, and twelve lines showed partial resistance (Table S7). GWAS indicated three genomic regions on chromosome 3DL (565.0 Mb), 6DS (7.2 Mb) and 7DL (597.9 Mb) were associated with greenbug resistance in TXE lines (Fig. 6a and Tables S8 and S9). A region on chromosome 7DL showed the largest effect and explained 48.7% of the trait variation in TXE lines, and the regions on 3DL and 6DS explained 10.7% and 15.3% of phenotypic variation, respectively. For reaction to Hessian fly, data were obtained from 219 TXE lines with 166 susceptible, 18 resistant, and 35 partially resistant (Table S7). Only the genomic region on 1AS at 7.8 Mb were significantly associated with the resistance and explained 17.0% of trait variation in TXE lines (Fig. 6b and Tables S8 and S9).

SNP allele drift in TXE lines due to breeding selection

By comparing frequency of major alleles in 70,525 SNPs between TXE groups of 2009–2011 (old) and 2015–2017 (new), allele frequencies of 10,034 SNPs decreased. Meanwhile, allele frequency in a different set of 974 SNPs each increased over 20% in the new TXE group. Of the SNPs with allele frequency decreasing in the newly developed TXEs, the allele genotype of 2,000 SNPs that have been the major alleles in TXE group of 2009–2011 were changed to minor allele. Whereas in SNPs with allele frequency increasing in the newly developed TXEs, 300 SNPs changed the allele status from minor in the group of 2009–2011 to major in the group of 2015–2017 (Table S10). Comparing the SNPs that had significant associations in increasing yield, several were located in the vicinity of the SNPs that had allele status changing from minor to major in new TXEs, such as the ones on 2A at 101.9 Mb, 3B at 245.3 and 738.3 Mb, 6D at 454.1 Mb, 7A at 620.3 Mb, 7D at 621.2 Mb, and 7B at 654.1, 711.9 and 741.0 Mb. Each of these alleles showed the potential of increasing yield by 281.8–611.3 Kg/Ha (Table S5). Among the 74 SNPs significantly associated with yield in groups G1 or G2, there were trends that the newer lines or cultivars had more favorite alleles for increasing yield (Table S6). Based on the pseudomolecule physical position indicated in the reference wheat genome sequence (IWGSC 2014), markers with allele frequency changing over 20% were mostly located at the distal sides of the chromosomes (Fig. 4), and agreed with the higher rate of recombination observed at the distal regions of the chromosomes.

Genomic prediction in TXE lines and model validation using advanced breeding lines

Genomic prediction models were tested in three situations that randomly picked 60%, 70% and 80% of TXE lines as training populations to predict the remaining TXE lines. After 500 independent runs in each situation, prediction accuracy using yield data from correlation-based group G2 is higher than using data from correlation-based group G1 (Table 2). Average prediction accuracies in group G1 varied from 0.42 to 0.44 but that increased from 0.68 to 0.70 in group G2 as the size of training population increased. The lowest range of the prediction accuracies were 0.14–0.19 in group G1 and 0.49–0.52 in group G2 and the maximum prediction accuracies were 0.61–0.75 in group G1 and 0.81–0.89 in group G2. This indicated that yield data in correlation-based group G2 were more reliable for genomic prediction.

To validate the prediction models developed in two correlation-based groups, yield data in a set of 105 advanced breeding lines were collected in 2018 from four environments including rain-fed and irrigated location in Bushland, TX, irrigated location in Etter, TX, and the rain-fed location in McGregor, TX. All TXE lines were used as the training set and 105 advanced breeding lines were used as the testing set. From the common 37,975 SNPs between TXE and advanced breeding lines, a total of 5,542 SNPs that were at least 1-Mb apart on each chromosome were selected for genomic prediction, and the prediction accuracies using the models developed from the correlation-group G2 ranged from 0.12 to 0.29, but none of the predictions based on the models from correlation-group G1 were correlated with the observed yield (Table 3). This is consistent with previous results of models from correlation-group G2 which were more reliable for genomic prediction when using TXE lines as both training and testing sets.

Discussion

Wheat grain yield is a very complex trait and is affected by numerous genes involved in many different biological processes affecting plant development, photosynthesis, carbon mobilization, grain filling and maturity. The effect of each gene is very limited and varied under different environments. Testing grain yield in breeding lines thus demands major efforts and breeding resources since it needs to be done in many locations under multiple years with replications included. Using historical yield data in the past or current breeding lines or cultivars to conduct genome-wide association analysis and genomic prediction will provide a cost-effective way of identifying beneficial genes for increasing yield and cumulating favorite alleles for crop improvement. However, imbalanced historical data obtained during breeding are hard to use since each environmental condition is unique and cannot be repeated. Interactions between genotypes and environments vary at different times and locations, which greatly increased difficulties of identifying favorite alleles. In this study, we developed a strategy of using correlations among yield data of overlapped lines evaluated at different times and locations to group different environments that have showed interactions with similar magnitudes. The grouping is further tested through heritability estimation. The best linear unbiased estimation (BLUE) calculated in each correlation-based group was then used for GWAS and developing genomic prediction models, which were further validated through a set of advanced breeding lines. This research thus developed a method of using the imbalanced historical data for genetic studies.

There are numerous QTLs for yield and yield components identified from bread wheat trials worldwide. Chromosome regions with significant SNPs associated with yield from GWAS in this study were very close to the QTLs identified from previous research from bread wheat trials conducted in the US Great Plains or other regions (Table S5). Yield associated genomic regions in this research at 71.4 Mb on 2A, 21.8 Mb on 2D, 12.8 Mb on 3A, 42.6 Mb on 3B, 682.6 Mb on 6B, 627.2 Mb and 654.1 Mb on 7B, and 592.2 Mb on 7D are very close to the QTLs at 79.8 Mb on 2A, 15.7 Mb on 2D, 9.6 Mb on 3A, 48.6 Mb on 3B, 673.8 Mb on 6B, 617.0 Mb and 647.8 Mb on 7B, and 591.2 Mb on 7D that were associated with yield and yield components identified from a bi-parental mapping population derived from the cross between TAM 111 and TAM 112 (Yang et al. 2020). Particularly, the region around 591.2 Mb on 7D is harboring gene *Gb3* conferring greenbug resistance (Liu et al. 2014). Breeders found that the majority of the TAM 112 derivatives had a decent yield in dry environments as *Gb3* was kept (J Rudd, personal communication, 2020).

Yield-associated regions at 603.0 Mb on chromosome 3D and 25.4 Mb on 7B were very close to the QTLs associated with spikes per square meter at 603.8 Mb on 3D linked to *XIWA6485* and kernel per spike at 22.6–24.7 Mb on 7B linked to *XIWB71684* with favorite alleles from ‘TAM 111’ (Assanga et al. 2017). The yield-associated region at 603.0 Mb on 3D was very close to a QTL at 603.4 Mb on 3D of ‘ND 705’ that was associated with spikes per square meter and linked to *XIWB17317* (Kumar et al. 2019).

Yield-associated regions at 532.8 Mb on 1A and at 711.5 Mb on 7B from this study were physically close to a flour yield QTL at 533.4 Mb on 1A and a grain volume weight QTL at 709.6 Mb on 7B detected in a recombinant inbred mapping population derived from cross between TAM 111 and TAM 112 population (Yang et al. 2020). The yield-associated region at 531.3 Mb on 2D from this study was also identified in the mapping population derived from TAM 112/TAM 111 and associated with thousand kernel weight and kernel diameter (Dhakal et al. 2021; Yang et al. 2020). Since cultivars TAM 111 and TAM 112 were core parents used in the Texas A&M AgriLife Research wheat breeding programs, it is very possible that these favorite alleles were carried through generations due to selections.

The yield-associated region at 16.2 Mb on 7D was close to gene *TaGS3-D1* located in region 6.5–6.8 Mb on 7D affecting wheat kernel weight and length (Rasheed et al. 2016; Zhang et al. 2014). Two QTLs at 32.8 Mb and 47.5 Mb on 6B associated with thousand kernel weight (Zou et al. 2017) coincided with the two yield-associated regions at 32.8 and 47.5 Mb on 6B in this study. The regions at 709.2 Mb on 3A, 625.7 Mb on 5A, 633.9 Mb on 6B from this study were very close to gene *TaTGW6-A1* at 711.1 Mb on 3A, a QTL at 619.5 Mb on 5A for thousand kernel weight, and a QTL at 631.8 Mb on 6B for grain volume weight (Juliana et al. 2019).

The region at 597.9 Mb on 7DL showed significant association with greenbug resistance and is corresponding to *Gb3*, the gene known to be carried by germplasms used for developing TXE lines and present in cultivars such as TAM 110, TAM 112, 'TAM 115', and TAM 204 (Lazar et al. 1997; Liu et al. 2014; Rudd et al. 2014; Rudd et al. 2019; Weng and Lazar 2002). The other two regions on 3DL and 6DS with minor effects on greenbug resistance might be novel genes since no greenbug resistance have been reported from these two genomic regions. Hessian fly resistance was associated with a region on 1AS (7.8 Mb) in this study, the position coincided with a Hessian fly resistance QTL on 1AS in 'Duster' (PI 644016, Edwards et al. 2012) (Li et al. 2015). It is likely that Hessian fly resistance in TXE lines is derived from Duster since many TXE lines had this cultivar in their pedigree.

From this study, eight chromosome regions significantly associated with yield, along with several major genes in TXE lines such as wheat curl mite resistance genes Cmc_{TAM112} and *Cmc3* (Dhakal et al. 2018; Dhakal et al. 2017), seed storage protein subunit genes *Gli-B1* and *Glu*-D1, dwarf gene *Rht*-B1, and grain weight and length gene *TaGS-D1* (Liu et al. 2014; Zhang et al. 2014). These mostly coincided with regions in which allele frequencies were greatly increased in the newly developed lines (Fig. 4b; Table S10), which may be a good indication of accumulating favorite alleles during selection. Particularly, the recently released cultivars fall into different sub-populations (Figs. 3 and S2) showing improved yield, disease and/or insect resistance, drought tolerance, and enhanced baking and milling quality attributes. As aforementioned, cultivars TAM 111 (Lazar et al. 2004) and TAM 112 (Rudd et al. 2014) were used as parental lines in new releases due to their high yield and superior drought tolerance in addition to the greenbug and wheat curl mite resistance carried in TAM 112. For example, cultivar TAM 114 derived from crosses of using TAM 111 as parents showed excellent baking and milling quality, and intermediate resistance to Hessian fly (Rudd et al. 2018), and cultivar TAM 204 selected from the crosses involving TAM 112 showed a good level of resistance to greenbug, Hessian fly and wheat curl mite in addition to the high grain yield (Rudd et al. 2019). The newly released TAM 115 is also a selection from crosses involving TAM 112 and showed high yield, good drought tolerance and resistance to greenbug and wheat curl mite (Rudd et al., not published). Therefore, further research focusing on the regions where allele frequency greatly increased in those newly developed TXE lines may provide an efficient way of revealing beneficial alleles for wheat improvement.

Of the two correlation-based groups G1 and G2 developed through historical yield data of TXE lines, results from GWAS and genomic prediction both indicated that yield data from group G2 may be more reliable. This is supported by the fact that group G2 contained environments either in north Texas under irrigated conditions or from south Texas with relatively higher level of rainfall and thus had better growing conditions. On the other hand, the group G1 included mainly dryland environments with severer drought stress, which greatly limited the expression of yield potential in each line and led to a much narrow yield variation (Fig. 1). Similarly, GWAS detected fewer genomic regions significantly associated with yield in group G1 than in G2 (Table S5), and genomic prediction models in two groups using 60–80% of lines as training set also pointed to lower prediction accuracy in group G1 (0.14–0.75) than in G2 (0.49–0.89) (Table 2). Validation of genomic prediction models through a set of advanced breeding lines also indicated that predictions made using yield data of group G2 showed stronger correlation with the observed data (Table 3). Therefore, the strategy of combining correlation and heritability estimates to group data from different environments used in this study also provided a way of selecting appropriate data from diverse environments for genetic analysis.

Declarations

Author contributions

CC analyzed the data, summarized results and prepared the manuscript draft. RM and CDJ conducted ddRADSeq library preparation and DNA sequencing. S.W. conducted DNA sequence alignment, SNP call and SNP imputation. JCR, QX and AMHI supervised the field data collection, and helped on manuscript revision. RD, JAB, SB, BS, GO, XL, KEJ, MC and KH helped on data collection. HD and ZSZ helped on data analysis. S.L. designed and supervised the research project and revised the manuscript. All authors reviewed and approved the submission.

Funding

This research was supported partially by Texas A&M AgriLife Research, Texas Wheat Producers Board, U.S. Department of Agriculture, Agricultural Research Service, and National Research Initiative Competitive Grants 2017-67007-25939, 2019-67013-29172, and 2021-67013-33940.

Conflict of interest statement

The authors declare that they have no conflicts of interest.

Acknowledgements

Great thanks to Dr. Shuhao Yu in Texas A&M AgriLife Research at Amarillo for providing critical review to this manuscript.

References

1. Agarwal M, Shrivastava N, Padh H (2008) Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep 27:617–631
2. Alvarado G, Rodríguez FM, Pacheco A, Burgueño J, Crossa J, Vargas M, Pérez-Rodríguez P, Lopez-Cruz MA (2020) META-R: A software to analyze data from multi-environment plant breeding trials. The Crop Journal 8:745–756
3. Assanga SO, Fuentealba M, Zhang G, Tan C, Dhakal S, Rudd JC, Ibrahim AMH, Xue Q, Haley S, Chen J, Chao S, Baker J, Jessup K, Liu S (2017) Mapping of quantitative trait loci for grain yield and its components in a US popular winter wheat TAM 111 using 90K SNPs. PLOS ONE 12:e0189669
4. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631

5. Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357:93–97
6. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one 3:e3376
7. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using *lme4*. J Stat Softw 67:48
8. Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090
9. Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, Jadon V, Tyagi A, Mushtaq M, Jain N, Singh PK (2016) Genomic selection in the era of next generation sequencing for complex traits in plant breeding. Frontiers in genetics 7:221
10. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635
11. 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. The American Journal of Human Genetics 81:1084–1097
12. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140
13. Chen MS, Echegaray E, Whitworth RJ, Wang H, Sloderbeck PE, Knutson A, Giles KL, Royer TA (2009) Virulence analysis of Hessian fly populations from Texas, Oklahoma, and Kansas. J Econ Entomol 102:774–780
14. Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences 363:557–572
15. Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS one 3:e3395
16. Dawson JC, Endelman JB, Heslot N, Crossa J, Poland J, Dreisigacker S, Manès Y, Sorrells ME, Jannink JL (2013) The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crops Research 154:12–22
17. Dhakal S, Liu X, Girard A, Chu C, Yang Y, Wang S, Xue Q, Rudd JC, Ibrahim AMH, Awika JM, Jessup KE, Baker JA, Garza L, Devkota RN, Baker S, Johnson CD, Metz RP, Liu S (2021) Genetic dissection of end-use quality traits in two widely-adapted wheat cultivars ‘TAM 111’ and ‘TAM 112’. Crop Science (in press)
18. Dhakal S, Tan C-T, Anderson V, Yu H, Fuentealba MP, Rudd JC, Haley SD, Xue Q, Ibrahim AMH, Garza L, Devkota RN, Liu S (2018) Mapping and KASP marker development for wheat curl mite resistance in “TAM 112” wheat using linkage and association analysis. Mol Breeding 38:119

19. Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation genetics resources 4:359–361

20. Edwards JT, Hunger RM, Smith EL, Horn GW, Chen MS, Yan L, Bai G, Bowden RL, Klatt AR, Rayas-Duarte P, Osburn RD, Giles KL, Kolmer JA, Jin Y, Porter DR, Seabourn BW, Bayles MB, Carver BF (2012) 'Duster' wheat: a durable, dual-purpose cultivar adapted to the Southern Great Plains of the USA. Journal of Plant Registrations 6:1–12

21. Elshire RJ, Glaubitz JC, Qi Sun JA, Poland K, Kawamoto ES, Buckler SE, Mitchell (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6

22. Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4:250–255

23. Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, Sorrells ME, Jannink JL (2014) Optimal design of preliminary yield trials with genome-wide markers. Crop Sci 54:48–59

24. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

25. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164

26. Flint-Garcia SA, Thornsberry JM, Buckler IVES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

27. Gao X, Becker LC, Becker DM, Starmer JD, Province MA (2010) Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol 34:100–105

28. He S, Schulthess AW, Mirdita V, Zhao Y, Korzun V, Bothe R, Ebmeyer E, Reif JC, Jiang Y (2016) Genomic selection in a commercial winter wheat population. Theoretical applied genetics 129:641–651

29. IWGSC IWGSC (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345

30. Jia H, Wan H, Yang S, Zhang Z, Kong Z, Xue S, Zhang L, Ma Z (2013a) Genetic dissection of yield-related traits in a recombinant inbred line population created using a key breeding parent in China's wheat breeding. Theor Appl Genet 126:2123–2139

31. Jia J, Zhao S, Kong X, Li Y, Guangyao Zhao WH, Rudi Appels M, Pfeifer Y, Tao X, Zhang R, Jing C, Zhang Y, Ma L, Gao C, Gao M, Spannagl, Klaus FX, Mayer D, Li S, Pan, Fengya Zheng, Qun Hu, Xianchun Xia, Jianwen Li, Q, Liang J, Chen et al (2013b) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature doi:10.1038/nature12028

32. Juliana P, Poland J, Huerta-Espino J, Shrestha S, Crossa J, Crespo-Herrera L, Toledo FH, Govindan V, Mondal S, Kumar U, Bhavani S, Singh PK, Randhawa MS, He X, Guzman C, Dreisigacker S, Rouse MN, Jin Y, Pérez-Rodríguez P, Montesinos-López OA, Singh D, Mokhlesur Rahman M, Marza F, Singh RP (2019) Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat Genet 51:1530–1539

33. Kumar A, Mantovani EE, Simsek S, Jain S, Elias EM, Mergoum M (2019) Genome wide genetic dissection of wheat quality and yield related traits and their relationship with grain shape and size traits in an elite × non-adapted bread wheat cross. PLOS ONE 14:e0221826

34. Lazar MD, Worrall WD, Peterson GL, Porter KB, Rooney LW, Tuleen NA, Marshall DS, McDaniel ME, Nelson LR (1997) Registration of 'TAM 110' wheat. Crop Sci 37:2

35. Lazar MD, Worrall WD, Peterson GL, Fritz AK, Marshall D, Nelson LR, Rooney LW (2004) Registration of ‘TAM 111’ wheat. Crop Sci 44:355–356

36. Letunic I, Bork P (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research 47:W256–W259

37. Li G, Wang Y, Chen M-S, Edae E, Poland J, Akhunov E, Chao S, Bai G, Carver BF, Yan L (2015) Precisely mapping a major gene conferring resistance to Hessian fly in bread wheat using genotyping-by-sequencing. BMC Genom 16:108

38. Liu S, Griffey C, Hall M, McKendry A, Chen J, Brooks W, Brown-Guedira G, Sanford D, Schmale D (2013) Molecular characterization of field resistance to Fusarium head blight in two US soft red winter wheat cultivars. Theor Appl Genet 126:2485–2498

39. Liu S, Rudd JC, Bai G, Haley SD, Ibrahim AMH, Xue Q, Hays DB, Graybosch RA, Devkota RN, St. Amand P (2014) Molecular markers linked to important genes in hard winter wheat. Crop Sci 54:1304–1321

40. Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12:e1005767

41. Marulanda JJ, Melchinger AE, Würschum T (2015) Genomic selection in bi-parental populations: assessment of parameters for optimum estimation set design. Plant Breeding 134:623–630

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

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

44. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21:2194–2202

45. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one 7:e37135

46. Porter KB, Gilmore EC, Tuleen NA (1980) Registration of ‘TAM 105’ wheat. Crop Sci 20:114–114

47. Porter KB, Worrall WD, Gardenhire JH, Gilmore EC, McDaniel ME, Tuleen NA (1987) Registration of ‘TAM 107' wheat. Crop Sci 27:818–819

48. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics155:945 – 59

49. Rafalski JA (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174–180

50. Rasheed A, Wen W, Gao F, Zhai S, Jin H, Liu J, Guo Q, Zhang Y, Dreisigacker S, Xia X, He Z (2016) Development and validation of KASP assays for genes underpinning key economic traits in bread wheat. Theoretical Applied Genetics 129:18

51. Rudd JC, Devkota RN, Baker JA, Peterson GL, Lazar MD, Bean B, Worrall D, Baughman T, Marshall D, Sutton R, Rooney LW, Nelson LR, Fritz AK, Weng Y, Morgan GD, Seabourn BW (2014) 'TAM 112' wheat, resistant to greenbug and wheat curl mite and adapted to the dryland production system in the southern high plains. Journal of Plant Registrations 8:291–297

52. Rudd JC, Devkota RN, Ibrahim AM, Baker JA, Baker S, Lazar MD, Sutton R, Simoneaux B, Opena G, Rooney LW, Awika JM, Liu S, Xue Q, Bean B, Duncan RW, Seabourn BW, Bowden RL, Jin Y, Chen MS, Graybosch RA (2018) 'TAM 114' wheat, excellent bread-making quality hard red winter wheat cultivar adapted to the southern high plains. Journal of Plant Registrations 12:367–372

53. Rudd JC, Devkota RN, Ibrahim AM, Baker JA, Baker S, Sutton R, Simoneaux B, Opena G, Hathcoat D, Awika JM, Nelson LR, Liu S, Xue Q, Bean B, Neely CB, Duncan RW, Seabourn BW, Bowden RL, Jin Y, Chen M-S, Graybosch RA (2019) 'TAM 204' wheat, adapted to grazing, grain, and graze-out production systems in the southern high plains. Journal of Plant Registrations

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

55. Simmonds J, Scott P, Leverington-Waite M, Turner AS, Brinton J, Korzun V, Snape J, Uauy C (2014) Identification and independent validation of a stable yield and thousand grain weight QTL on chromosome 6A of hexaploid wheat (Triticum aestivum L.). BMC plant biology 14:1–13

56. Stewart C, Via LE (1993) A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. Biotechniques 14:748–751

57. Sun J, Poland JA, Mondal S, Crossa J, Juliana P, Singh RP, Rutkoski JE, Jannink J-L, Crespo-Herrera L, Velu G (2019) High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. Theor Appl Genet 132:1705–1720

58. Tsai H-Y, Janss LL, Andersen JR, Orabi J, Jensen JD, Jahoor A, Jensen J (2020) Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. Scientific reports 10:1–15

59. Tyagi S, Mir R, Kaur H, Chhuneja P, Ramesh B, Balyan H, Gupta P (2014) Marker-assisted pyramiding of eight QTLs/genes for seven different traits in common wheat (*Triticum aestivum* L.). Mol Breeding 34:167–175

60. Verges VL, Van Sanford DA (2020) Genomic selection at preliminary yield trial stage: training population design to predict untested lines. Agronomy 10:60

61. Weng Y, Lazar M (2002) Amplified fragment length polymorphism- and simple sequence repeat-based molecular tagging and mapping of greenbug resistance gene *Gb3* in wheat. Plant Breeding 121:218–223

62. Yang Y, Dhakal S, Chu C, Wang S, Xue Q, Rudd JC, Ibrahim AMH, Jessup K, Baker J, Fuentealba MP, Devkota R, Baker S, Johnson CD, Metz R, Liu S (2020) Genome wide identification of QTL associated

with yield and yield components in two popular wheat cultivars ‘TAM 111’ and ‘TAM 112’. PLOS ONE 15

63. Zhang Y, Liu J, Xia X, He Z (2014) *TaGS-D1*, an ortholog of rice *OsGS3*, is associated with grain weight and grain length in common wheat. Mol Breeding 34:1097–1107
64. Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. The Plant Genome 1:5–20
65. Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. Gigascience 6:gix097
66. Zou J, Semagn K, Iqbal M, N’Diaye A, Chen H, Asif M, Navabi A, Perez-Lara E, Pozniak C, Yang R-C, Randhawa H, Spaner D (2017) Mapping QTLs controlling agronomic traits in the ‘Attila’ × ‘CDC Go’ spring wheat population under organic management using 90K SNP array. Crop Sci 57:365–377

Tables

Table 1

List of datasets in two correlation-based groups formed by correlation analysis and broad-sense heritability estimation using yield data of 227 Texas Elite (TXE) breeding lines collected during 2009–2018

Year	Yield dataset	Dataset grouped through correlation			Dataset group adjusted by heritability test		
		Correlation-Group 1	Correlation-Group 2	Non-correlated	Correlation-Group 1	Correlation-Group 2	abandoned
2009	9	5	3	1	7	2	0
2010	14	6	5	3	7	6	1
2011	10	4	4	2	6	4	0
2012	14	3	8	3	3	9	2
2013	9	5	3	1	2	6	1
2014	10	2	7	1	3	7	0
2015	6	1	4	1	3	3	0
2016	9	6	3	0	6	1	2
2017	11	3	5	3	3	5	3
2018	10	1	6	3	1	3	6
Total	102	36	48	18	41	46	15

Table 2

Prediction accuracy using different portions of TXE lines as training and testing sets in two correlation-based groups G1 and G2

Percentage ration of training : testing	Correlation-based group G1			Correlation-based group G2		
	Average	minimum	maximum	Average	minimum	maximum
60% : 40%	0.42 ± 0.07	0.19	0.61	0.66 ± 0.04	0.52	0.81
70% : 30%	0.45 ± 0.08	0.14	0.65	0.68 ± 0.04	0.52	0.82
80% : 20%	0.46 ± 0.10	0.18	0.75	0.69 ± 0.06	0.49	0.89

Table 3

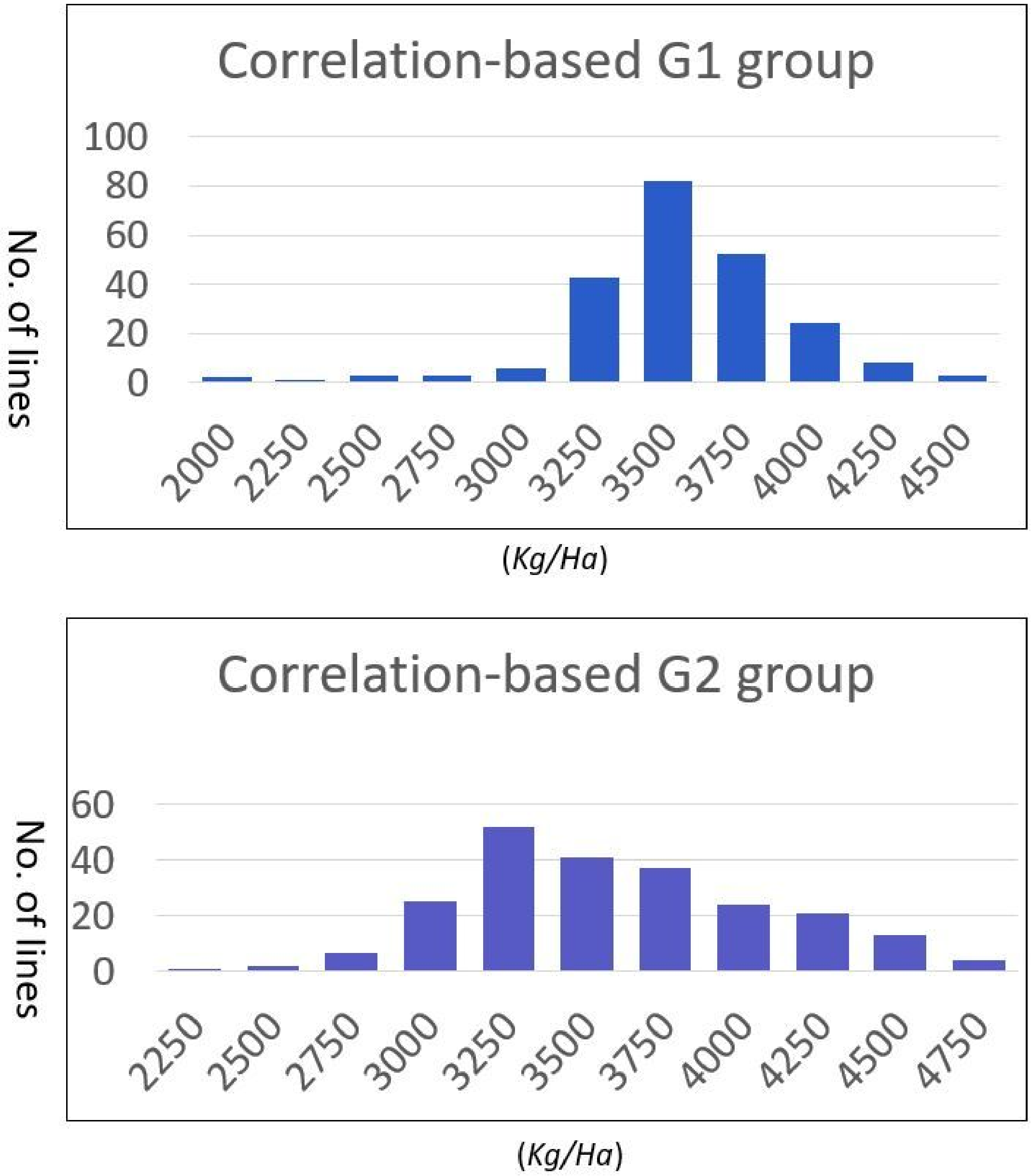
Prediction validation using yield data obtained from a set of advanced breeding lines in 2019 based on genomic prediction models developed in two correlation-based groups G1 and G2

Model	Environment [a]			
	BD	BI	EI	MCG
Correlation-based group G1	-0.06	-0.03	0.01	-0.06
Correlation-based group G2	0.29	0.12	0.14	0.23

[a] BD and BI mean rain-fed and irrigated land in location at Bushland, TX, respectively. EI means irrigated land in location at Etter, TX. MCG means rain-fed land location in McGregor, TX.

Figures

Fig. 1


Correlation-based G1 group

Correlation-based G2 group

Figure 1

Distribution of the best linear unbiased estimations (BLUEs) for grain yield in the two correlation-based groups G1 and G2. BLUEs were calculated using the R package lme4 (Bates et al. 2015).

Fig. 2

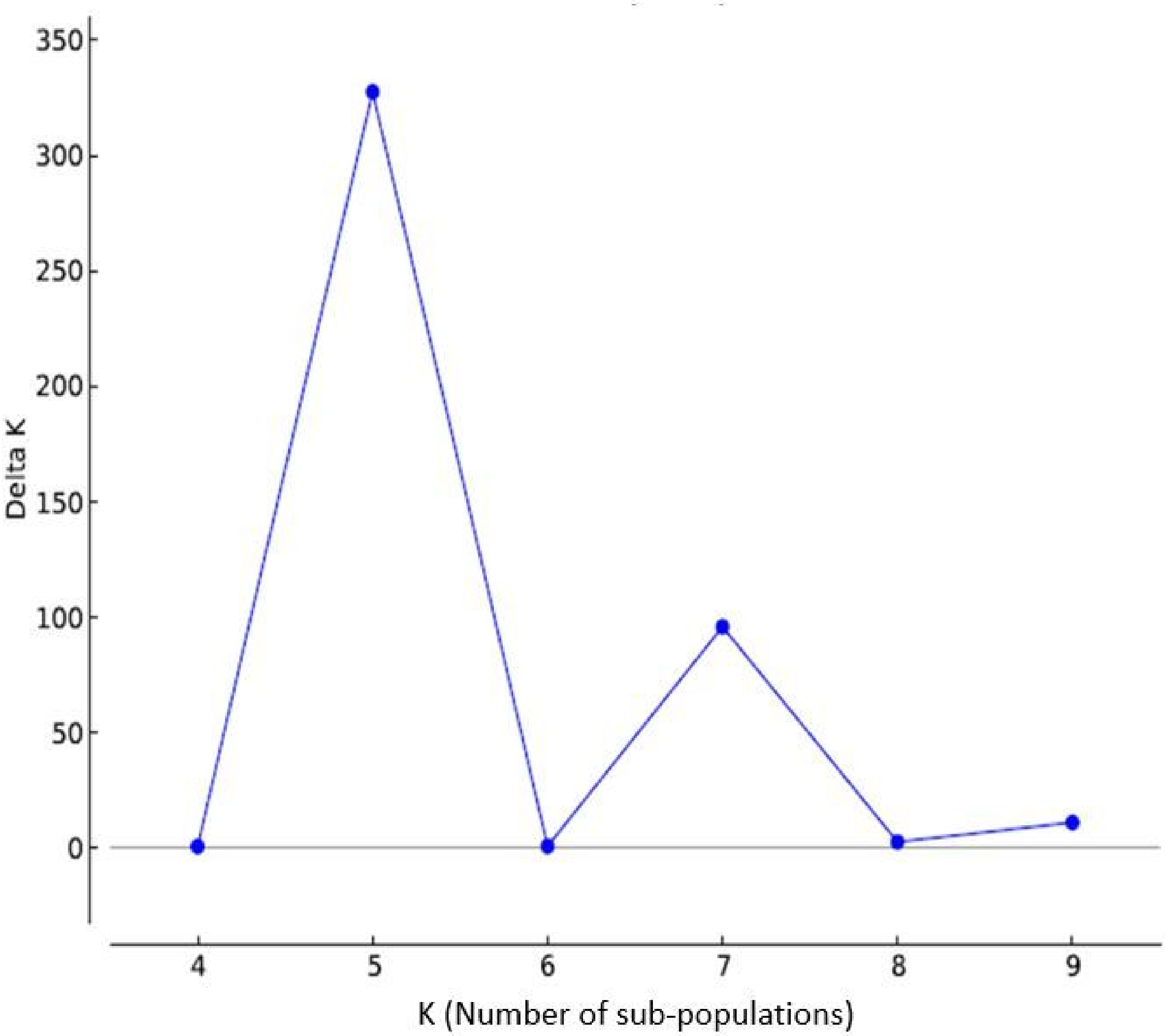


Figure 2

Population structure analysis using Structure v2.3.4 (Evanno et al. 2005; Pritchard et al. 2000) based on the most reliable set of 8,401 SNPs in 227 TXE lines. Five sub-populations were contained in the collection. Number of sub-populations were determined using Delta K method through Structure Harvester (Earl and vonHoldt 2012).

Fig. 3

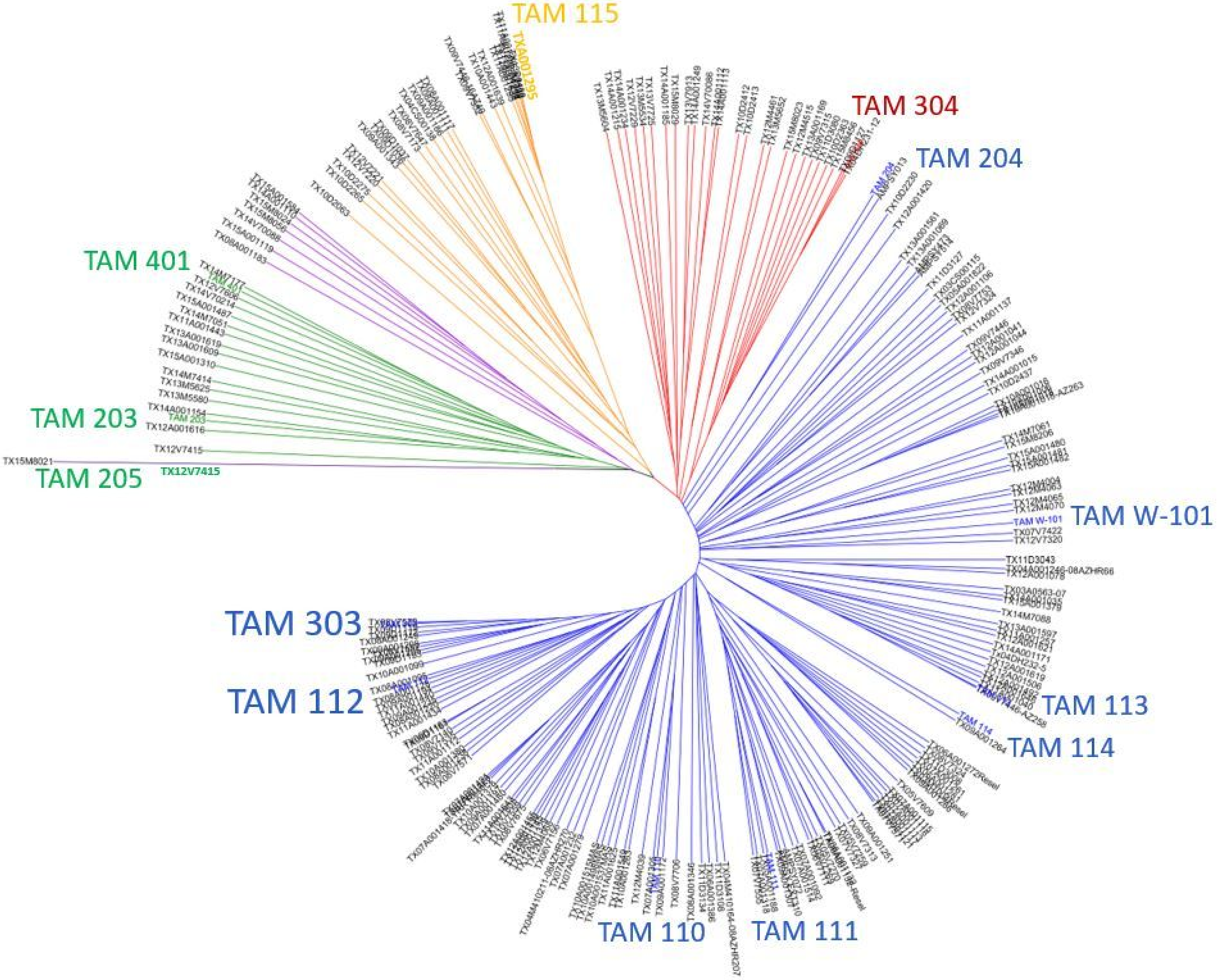
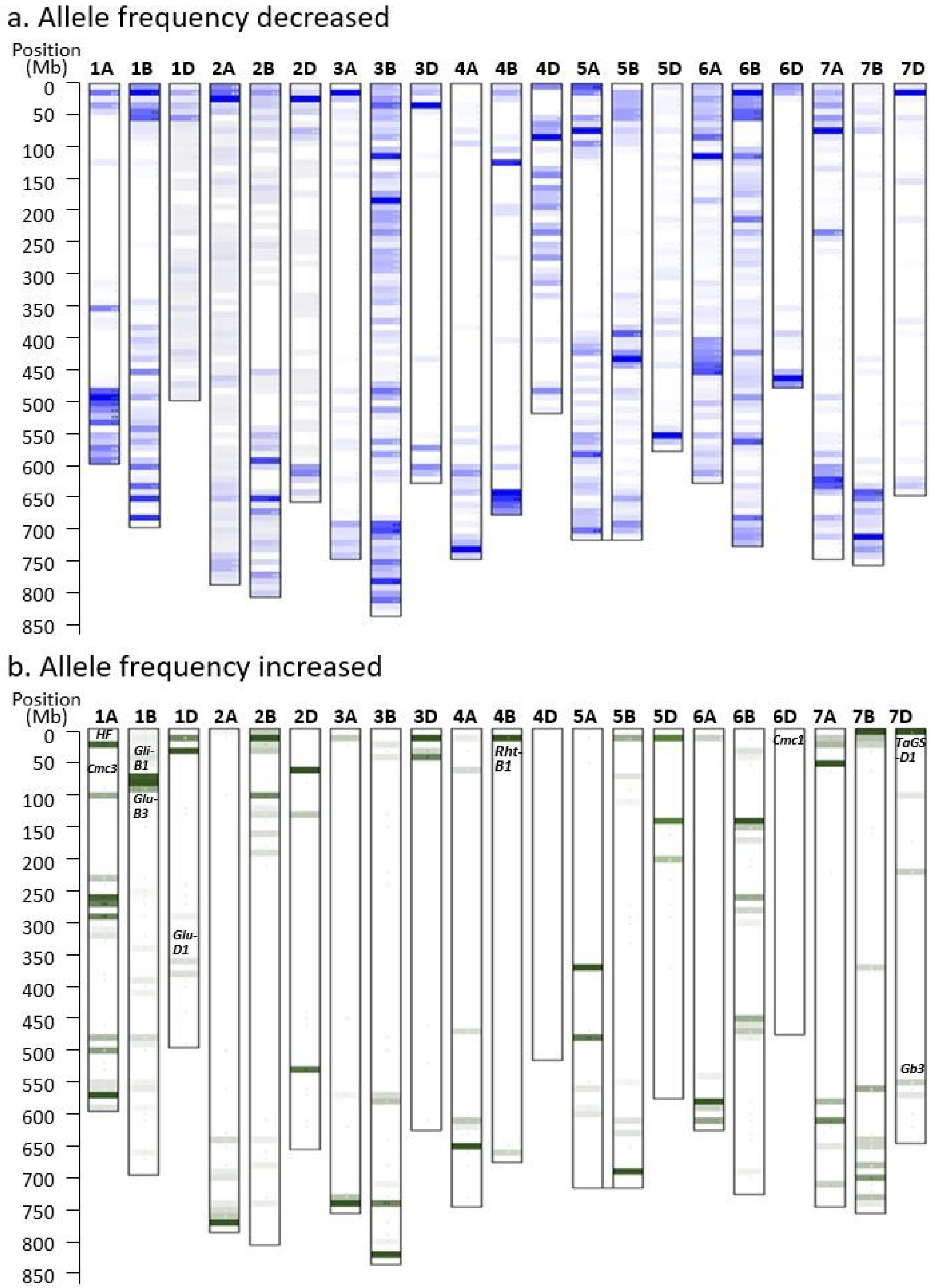


Figure 3

Phylogenetic tree developed using 70,525 imputed SNPs in 227 TXE lines. The released cultivars were indicated in the corresponding clusters. Phylogenetic tree was produced using Tassel v5.0 (Bradbury et al. 2007) through UPGMA (unweighted pair group method with arithmetic mean) hierarchical clustering method and was drawn using the Interactive Tree Of Life (iTOL v5) (Letunic and Bork 2019).

Fig. 4


a. Allele frequency decreased

b. Allele frequency increased

Figure 4

Diagram of allele frequency change among sets of TXE lines developed during 2009 – 2011 (old TXE, 92 lines) and 2015 – 2017 (newly developed TXE, 68 lines). a) allele frequency decreased in newly developed TXEs. b) allele frequency increased in newly developed TXEs with some major genes indicted in the corresponding position according to previous research (Dhakal et al. 2018; Liu et al. 2014; Zhang et al. 2014). Physical position of SNPs was determined according to pseudomolecule position in wheat reference genome v1.0 (IWGSC 2014). Darker regions indicated more markers have frequency changed.

Fig. 5

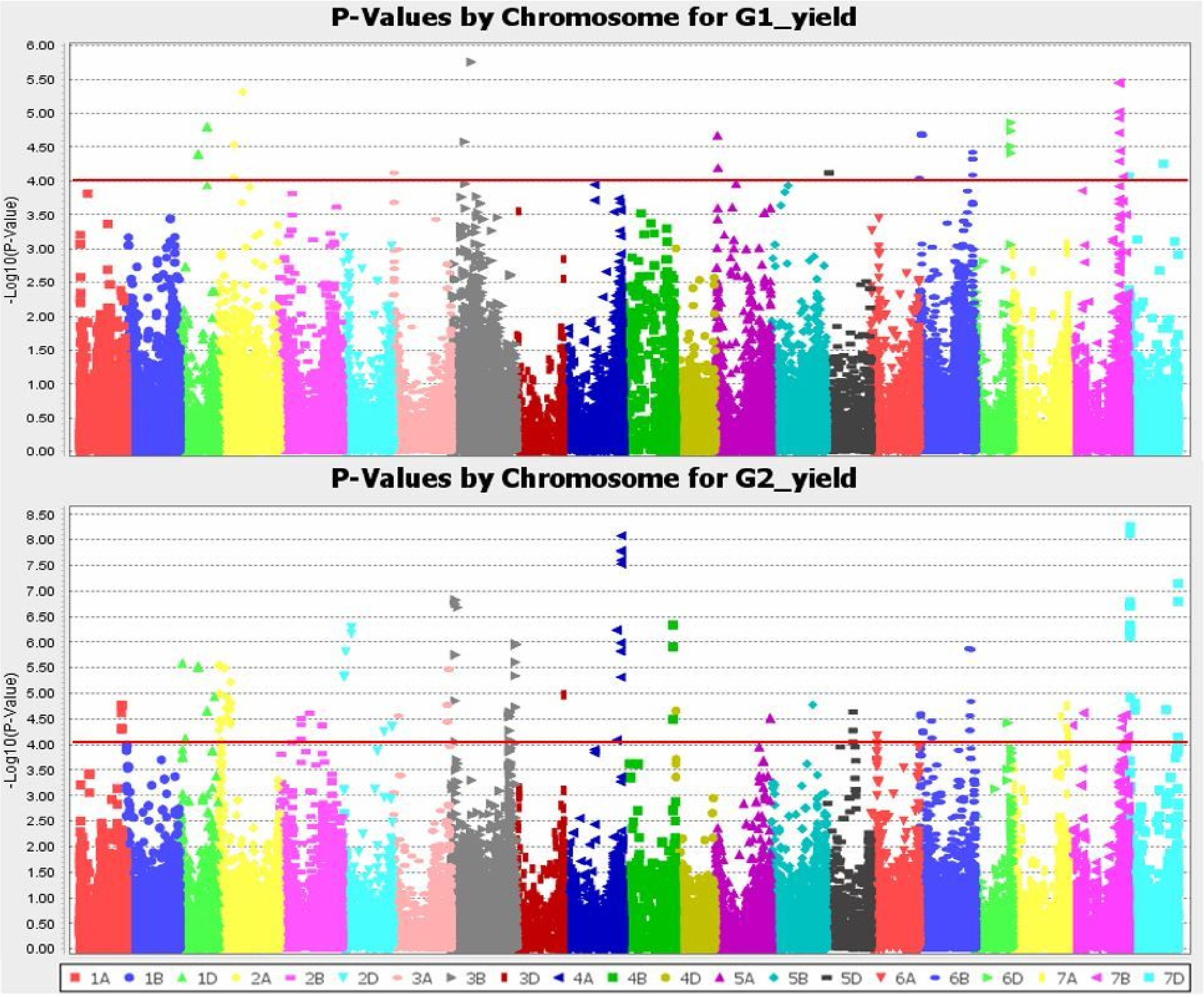


Figure 5

GWAS in TXE collections using Tassel v5.0 identified genomic regions significantly associated with grain yield in correlation-based groups G1 and G2. Critical threshold was set at LOD = 4.0.

Fig. 6

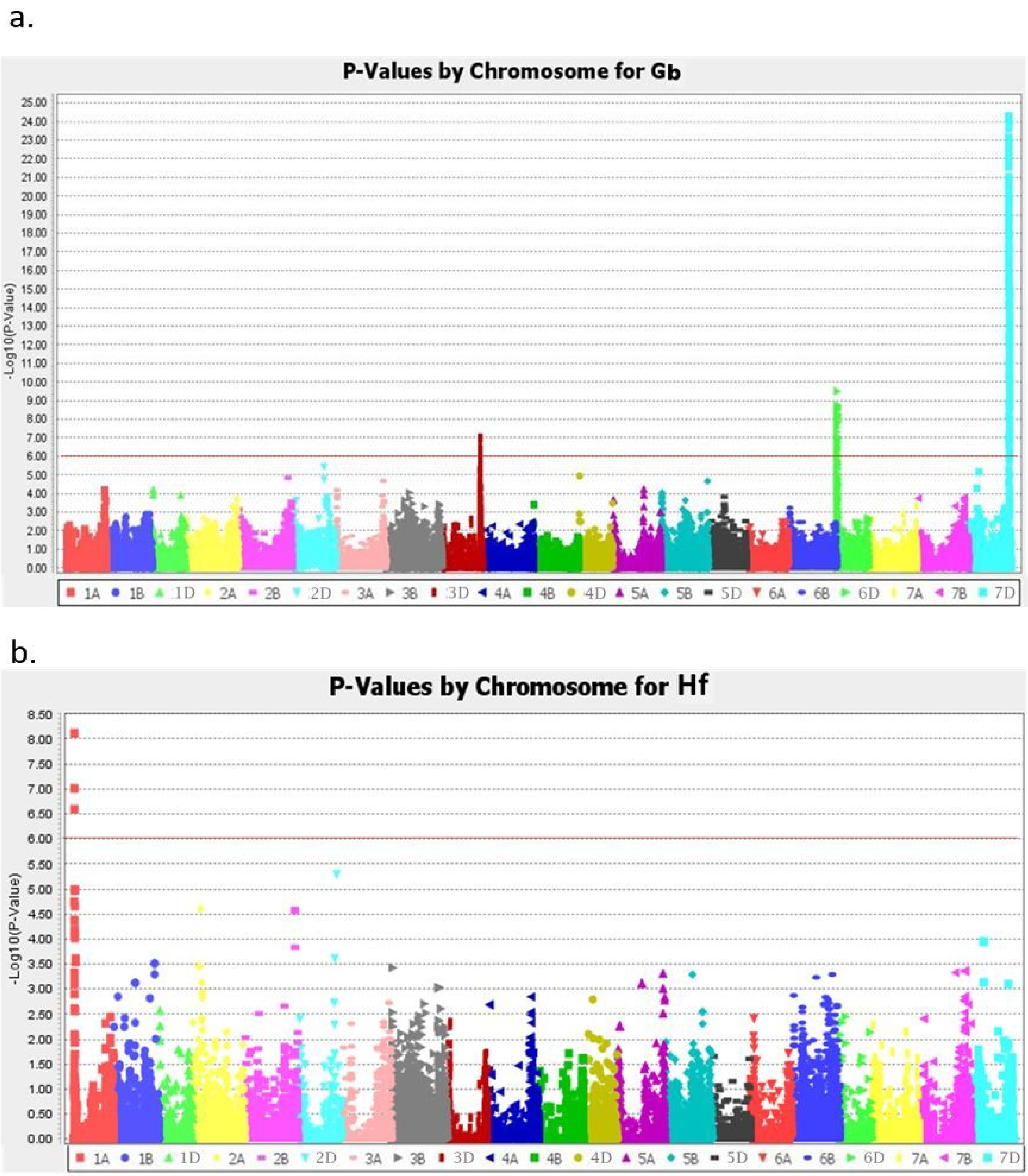


Figure 6

GWAS in TXE collections using Tassel v5.0 identified genomic regions significantly associated with greenbug resistance (a) and Hessian fly resistance (b). Critical threshold was set at LOD = 6.0.

Supplementary Files

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

- Supplementaryfigures.pptx
- Supplementarytables.xlsx